



Artificial Intelligence Governance under Change Foundations, Facets, Frameworks

By

Matthijs M. Maas

This dissertation submitted in partial fulfilment of the requirements for
the degree of PhD in Law, Faculty of Law, University of Copenhagen.

Supervisor: Prof. Hin-Yan Liu

Co-supervisor: Prof. Kristian Cedervall Lauta

Submitted on: September 30th, 2020

Defended on: April 21st, 2021

Abstract

This dissertation explores how we may govern a changing technology, in a changing world, using governance systems that may themselves be left changed.

Artificial Intelligence (AI) has made remarkable progress in the past decade, and is anticipated to become an increasingly disruptive, even transformative technology. AI can be functionally understood as a diverse portfolio of computational techniques to improve the accuracy, speed, or scale of machine decision-making, producing capabilities that can support, substitute for-, or improve upon human task performance. The resulting breadth of application makes AI promising—and challenging—in so many domains of life.

In recent years diverse AI applications—from facial recognition to automated legal decisionmaking, and from computational propaganda to Lethal Autonomous Weapons Systems—have raised deep ethical, political, legal and security concerns. With growing public and policymaker attention has come a wave of governance initiatives and proposals. Nonetheless, global governance for AI remains relatively fragmented and incipient. At this cross-roads, this dissertation takes up the research question, *“How should global governance for artificial intelligence account for change?”*

To answer this question, this dissertation draws together scholarship on technology regulation, (international) law, and global governance, in order to unpack three facets of ‘change’ that will prove critical to the global governance of AI. These three facets of change are examined through the conceptual lenses of *Sociotechnical Change*, *Governance Disruption*, and *Regime Complexity*.

Sociotechnical Change (Chapter 4) explores how and why technological change in AI produces societal changes that create a *rationale* for regulatory intervention, and how we can productively characterize the appropriate *targets* for AI governance. Along with material features, I distinguish six problem logics which highlight different governance solutions and conditions.

Governance Disruption (Chapter 5) addresses when, where and why certain AI capabilities might drive or demand change in the substance (*Development*), tools or processes (*Displacement*) or political scaffolding (*Destruction*) of global governance itself, and what are the implications for global regime complexity.

Regime Complexity (Chapter 6) helps focus attention on how prospective AI regimes are shaped by underlying changes in the broader global governance architecture. It provides insight into the (1) *origins* or foundations of AI regimes; the (2) *topology* of the AI ‘regime complex’; its (3) *evolution* towards integration or fragmentation; (4) the functional *consequences* of these paths; and (5) *strategies* for managing the AI regime complex.

Through these three lenses, this dissertation explores key considerations, insights and tradeoffs for AI governance (Chapter 7). It argues that AI governance needs to shift or adopt novel strategies—in conceptual approach, instrument choice, and instrument design—to ensure the *efficacy* of AI regimes in tracking AI’s sociotechnical impacts, their *resilience* to future AI-driven disruption to the tools, norms or broader conditions of governance, and their *coherence*. In this way, AI governance regimes may remain fit for change.

Resume

Afhandlingen undersøger hvordan vi potentielt kan forvalte de teknologiske forandringer i en verden i forandring ved brug af reguleringsværktøjer og forvaltningssystemer, der selv står over for forandring.

Kunstig intelligens (AI) har undergået bemærkelsesværdig forandringer i det seneste årti, og forventes i stigende grad at skabe opbrud. AI kan forstås som en mangfoldig portefølje af dataadrevne beregningsteknikker skabt til at forbedre præcision, hastighed og omfanget af maskinel beslutningskraft, hvorved der opstår egenskaber, der kan supportere, substituere eller forbedre menneskelig opgaveudførelse. Teknologiens anvendelsesomfang gør AI både lovende og udfordrende.

I de senere år har den stigende anvendelse af AI – fra ansigtsgenkendelse til automatiserede juridiske afgørelser, og fra dataforen propaganda til dødbringende autonome våbensystemer – rejst spørgsmål af dybdegående etisk, politisk, juridisk og sikkerhedsmæssig karakter. Med øget opmærksomhed fra offentligheden og lovgivere, er der kommet nye, globale initiativer til forvaltning af AI fra forskellige interesser. Ikke desto mindre vedbliver den globale forvaltning at være i sin spæde begyndelse, og ganske fragmenteret. I lyset af den forestående teknologiske brydningstid vil denne afhandling fokusere på strukturelle udfordringer ved global forvaltning af AI, med afsæt i spørgsmålet: "Hvordan bør global forvaltning af kunstig intelligens tage højde for forandring?"

For at besvare dette spørgsmål vil denne afhandling sammenholde en række grene af akademisk faglighed for så vidt angår teknologisk regulering, (international) lovgivning og global forvaltning, for at afdække de tre facetter af 'forandring', der vil vise sig at være afgørende for den globale forvaltning af AI. De tre forandringsfacetter belyses via tre linser bestående af Socioteknisk forandring (*Sociotechnical Change*), Forvaltningdisruption (*Governance Disruption*) og Forvaltningskompleksitet (*Regime Complexity*).

Socioteknisk forandring (kapitel 4) tager afsæt i spørgsmål om hvornår, hvordan og hvorfor teknologisk forandring af AI skaber samfundsmæssige omvæltninger, der kræver (skaber et rationale for) reguleringsmæssig indblanding, og hvordan vi på produktiv vis kan karakterisere de relevante mål for forvaltning af AI. Sammen med de materielle træk skelner jeg mellem seks problemlogikker, der fremhæver forskellige forvaltningsmæssige løsninger og betingelser.

Forvaltningdisruption (kapitel 5) belyser hvornår, hvor og hvorfor visse AI-egenskaber kan vise sig at skabe forandring i international lovgivning (Udvikling), værktøjer og processer (Forskydning) eller de politiske strukturer i selve den globale forvaltning (Destruktion).

Forvaltningskompleksitet (kapitel 6) tager afsæt i, hvordan potentielle AI-regimer bliver formet af underliggende forandringer i den bredere globale forvaltningsarkitektur. Dette giver indsigt i (1) oprindelse eller fundament for AI-regimer; (2) topologien af AI 'forvaltningskomplekset'; (3) kompleksitetens evolution mod integrering eller fragmentering; (4) de funktionelle konsekvenser heraf, og (5) strategier for at håndtere AI-regime komplekset.

Denne afhandling vil, igennem disse tre linser, undersøge centrale overvejelser, indsigter og kompromiser for forvaltning af AI (kapitel 7). Afhandlingen argumenterer, at forvaltning af AI skal ændres eller vælge nye strategier – både hvad angår den konceptuelle tilgang, hvad angår valg af værktøj samt værktøjernes design – for at sikre, at AI-regimer effektivt følger AIs sociotekniske indvirkning og standhaftighed over for AI-drevet disruption til værktøjerne, normerne og de bredere betingelser for forvaltning, såvel som for deres sammenhængskraft. Gennem disse strategier kan forvaltningen af AI håndtere forandring.

Acknowledgements

How do we govern a changing technology, in a changing world, using governance systems that may themselves be left changed?

The brief answer is—together.

It has been a privilege and a cherished opportunity to get to spend these past years exploring an issue I both deeply care about and am unendingly fascinated by. However, my true fortune has been to undertake this journey with the support of a community of so many wonderful people. I am deeply grateful to many individuals for their support, feedback and patience throughout.

I would like to express my thanks to my supervisor, Hin-Yan Liu—who took a chance on a wandering student at a conference back in 2016, and who has since then encouraged me to develop these ideas fully. Throughout this project, he provided invaluable reminders to not get too attached to early ideas, and to regularly take a step back and consider my project in a new light. I would also like to thank my co-supervisor, Kristian Lauta, for his infective enthusiasm, and for his facility with clarifying metaphors that really would have made him a wonderful poet if he were not already a great scholar and educator.

I am grateful to Seán Ó hÉigearthaigh, who set me on this path years ago; and to Seth Baum and Allan Dafoe, who provided invaluable advice and opportunities over several years, helping sustain my passion and prepare me for this PhD. I would not have been in a position to undertake this project if not for them. My thanks here also go to Tim Sweijs and Stephan De Spiegeleire at The Hague Centre for Strategic Studies, for our early collaboration on a 2017 study of military AI.

All scholarship is a process of cultivating new shoots of insight within a rich existing ecology of ideas. Mine is no different, and I have learned immensely from the expertise of scholars who are true leaders in their fields, including Roger Brownsword, Thomas Burri, John Danaher, Thore Husfeldt, Karin Kuhlemann, Osonde Osoba, Kenneth Payne and Jacob Turner. Cass Sunstein’s feedback during a 2018 Holberg Prize workshop not only improved that paper, but seeded the ideas for several new ones within one hour. Moreover, I have gained immeasurably from conversations in Copenhagen with Léonard van Rompaey, Andrew Mazibrada, Lena Trabucco, Jacob Livingston Slosser and Nathan Clark, and abroad with Beth Barnes, Søren Elverlin, Gavin Leech, Robert de Neufville, Reinmar Nindler, Nathan Sears, and Maaike Verbruggen.

My early thinking on the global governance of AI benefited tremendously from my participation, from 2016, in a research network organized by Allan Dafoe; during the PhD, I had the opportunity to undertake a research stay with his team at the Centre for the Governance of AI in Oxford, where I finished and submitted paper [I]. I am deeply grateful for the people and the vibrant intellectual community there. In particular, I have gained from the insights from Ashwin Acharya, Markus Anderljung, Jan Brauner, Miles Brundage, Jeffrey Ding, Sophie-Charlotte Fischer, Carrick Flynn, Jade Leung, Gregory Lewis, Rose Hadshar, Carina Prunkl, Anders Sandberg, and Baobao Zhang. The Centre for the Study of Existential Risk organized the 2016 conference where I met Dr. Liu and found the opportunity to undertake this PhD, and I have

since profited immensely from conversations with other CSER scholars, including Shahar Avin, Haydn Belfield, Adrian Currie, Olaf Corry, Martina Kunz, Shin-Shin Hua, Jess Whittlestone.

I am indebted to my co-authors on various papers and projects these past years—especially Seth Baum, Peter Cihon, Keith John Hayward, Luke Kemp, John-Clark Levin, and Charlotte Stix. I have learned so much from them, and I am thankful for the way they have put up with my prose and extensive footnoting: they are absolved of stylistic responsibility for this dissertation. The same goes for Thomas Basbøll, who taught me the best writing methodology that I was never able to properly implement (but which nonetheless left its mark).

Substantively, I have had the privilege of having early drafts of this dissertation reviewed by a range of people who have been enormously generous with their time and attention. David Galbreath twice provided extensive critical comments and guiding feedback on an early draft at a critical moment, transforming. Moreover, for providing detailed thoughts and feedback on the chapters of this final dissertation in these final critical months, I owe thanks to Roger Brownsword, Miles Brundage, Laura Emily Christiansen, Peter Cihon, Jeffrey Ding, James Fox, Eugenio Vargas Garcia, Ross Gruetzmacher, Olle Häggström, Gavin Leech, John-Clark Levin, Sam Lusk, Wilfried Maas, Suvradip Maitra, Sebastian Porsdam Mann, Andrew Mazibrada, Eugenie Carolina van der Meulen, Alessandro Monti, Robert de Neufville, Joe Parrino, Carina Prunkl, Max Reddel, Beatriz Martinez Romera, Léonard van Rompaey, Luisa Scarcella, Zachary Shaw, Jonas Schuett, Jaime Sevilla, Charlotte Siegmann, Lena Trabucco, Theodora Valkanou, Magnus Vinding, Jess Whittlestone, and Kim Zwitserloot. Their comments have made this work immensely better. Any remaining errors are all my own.

My colleagues at the University of Copenhagen have made this process not just an intellectually stimulating journey, but also a thoroughly enjoyable one. For this, I am grateful to Anders Henriksen, Amnon Lev, Beatriz Martinez Romera, Keith John Hayward, Miriam Cullen, Jens Elo Rytter, and Helle Krunk—*as well as my fellow travellers, Léonard van Rompaey, Alessandro Monti, Linnéa Nordlander, Kathrine Elmose Jørgensen, Theodora Valkanou, Annemette Fallentin Nyborg, Sue Anne Teo, William Hamilton Byrne, Katarina Hovden, Ida Gundersby Rognlien, Berdien van der Donk, Gunes Ünvar, Werner Schäfke-Zell, Sarah Scott Ford and Maxim Usynin.*

Finally, to my family, who have supported me through all in this and in life—I would like to especially thank my parents, Patricia and Wilfried, and my brothers, Jonathan and Olivier. You know me so well. I owe this all to you.

I am most of all grateful to Christina Korsgaard, who has helped make the years writing this dissertation some of the happiest of my life.

Declaration

This document constitutes the integrative framework for my paper-based PhD Dissertation. This framework comprises part of my academic work undertaken during the 3-year program, between October 2017 and September 2020. This work is as follows.

Publications presented in this dissertation:

- [I] **Maas, Matthijs M.** "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons." *Contemporary Security Policy* 40, no. 3 (February 6, 2019): 285–311. <https://doi.org/10.1080/13523260.2019.1576464>.
- [II] **Maas, Matthijs M.** "Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot." *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57. <https://doi.org/10.1163/18781527-01001006>.
- [III] **Maas, Matthijs M.** "International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order." *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56.
- [IV] Cihon, Peter, **Matthijs M. Maas**, and Luke Kemp. "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History." In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, 228-34. New York, NY, USA: ACM, 2020. <https://doi.org/10.1145/3375627.3375857>.

Other publications not directly covered in this dissertation:

- **Maas, Matthijs M.** "Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 223–228. AIES '18. New York, NY, USA: Association for Computing Machinery, 2018. <https://doi.org/10.1145/3278721.3278766>.
- **Maas, Matthijs M.** "Two Lessons from Nuclear Arms Control for the Responsible Governance of Military Artificial Intelligence." In *Envisioning Robots in Society – Power, Politics, and Public Space*, Vol. 311. Frontiers in Artificial Intelligence and Applications. IOS Press, 2018. <https://doi.org/10.3233/978-1-61499-931-7-347>.
- Liu, Hin-Yan, Kristian Cedervall Lauta, and **Matthijs Michiel Maas**. "Governing 'Boring' Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research." *Futures* 102 (2018): 6–19. <https://doi.org/10.1016/j.futures.2018.04.009>.
- Baum, Seth D., Stuart Armstrong, Timoteus Ekenstedt, Olle Häggström, Robin Hanson, Karin Kuhlemann, **Matthijs M. Maas**, James D. Miller, Markus Salmela, Anders Sandberg, Kaj Sotala, Phil Torres, Alexey Turchin, and Roman V. Yampolskiy. "Long-Term Trajectories of Human Civilization." *Foresight* 21, no. 1 (February 11, 2019): 53–83. <https://doi.org/10.1108/FS-04-2018-0037>.
- Hayward, Keith J., and **Matthijs M Maas**. "Artificial Intelligence and Crime: A Primer for Criminologists." *Crime, Media, Culture*, June 30, 2020, 1741659020917434. <https://doi.org/10.1177/1741659020917434>.

- Liu, Hin-Yan, Kristian Cedervall Lauta, and **Matthijs M. Maas**. “Apocalypse Now? Initial Lessons from COVID-19 for the Governance of Existential and Global Catastrophic Risks” *Journal of International Humanitarian Legal Studies* 11, no.2 (2020): 1-16. <https://doi.org/10.1163/18781527-01102004>
- Levin, John-Clark, and **Matthijs M. Maas**. “Roadmap to a Roadmap: How Could We Tell When AGI is a ‘Manhattan Project’ Away?” *Proceedings of ECAI/EPAI 2020*. http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf
- Liu, Hin-Yan, **Matthijs Maas**, John Danaher, Luisa Scarcella, Michaela Lexer, and Leonard Van Rompaey. “Artificial Intelligence and Legal Disruption: A New Model for Analysis.” *Law, Innovation and Technology* (September 16, 2020): 1–54. <https://doi.org/10.1080/17579961.2020.1815402>.

Policy documents and non-peer reviewed publications:

- Kemp, Luke, Peter Cihon, **Matthijs Michiel Maas**, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. “Submission to the UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance.” Centre for the Study of Existential Risk, February 26, 2019.
- Belfield, Haydn, José Hernández-Orallo, Seán Ó hÉigeartaigh, **Matthijs M. Maas**, Alexa Hagerty, and Jess Whittlestone. “Consultation on the White Paper on AI – a European Approach”. *Submission to European Commission public consultation*. Centre for the Study of Existential Risk. <https://www.cser.ac.uk/news/response-european-commissions-consultation-ai/> - June 15, 2020.

Other articles currently (as of September 2020) under review or in final preparation:

- Liu, Hin-Yan, and **Matthijs M. Maas**. “Solving for X? Towards a problem-finding framework that grounds long-term governance strategies for artificial intelligence” (forthcoming in *Futures*).
- Stix, Charlotte, and **Matthijs M. Maas**. “The Case for an ‘Incompletely Theorized Agreement’ on AI Policy” (under review).
- Cihon, Peter, **Matthijs M. Maas**, and Luke Kemp. “Fragmentation and the Future: Investigating Architectures for International AI Governance” (forthcoming in *Global Policy*).
- **Maas, Matthijs M.**, and Charlotte Stix. “Cross-community collaboration to address AI impacts and risks: an Incompletely Theorized Agreement for AI policy” (submitted to conference).

Contents

CHAPTER 1. INTRODUCTION	1
1.1 Artificial Intelligence at the Edge of Governance?.....	1
1.1.1 Scoping AI challenges.....	3
1.1.2 Global Governance for AI: Tensions, Developments, Avenues.....	5
1.1.3 Technology, Governance, and Three Questions of ‘Change’	11
1.2 The Research and the Argument.....	15
1.2.1 Research Questions	15
1.2.2 Thesis	18
1.2.3 Research Objectives and Original Contributions.....	20
1.3 Context and approach	24
1.3.1 On the papers composing the present dissertation.....	24
1.3.2 Summary of Methodology.....	27
1.4 Structure of dissertation.....	28
PART I: FOUNDATIONS OF CHANGE.....	33
CHAPTER 2. GLOBAL GOVERNANCE FOR AI: STAKES, CHALLENGES, STATE	34
2.1 Stakes: AI matters and will drive change.....	34
2.1.1 Defining AI Definitions: Three Purposes.....	34
2.1.2 Does AI actually matter? Reasons for caution	39
2.1.3 Progress and promise: a lasting AI Summer?	41
2.1.4 Limits and problems: a coming AI autumn?	46
2.1.5 AI as high-variance technology	54
2.1.6 Change from Further AI Progress: AI between hype and counterhype	56
2.1.7 Change from Algorithmic Proliferation: disruptive AI is already here.....	71
2.1.8 Reflections: how does AI matter?	80
2.2 Challenges: AI needs global governance	83
2.2.1 Scoping AI challenges.....	83
2.2.2 Do these challenges need global cooperation?	86
2.2.3 Is AI global governance possible? Barriers and Levers	89
2.3 State: the global AI governance landscape is insufficient	94
2.3.1 Extending or applying existing International Law.....	94
2.3.2 AI governance: ongoing developments, prospects, and gaps	107
2.4 Conclusion: the world needs global governance for AI.....	122
CHAPTER 3. METHODOLOGY, THEORY, AND ASSUMPTIONS	123
3.1 Methodology and approach	123
3.2 Theoretical Foundations and Rationales	125
3.2.1 Sociotechnical change.....	126
3.2.1 Governance disruption	127
3.2.2 Regime complexity.....	129

3.3	Epistemological Approach.....	132
3.3.1	Scope and use of examples	132
3.3.2	Present and future: on speculation and anticipatory governance.....	134
3.4	Ontological Assumptions	137
3.4.1	The global governance architecture	137
3.4.2	Foundations and dynamics of international regimes.....	140
3.4.3	Technology and society	142
3.5	Methods.....	145
3.6	Conclusion: Methodology, Theory, Assumptions.....	145
PART II. FACETS OF CHANGE		146
CHAPTER 4. SOCOTECHNICAL CHANGE: AI AS GOVERNANCE RATIONALE AND TARGET		147
4.1	Technology and Change	150
4.1.1	A definition of Technology	150
4.1.2	Change in Technology and Change from Technology	152
4.1.3	The concept of sociotechnical change	153
4.2	Sociotechnical change as Governance Rationale.....	154
4.2.1	Thresholds: sociotechnical change and affordances	154
4.2.2	Magnitude of ‘disruptive’ sociotechnical change	155
4.2.3	Governance rationales.....	157
4.2.4	The epistemic limits of anticipating sociotechnical change	159
4.3	Sociotechnical change as Governance Target	163
4.3.1	Regulatory textures and the role of materiality.....	163
4.3.2	Regulatory surfaces and problem logics	166
4.4	Governing AI Sociotechnical Change: a taxonomy of Problem Logics	166
4.4.1	Ethical challenges.....	169
4.4.2	Security threats	171
4.4.3	Safety risks	173
4.4.4	Structural shifts.....	178
4.4.5	Common benefits	184
4.4.6	Governance Disruption.....	185
4.5	Evaluating Sociotechnical Change for AI governance.....	186
4.5.1	Uses: regulatory triage, tailoring, timing, & design	186
4.5.2	Limits: operationalisability, scale, prediction	187
4.6	Conclusion: AI governance for sociotechnical change.....	188
CHAPTER 5. TECHNOLOGY-DRIVEN LEGAL CHANGE: AI AS GOVERNANCE DISRUPTOR.....		190
5.1	International law and technology: an abridged history	193
5.2	An Overview of AI Governance Disruption	195
5.3	Development	197
5.3.1	New Governance Gaps.....	198
5.3.2	Conceptual Uncertainty & Ambiguity	201
5.3.3	Incorrect Scope of Application of Existing Laws	202

5.3.4	Rendering obsolete core assumptions of governance	205
5.3.5	Altering problem portfolio	208
5.3.6	Carrying Through Development: Reflections and Risks.....	209
5.4	Displacement	218
5.4.1	Automation of rule creation, adjudication or arbitration	220
5.4.2	Automation of monitoring & enforcement.....	224
5.4.3	The replacement of international ‘law’? AI and shifts in regulatory modality	227
5.5	Destruction.....	229
5.5.1	Erosion: AI as conceptually or politically intractable puzzle for development	229
5.5.2	Decline: AI as persistent threat to governance architectures	234
5.6	Conclusion: AI and Governance Disruption	239
CHAPTER 6. CHANGES IN GOVERNANCE ARCHITECTURE: REGIME COMPLEXITY AND AI		241
6.1	Regime Theory and the Fragmentation of Governance.....	242
6.2	The Concept of a Regime Complex	244
6.3	The effects and importance of regime complexity	246
6.3.1	Critiques of regime complexity: dysfunction, inequality, strategic vulnerability.....	246
6.3.2	Defences of regime complexity: flexible problem-solving, democratic dynamics	247
6.4	Analysing a Regime Complex.....	249
6.5	Conclusion: AI and regime complexity	251
PART III. FRAMEWORKS FOR CHANGE		252
CHAPTER 7. AI GOVERNANCE IN FIVE PARTS		253
7.1	Origins: the Purpose, Viability and Design of an AI regime	256
7.1.1	Regime purpose: Sociotechnical changes and governance rationales	256
7.1.2	Regime viability: Regime theory and governance foundations.....	267
7.1.3	Regime design: Governance Disruption and Regime Brittleness.....	277
7.2	Topology: the state of the AI governance architecture.....	281
7.2.1	Demographics	282
7.2.2	Organisation: density and links.....	282
7.2.3	Interactions: gaps, conflicts, cooperation, synergies	284
7.2.4	Scope: macro, meso, micro.....	285
7.3	Evolution: trends and trajectories in AI regime complexity	285
7.3.1	Caveats: the limits of predicting or designing a regime complex	285
7.3.2	General drivers of regime fragmentation	286
7.3.3	AI governance disruption at the intersection with regime complexity	292
7.4	Consequences: effects of regime complexity on AI governance.....	297
7.4.1	Should AI Governance be Centralised or Decentralised? Trade-offs from History	297
7.4.2	Political power	299
7.4.3	Efficiency and participation	300
7.4.4	Slowness of establishment or take-off	302
7.4.5	Brittleness vs. adaptation	303
7.4.6	Breadth vs. depth dilemma	305

7.4.7	Forum shopping and strategic vulnerability	308
7.4.8	Policy coordination.....	309
7.5	Strategies: AI regime efficacy, resilience, and coherence	313
7.5.1	AI's Sociotechnical Change and Strategies for Efficacy.....	315
7.5.2	AI Governance Disruption and Strategies for Resilience	318
7.5.3	AI Regime Complexity and Strategies for Coherence	325
7.5.4	An overview of strategies	329
7.6	Conclusion: AI Governance in Five Parts	331
CONCLUSION.....		334
BIBLIOGRAPHY.....		343
APPENDIX: PAPERS [I – IV].....		384

List of Tables

Table 2.1. Overview: how does AI matter? Anticipatory and sceptical arguments.....	81
Table 4.1. Taxonomy of AI problem logics under Sociotechnical Change	168
Table 4.2. Types of AI-induced structural shifts (macro vs. micro), with examples	180
Table 5.1. A Governance Disruption Framework	196
Table 7.1. Overview of analysis of AI Governance Regime.....	255
Table 7.2. Military AI as governance target: considering material and political features	264
Table 7.3. Overview of Strategies for AI regime efficacy, resilience, and coherence	330

List of Figures

Figure 1.1. Conceptual sketch: AI Governance and Three Facets of Change	23
Figure 1.2. Structure and build-up of chapters	29
Figure 2.1. Drezner's typology of Technological Innovation.	77
Figure 3.1. AI Governance and Three Facets of Change (revisited)	125
Figure 4.1. Conceptual sketch: Sociotechnical Change.....	148
Figure 5.1. Conceptual sketch: Governance Disruption	191
Figure 6.1. Conceptual sketch: Regime Complexity.....	242
Figure 7.1. AI Governance and Three Facets of Change (revisited again)	254
Figure 7.2. Fragmented membership in international AI policy initiatives	283

List of Acronyms

ABM	—	Anti-Ballistic Missile (Treaty)
AGI	—	Artificial General Intelligence
AI	—	Artificial Intelligence
ATT	—	Arms Trade Treaty
AWS	—	Autonomous Weapons Systems
CAHAI	—	Council of Europe Ad Hoc Committee on Artificial Intelligence
CAIS	—	Comprehensive AI Services
CBM	—	Confidence-Building Measures
CBRN	—	Chemical, Biological, Radiological and Nuclear
CCGAI	—	Coordinating Committee for the Governance of AI (<i>proposal</i>)
CCTV	—	Closed-Circuit Television
CCW	—	Convention on Certain Conventional Weapons
CEB	—	UN Chief Executives Board for Coordination
CIL	—	Customary International Law
CoE	—	Council of Europe
CS	—	Computer Science
CTBT	—	Comprehensive Test Ban Treaty
DARPA	—	Defense Advanced Research Projects Agency
DETF	—	Digital Economy Task Force
EU	—	European Union
FLOPS	—	Floating Point Operations Per Second
FOBS	—	Fractional Orbital Bombardment System
G7	—	Group of Seven
G20	—	Group of Twenty
GATT	—	General Agreement on Tariffs and Trade
GAN	—	Generative Adversarial Network
GDPR	—	General Data Protection Regulation
GGE	—	Group of Governmental Experts
GPAI	—	Global Partnership on AI
GPT	—	General-Purpose Technology
GPT-2	—	Generative Pretrained Transformer 2 (2019 OpenAI language model)
GPT-3	—	Generative Pretrained Transformer 3 (2020 OpenAI language model)
GPU	—	Graphics Processing Units
HBP	—	Human Brain Project
HLAI	—	Human-Level AI
HLEG	—	High-Level Expert Group on Artificial Intelligence
HLMI	—	High-Level Machine Intelligence
IAEA	—	International Atomic Energy Agency
IAIO	—	International Artificial Intelligence Organisation (<i>proposal</i>)
ICAO	—	International Civil Aviation Organization
ICBM	—	Intercontinental Ballistic Missile
ICC	—	International Criminal Court
ICJ	—	International Court of Justice
ICRAC	—	International Committee for Robot Arms Control
IEEE	—	Institute for Electrical and Electronics Engineers
IHL	—	International Humanitarian Law

IHRL	—	International Human Rights Law
IL & IR	—	International Law & International Relations
ILC	—	International Law Commission
ILO	—	International Labour Organization
IMO	—	International Maritime Organization
IMS	—	International Monitoring System
INF	—	Intermediate-Range Nuclear Forces (Treaty)
IoT	—	Internet of Things
IPCC	—	Intergovernmental Panel on Climate Change
ISO	—	International Organization for Standardization
ISR	—	Intelligence, Surveillance, and Reconnaissance
IT	—	Information Technology
ITER	—	International Thermonuclear Experimental Reactor
ITU	—	International Telecommunications Union
LAWS	—	Lethal Autonomous Weapons Systems
MHC	—	Meaningful Human Control
ML	—	Machine Learning
MTCR	—	Missile Technology Control Regime
NAT	—	Normal Accident Theory
NGAD	—	Next-Generation Air Dominance
NGO	—	Non-Governmental Organization
NLP	—	Natural Language Processing
NPT	—	Nuclear Non-Proliferation Treaty
NSG	—	Nuclear Suppliers Group
OECD	—	Organisation for Economic Co-operation and Development
OoD	—	Out-of-Distribution (inputs to neural networks)
OST	—	Outer Space Treaty
PLA	—	People's Liberation Army
RL	—	Reinforcement Learning
SALT	—	Strategic Arms Limitation Talks
TAI	—	Transformative AI
TBT	—	Agreement on Technical Barriers to Trade
UAV	—	Unmanned Aerial Vehicle
UDHR	—	Universal Declaration of Human Rights
UN	—	United Nations
UNCLOS	—	United Nations Convention on the Law of the Sea
UNGA	—	United Nations General Assembly
UNEP	—	United Nations Environment Programme
UNESCO	—	United Nations Educational, Scientific and Cultural Organization
UNICRI	—	United Nations Interregional Crime and Justice Research Institute
UNIDIR	—	United Nations Institute for Disarmament Research
UNODC	—	United Nations Office on Drugs and Crime
UNU	—	United Nations University
WAMI	—	Wide-Area Motion Imagery
WHO	—	World Health Organization
WTO	—	World Trade Organization

Chapter 1. Introduction

New technologies come and—often fade into the background. They stay, for better or worse, and they each leave their marks, be they large or small: on our industries and our ecosystem; on our fiction and our values; on our views of ourselves and our relations to others. They alter how we constitute orders of law, and how we enter into the chaos of war.

At times, the impact of a technology is broad. It does not have a dramatic impact in any one sector, but it drives many small changes across society. At other times, the impact of a technology is deep. It is highly disruptive in a particular area, even if it does not affect many others. A rare few technologies have an impact on the world that is both unusually broad and deep. Artificial Intelligence ('AI') may be one such transformative technology. This prospect may be ground for anticipation—but also for caution.

To govern a transformative technology is to reckon with questions of change: in technology; in society; and in governance itself. This coming decade will see a growing need and opportunity to ensure that global governance is up to the task of governing the diverse challenges created by AI technology. In doing so, governance strategies and initiatives will have to reckon with these questions of change. This dissertation explores how we can govern changing AI, in a changing world, using governance systems that may themselves be subject to change.

1.1 Artificial Intelligence at the Edge of Governance?

Scientifically, the field of AI has been characterised as “making machines intelligent, [where] intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.”¹ Technically, ‘AI’ is an umbrella term that includes both traditional rule-based symbolic AI as well as the data-driven machine learning (ML) approaches that are responsible for the recent surge in AI progress and attention. Functionally, AI can be described as a varied suite of computational *techniques* which are able to improve the accuracy, speed, and/or scale of machine decision-making across diverse information-processing or decision-making contexts. The resulting *capabilities* can accordingly be used in order to support, substitute for-, and/or improve upon human performance in diverse tasks, enabling their deployment in *applications* across various domains, and resulting in diverse *societal impacts*.

While modern AI techniques are today still subject to many limitations and restrictive preconditions, they are also increasingly able to demonstrate significant performance in diverse tasks, such as those involving data classification, pattern recognition, prediction, optimisation, anomaly detection, or autonomous decision-making.² These functions may be individually narrow,

¹ NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE: A HISTORY OF IDEAS AND ACHIEVEMENTS xiii (2010). For a classic overview, see STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (3rd ed. 2016). For further discussion of definitions of AI (and the underlying purposes of definitions), see also Chapter 2.1.1.

² For accessible introductions, see also: PAUL SCHARRE & MICHAEL C HOROWITZ, *Artificial Intelligence: What Every Policymaker Needs to Know* 23 (2018), <https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know>; GREG ALLEN, *Understanding AI Technology* 20 (2020). For definitions in a legal

but the fact that they are relatively domain-agnostic ensures their relevance across many roles and industries. A facility at recognizing data patterns beyond the abilities of the naked human eye is useful whether the context is medical diagnosis, stock trading, cybercrime detection, or military perimeter control. Moreover, in many applications, AI systems can reach equivalent- or better-than-human performance at such tasks or, at least, can perform to such a standard that scalability and cost-effectiveness considerations may support the technology's deployment in those contexts.³

As a result, AI has in recent years come into its own as a widely applicable set of technologies, with applications in a diverse array of sectors ranging from healthcare to finance, from education to security, and even the scientific process itself.⁴ While the efficacy of today's AI technology is not without its problems and limits, there can be little doubt that AI's development and proliferation will impact many or all sectors of society.

Indeed, the emergence of increasingly capable AI systems might spell a landmark development in history. After all, if many human achievements and successes are the result of our 'intelligent' or adaptive behaviour (broadly defined), then the creation of technologies that manage to automate this capacity is likely to be a significant step.⁵ Even if one is a sceptic, and expects that these technologies will only automate 'facets' of human intelligence, or that in a philosophical sense they can never recreate 'true' human intelligence but may 'only' mimic or approach its external 'appearance', the mere functionalities unlocked by such capabilities would be more than sufficient to drive significant shifts and upheavals across all societies that use them. The potential of AI as a generally enabling technology has consequently been compared by some to 'fire', 'electricity', 'nuclear power' or the 'internal combustion engine'.⁶ If machine 'intelligence'

context, see Jonas Schuett, *A Legal Definition of AI*, ARXIV190901095 Cs (2019), <http://arxiv.org/abs/1909.01095> (last visited Jan 6, 2020); JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE 7–21 (2018); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS LAW REV. 653 (2017).

³ For overviews of AI system performance, see Peter Eckersley & Yomna Nasser, *EFF AI Progress Measurement Project (2017-)*, ELECTRONIC FRONTIER FOUNDATION (2017), <https://www.eff.org/ai/metrics> (last visited Jun 22, 2020).; as well as RAYMOND PERRAULT ET AL., *The AI Index 2019 Annual Report* (2019), https://hai.stanford.edu/sites/g/files/shiybj10986/f/ai_index_2019_report.pdf.

⁴ See generally Maithra Raghu & Eric Schmidt, *A Survey of Deep Learning for Scientific Discovery*, ARXIV200311755 CS STAT (2020), <http://arxiv.org/abs/2003.11755> (last visited Jun 29, 2020). For instance, in one study, the IBM Watson system used language processing algorithms to process thousands of peer-reviewed medical articles on the neurodegenerative disorder amyotrophic lateral sclerosis (ALS). On this basis, it correctly predicted five previously unknown genes related to the disease; Nadine Bakkar et al., *Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis*, 135 ACTA NEUROPATHOL. (BERL.) 227–247 (2018). In another case, an unsupervised Natural Language Processing (NLP) system deployed on the scientific materials science literature was able to capture complex materials science concepts such as the underlying structure of the periodic table, and on the basis of past literature was able to 'predict' later scientific findings by recommending new materials for application, ahead of their eventual discovery. Vahe Tshitoyan et al., *Unsupervised word embeddings capture latent knowledge from materials science literature*, 571 NATURE 95–98 (2019).

⁵ Ross Gruetzmacher & Jess Whittlestone, *Defining and Unpacking Transformative AI*, ARXIV191200747 Cs (2019), <http://arxiv.org/abs/1912.00747> (last visited Jan 6, 2020); Edward Parson et al., *Artificial Intelligence in Strategic Context: An Introduction*, AI PULSE (2019), <https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/> (last visited Feb 26, 2019). In the context of military strategic decision-making, see Kenneth Payne, *Artificial Intelligence: A Revolution in Strategic Affairs?*, 60 SURVIVAL 7–32 (2018).

⁶ Lauren Goode, *Google CEO Sundar Pichai says AI is more profound than electricity or fire*, THE VERGE (2018), <https://www.theverge.com/2018/1/19/16911354/google-ceo-sundar-pichai-ai-artificial-intelligence-fire-electricity>

turns out to be even half as impactful as these past technologies, that would still suffice to make it one of the key developments of this century, and a key driver of societal progress—or decline.

1.1.1 Scoping AI challenges

While many hail the benefits that AI technology could bring, others have begun to express concern about the technology's diverse challenges. As a technology aimed at automating various human activities either in part or in full, AI, almost by definition, has the potential to touch upon virtually any domain of human activity. The use of AI systems is driving considerable social changes—for better and for worse—across a range of dimensions.⁷

At the domestic level, applications of AI technology have raised fundamental ethical concerns in contexts ranging from online advertising to government administration, and from policing to healthcare systems, many of which relate to questions of machine bias, explainability, and accountability.⁸ There are concerns surrounding the ability of AI technology to support or enable pervasive forms of disinformation or online manipulation, which could threaten democratic norms and practices.⁹ The integration of AI systems into various infrastructures or cyber-physical systems such as self-driving cars has introduced a variety of novel safety risks.¹⁰ AI also gives rise to new security threats, as malicious actors are able to exploit AI either as a tool or as a vulnerable attack surface, to scale up old attacks, or to carry out entirely new types

jobs-cancer (last visited Sep 12, 2018); Shana Lynch, *Andrew Ng: Why AI Is the New Electricity*, STANFORD GRADUATE SCHOOL OF BUSINESS, 2017, <https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity> (last visited Oct 22, 2018); Michael C. Horowitz, *Artificial Intelligence, International Competition, and the Balance of Power*, TEXAS NATIONAL SECURITY REVIEW, 2018, <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/> (last visited May 17, 2018).

⁷ For a more detailed discussion of these, see also the survey of AI challenges in Chapter 2.2.1., as well as the taxonomy of how these can be mapped onto six distinct ‘problem-logics’, introduced in Chapter 4.4.

⁸ See KATE CRAWFORD ET AL., *AI Now 2019 Report* 100 (2019), https://ainowinstitute.org/AI_Now_2019_Report.pdf. And an informal compilation, see: Roman Lutz, *Learning from the past to create Responsible AI: A collection of controversial, and often unethical AI use cases*, ROMAN LUTZ ,<https://romanlutz.github.io/ResponsibleAI/> (last visited Jun 22, 2020). On bias in facial recognition systems, see also the landmark study Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 in PROCEEDINGS OF MACHINE LEARNING RESEARCH 1–15 (2018), <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.

⁹ Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCIENTIFIC AMERICAN, 2017, <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/> (last visited May 29, 2017); Danielle Citron & Robert Chesney, *Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics*, 98 FOREIGN AFFAIRS, 2019, <https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war> (last visited Jun 26, 2019).

¹⁰ Dario Amodei et al., *Concrete Problems in AI Safety*, ARXIV160606565 Cs (2016), <http://arxiv.org/abs/1606.06565> (last visited May 13, 2017).

of crimes.¹¹ Accordingly, many uses of AI create foundational challenges for core societal values and for human rights.¹²

Globally, there are growing concerns over the pervasive disruption of economic and social stability,¹³ including through large-scale work displacement and possible changes in the nature of work itself.¹⁴ Moreover, the rise of diverse military uses of AI—including but not limited to ‘lethal autonomous weapons systems’ (LAWS)—has led to concerns regarding its potential incompatibility with human dignity as well as its impact on human control over general dynamics of military escalation or instability.¹⁵ Moreover, AI technology is projected to contribute to widespread challenges in political economy, global power relations, surveillance, and authoritarianism.¹⁶

This is an incomplete survey of a broad and rapidly expanding issue spectrum. Underlying and bracketing this plethora of challenges is the fact that AI technologies may also produce changes in- and to systems of law and regulation themselves. The use of algorithms in legal decision-making or prediction might speed up and expand access to legal processes or increase compliance with important regulations. However, such use can also challenge existing legal doctrines and concepts or more fundamentally alter the practices, processes, or assumptions of governance orders. In doing so, AI technology may alter the underlying logics and assumptions of governance systems, or even our concept of ‘law’ itself.¹⁷ Such steps may well be beneficial; nevertheless, they should be critically interrogated. This is especially because, beyond their impact on the broader systems of law, these changes might also alter the way we can or should approach the governance of (AI) technology itself.

¹¹ Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV180207228 Cs (2018), <http://arxiv.org/abs/1802.07228> (last visited Feb 21, 2018); Thomas C. King et al., *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, SCI. ENG. ETHICS (2019), <https://doi.org/10.1007/s11948-018-00081-0> (last visited Feb 21, 2019); Keith J Hayward & Matthijs M Maas, *Artificial intelligence and crime: A primer for criminologists*, CRIME MEDIA CULT. 1741659020917434 (2020); M. Caldwell et al., *AI-enabled future crime*, 9 CRIME SCI. 14 (2020).

¹² Q. C. VAN EST, J. GERRITSEN & L. KOOL, *Human rights in the robot age: challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality* (2017), <https://research.tue.nl/en/publications/human-rights-in-the-robot-age-challenges-arising-from-the-use-of-> (last visited May 22, 2019).

¹³ Nick Bostrom, Allan Dafoe & Carrick Flynn, *Public Policy and Superintelligent AI: A Vector Field Approach*, in *ETHICS OF ARTIFICIAL INTELLIGENCE* (S.M. Liao ed., 2019), <http://www.nickbostrom.com/papers/aipolicy.pdf> (last visited May 13, 2017).

¹⁴ MARY L. GRAY & SIDDHARTH SURI, *GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS* (2019); Carl Benedikt Frey & Michael A. Osborne, *The future of employment: How susceptible are jobs to computerisation?*, 114 TECHNOL. FORECAST. SOC. CHANGE 254–280 (2017). Though for a more optimistic take, see JOHN DANAHER, *AUTOMATION AND UTOPIA: HUMAN FLOURISHING IN A WORLD WITHOUT WORK* (2019).

¹⁵ AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY, (Nehal Bhuta et al. eds., 2016); Jürgen Altmann & Frank Sauer, *Autonomous Weapon Systems and Strategic Stability*, 59 SURVIVAL 117–142 (2017).

¹⁶ ALLAN DAFOE, *AI Governance: A Research Agenda* 52 (2018), <https://www.fhi.ox.ac.uk/govaiagenda/>.

¹⁷ Benjamin Alarie, *The path of the law: Towards legal singularity*, 66 UNIV. TOR. LAW J. 443–455 (2016); Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 INDIANA LAW J. 1401–1447 (2017); Brian Sheppard, *Warming up to inscrutability: How technology could challenge our concept of law*, 68 UNIV. TOR. LAW J. 36–62 (2018).

1.1.2 Global Governance for AI: Tensions, Developments, Avenues

Some of these above-mentioned issues may well be addressed effectively through policy at the national level. However, many of these issues will gain from some forms of global cooperation. Some critically require it.¹⁸ Public polls indicate growing global concern over AI development,¹⁹ and growing call for forms of policy or regulation to adequately address many of these challenges. It is clear that policymakers have their work cut out for them. However, the difficulty of addressing these challenges internationally is compounded by the (perceived) strategic stakes of the issue, and rising tensions not just around digital governance, but around the architecture of global cooperation more broadly.²⁰

As a result of high-profile breakthroughs, the promise of AI has hardly gone unnoticed. AI, some argue, will be ‘the biggest geopolitical revolution in human history’.²¹ Its prospective impact on national security has been compared to nuclear weapons, aircraft, and computing.²² “Whoever leads in artificial intelligence in 2030,” it has been claimed, “will rule the world until 2100.”²³ Even under more modest readings, AI has been perceived to offer considerable economic and scientific advantages. As such, over the past years, dozens of states have articulated national AI strategies,²⁴ and many have begun investing vast sums in both research and application, including the development of AI systems in strategic, military or cybersecurity applications.²⁵

¹⁸ Amanda Askell, Miles Brundage & Gillian Hadfield, *The Role of Cooperation in Responsible AI Development* 23 (2019).

¹⁹ See EDELMAN, *2019 Edelman AI Survey* (2019), https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_AI_Survey_Whitepaper.pdf (last visited Aug 11, 2020); EDELMAN, *2019 Edelman Trust Barometer: Trust in Technology* (2019), https://www.edelman.com/sites/g/files/aatuss191/files/2019-04/2019_Edelman_Trust_Barometer_Technology_Report.pdf. And in the US context, see BAOBAO ZHANG & ALLAN DAFOE, *Artificial Intelligence: American Attitudes and Trends* 111 (2019), <https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/>.

²⁰ THORSTEN JELINEK, WENDELL WALLACH & DANIL KERIMI, *Coordinating Committee for the Governance of Artificial Intelligence* (2020), https://www.g20-insights.org/policy_briefs/coordinating-committee-for-the-governance-of-artificial-intelligence/ (last visited Jul 8, 2020).

²¹ Kevin Drum, *Tech World: Welcome to the Digital Revolution*, FOREIGN AFFAIRS, 2018.

²² GREG ALLEN & TANIEL CHAN, *Artificial Intelligence and National Security* (2017), <http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf> (last visited Jul 19, 2017).

²³ Indermit Gill, *Whoever leads in artificial intelligence in 2030 will rule the world until 2100*, BROOKINGS (2020), <https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/> (last visited Jan 22, 2020). Note, however, that an often-repeated claim by Russian President Vladimir Putin, that “whoever rules AI rules the world”, may have been taken out of context: rather than an official statement of Russian foreign policy, this appears to have been an off-the-cuff comment which Putin made in the context of giving young Russian school children feedback on their science projects. Robert Wiblin, Keiran Harris & Allan Dafoe, *The academics preparing for the possibility that AI will destabilise global politics* (2018), <https://80000hours.org/podcast/episodes/allan-dafoe-politics-of-ai/> (last visited Aug 12, 2020).

²⁴ See for overviews Jessica Cussins, *National and International AI Strategies*, FUTURE OF LIFE INSTITUTE (2020), <https://futureoflife.org/national-international-ai-strategies/> (last visited Jun 22, 2020). TIM DUTTON, BRENT BARRON & GAGA BOSKOVIC, *Building an AI World: Report on National and Regional AI Strategies* 32 (2018), https://www.cifar.ca/docs/default-source/ai-society/buildinganaiworld_eng.pdf?sfvrsn=fb18d129_4.

²⁵ For instance, on the US side, the Pentagon has emphasized its intention to invest up to \$2 billion in the next five years to develop programs advancing AI—and spent around \$7.4 billion on AI in 2017. Drew Harwell, *Defense Department pledges billions toward artificial intelligence research*, WASHINGTON POST, September 7, 2018, <https://www.washingtonpost.com/technology/2018/09/07/defense-department-pledges-billions-toward-artificial-intelligence-research/> (last visited Jun 22, 2020); Julian E. Barnes & Josh Chin, *The New Arms Race in AI*, WALL STREET JOURNAL, March 2, 2018, <https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261> (last visited Nov 22, 2018). For discussion of funding for military AI projects, see: Justin Haner & Denise Garcia, *The Artificial*

Both the US and China have established AI technology as a lynchpin of their future strategic dominance.²⁶

As a result, some claim we are seeing a new global ‘arms race’ or even an impending ‘Cold War’ in AI,²⁷ and others anticipate the rise of ‘techno-nationalism’ around AI.²⁸ Even if AI’s technological trajectory were not, eventually, to bear out these ambitious promises, and even if zero-sum framings of global AI development as a ‘race’ might be misconceived or even hazardous,²⁹ such depictions are increasingly shaping global debates, and with it the stage for governance efforts.³⁰ In this environment, it should come as no surprise that surveys show that the public is concerned about the appropriate governance of AI technology,³¹ and recent years have seen increasing calls for appropriate societal and regulatory responses.

Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development, 10 GLOB. POLICY 331–337 (2019).

²⁶ In the US, this was initially articulated (in 2016) under the Obama administration; OFFICE OF SCIENCE AND TECHNOLOGY POLICY, *The National Artificial Intelligence Research and Development Strategic Plan* (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf (last visited Feb 26, 2017); This has since been rearticulated under the “American Artificial Intelligence Initiative”: OFFICE OF SCIENCE AND TECHNOLOGY POLICY, *American Artificial Intelligence Initiative: Year One Annual Report* 36 (2020), <https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf>. Meanwhile, China’s State Council in 2017 issued a plan that anticipated China becoming the world leader in the AI field by 2030: CHINA’S STATE COUNCIL, *A Next Generation Artificial Intelligence Development Plan* (Rogier Creemers et al. trans., 2017), <https://n-production.s3.amazonaws.com/documents/translation-fulltext-8.1.17.pdf> (last visited Oct 23, 2017). See also JEFFREY DING, *Deciphering China’s AI Dream: The context, components, capabilities, and consequences of China’s strategy to lead the world in AI* 44 (2018), https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf?platform=hootsuite. For general info on China and AI, one can refer to his ‘ChinAI’ newsletter, at <https://chinai.substack.com/>

²⁷ See Edward Moore Geist, *It’s already too late to stop the AI arms race—We must manage it instead*, 72 BULL. AT. SCI. 318–321 (2016); Haner and Garcia, *supra* note 25; KAI-FU LEE, *AI SUPERPOWERS: CHINA, SILICON VALLEY, AND THE NEW WORLD ORDER* (2018); Nicholas Thompson & Ian Bremmer, *The AI Cold War That Threatens Us All*, WIRED, 2018, <https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/> (last visited Nov 20, 2018); Michael Auslin, *Can the Pentagon Win the AI Arms Race?*, FOREIGN AFFAIRS, 2018, <https://www.foreignaffairs.com/articles/united-states/2018-10-19/can-pentagon-win-ai-arms-race> (last visited Nov 20, 2018); Barnes and Chin, *supra* note 25.

²⁸ Claudio Feijóo et al., *Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy*, TELECOMMUN. POLICY 101988 (2020). See also Ian Hogarth, *AI Nationalism*, IAN HOGARTH (2018), <https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism> (last visited Jul 23, 2018).

²⁹ For compelling critiques of this framing, see for instance Stephen Cave & Seán S. Ó hÉigearthaigh, *An AI Race for Strategic Advantage: Rhetoric and Risks*, in AAAI / ACM CONFERENCE ON ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY (2018), http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf (last visited Mar 12, 2018); Heather M. Roff, *The frame problem: The AI “arms race” isn’t one*, 0 BULL. AT. SCI. 1–4 (2019); Remco Zwetsloot, Helen Toner & Jeffrey Ding, *Beyond the AI Arms Race: America, China, and the Dangers of Zero-Sum Thinking*, FOREIGN AFFAIRS, 2018, <https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race> (last visited Nov 20, 2018); Elsa Kania, *The Pursuit of AI Is More Than an Arms Race*, DEFENSE ONE, 2018, <https://www.defenseone.com/ideas/2018/04/pursuit-ai-more-arms-race/147579/> (last visited Apr 26, 2018).

³⁰ This should be somewhat nuanced. A recent report by the Center for Security and Emerging Technology reviewed coverage of AI in 4,000 English-language articles over a 7-year period, and found that while a growing number of these framed AI development as a competition, these represented a declining proportion of all articles about AI, suggesting narratives around AI are becoming more diverse. ANDREW IMBRIE ET AL., *Mainframes: A Provisional Analysis of Rhetorical Frames in AI* (2020), <https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/> (last visited Aug 19, 2020).

³¹ ZHANG AND DAFOE, *supra* note 19.

Scholarship on society's relation to AI is, of course, not new. There is a long and extensive philosophical literature exploring the ethics of robots.³² There is also a growing body of work exploring questions of how to regulate AI and algorithms at the national level.³³ Initially, this focused on the regulation of robotics,³⁴ later shifting to use cases such as self-driving cars,³⁵ and increasingly focusing on the problems of algorithmic decision-making more broadly.³⁶ This scholarship also comes at a time of growing national regulatory initiatives focused on the regulation of various use cases and applications of AI in domestic jurisdictions.³⁷

At the international level, debates around the global governance of AI are more recent, having only started in the last eight years, and arguably having only picked up momentum since 2016. Initially, these debates focused predominantly on the rise of LAWS.³⁸ This is perhaps not surprising, given that the spectre of 'killer robots' is one that is both particularly visceral, as well as one close to the domain of international peace and security, a main concern of the global legal order. However, in recent years, there has been a growth in broader global AI governance efforts. There has been an explosive growth in the number of AI ethics codes issued,³⁹ and an accompanying rise in discussions exploring broader questions around the global governance of

³² PATRICK LIN, KEITH ABNEY & GEORGE A. BEKEY, ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS (2011); ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE, (Patrick Lin, Keith Abney, & Ryan Jenkins eds., 2017); John Tasioulas, *First Steps Towards an Ethics of Robots and Artificial Intelligence*, 7 J. PRACT. ETHICS 61–95 (2019); Vincent C. Müller, *Ethics of AI and robotics*, in STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2020), <https://plato.stanford.edu/entries/ethics-ai/#Sing> (last visited Apr 2, 2019).

³³ Cf. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 UC DAVIS LAW REV. 37 (2017); Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, HARV. J. LAW TECHNOL. (2016), <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf> (last visited Mar 5, 2018); Michael Guihot, Anne F. Matthew & Nicolas Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, VANDERBILT J. ENTERTAIN. TECHNOL. LAW (2017), <https://papers.ssrn.com/abstract=3017004> (last visited Jul 2, 2018); Ronald Leenes et al., *Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues*, 9 LAW INNOV. TECHNOL. 1–44 (2017); REGULATING ARTIFICIAL INTELLIGENCE, (Thomas Wischmeyer & Timo Rademacher eds., 2020), <https://www.springer.com/gp/book/9783030323608> (last visited Jan 7, 2020); TURNER, *supra* note 2.

³⁴ Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. LAW REV. 513–564 (2015); Neil M. Richards & William D. Smart, *How should the law think about robots?*, in ROBOT LAW (Ryan Calo, A. Froomkin, & Ian Kerr eds., 2016), <http://www.elgaronline.com/view/9781783476725.xml> (last visited Feb 8, 2019).

³⁵ This is an extensive literature, but see for example: Jessica S. Brodsky, *Autonomous Vehicle Regulation: How an Uncertain Legal Landscape May Hit the Brakes on Self-Driving Cars Cyberlaw and Venture Law*, 31 BERKELEY TECHNOL. LAW J. 851–878 (2016); JAMES M. ANDERSON ET AL., *Autonomous Vehicle Technology: a Guide for Policymakers* (2016), https://www.rand.org/pubs/research_reports/RR443-2.html (last visited Oct 17, 2017).

³⁶ This is again an extensive literature, but see Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 671 CALIF. LAW REV. (2016), <https://papers.ssrn.com/abstract=2477899> (last visited Jan 28, 2019); John Danaher, *The Threat of Algoracry: Reality, Resistance and Accommodation*, 29 PHILOS. TECHNOL. 245–268 (2016); Karen Yeung, *'Hypernudge': Big Data as a mode of regulation by design*, 20 INF. COMMUN. SOC. 118–136 (2017).

³⁷ For an overview in different countries, see REGULATION OF ARTIFICIAL INTELLIGENCE IN SELECTED JURISDICTIONS, 138 (2019), <https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf>.

³⁸ AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY, *supra* note 15; HUMAN RIGHTS WATCH, LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS (2012), https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf; Thomas Burri, *International Law and Artificial Intelligence*, 60 GER. YEARB. INT. LAW 91–108 (2017).

³⁹ Anna Jobin, Marcello Ienca & Effy Vayena, *The global landscape of AI ethics guidelines*, NAT. MACH. INTELL. 1–11 (2019); Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI* (2020), <https://dash.harvard.edu/handle/1/42160420> (last visited Jan 16, 2020); Yi Zeng, Enmeng Lu & Cunqing Huangfu, *Linking Artificial Intelligence Principles*, ARXIV181204814 Cs (2018), <http://arxiv.org/abs/1812.04814> (last visited Jan 30, 2019).

AI.⁴⁰ There are an increasing number of UN initiatives exploring the effects and implications of AI.⁴¹ Simultaneously, there are the first cautious steps towards institutionalisation. In 2019, the OECD issued a set of Principles for AI,⁴² which were subsequently endorsed by the G20.⁴³ Earlier, in 2018, France and Canada proposed an ‘International Panel on Artificial Intelligence’, modelled after the IPCC.⁴⁴ While this proposal was initially held up at the 2019 G7, it was re-founded in June 2020 as the ‘Global Partnership on AI’, with 15 founding members.⁴⁵

Clearly, this is an active and exciting time for AI governance. However, the field remains at a vulnerable and uncertain juncture. Many AI governance initiatives are still relatively incipient and fragmented. It remains deeply unclear what trajectory these global governance regimes may take in the coming years. In spite of the topic’s high public profile, regulatory success is certainly not guaranteed. Indeed, without impugning the achievements of international law and global governance over the past decades and centuries, it is also clear that global governance actors cannot or do not always rise to a challenge. Even today, numerous global problems—such as the accumulation of space junk or coral reef degradation, the recognition of professional qualifications for migrants, or the regulation of potential chemical endocrine disruptors—remain under-institutionalised,⁴⁶ or even locked into a ‘non-regime’ state.⁴⁷ Other governance areas, such as the environment or global financial regulation, have become subject to persistent governance ‘gridlock’.⁴⁸ Global cyberspace governance, intuitively perhaps one of the closest technological analogues for AI governance, has seen a number of remarkable multi-stakeholder successes, but

⁴⁰ For overviews, see James Butcher & Irakli Beridze, *What is the State of Artificial Intelligence Governance Globally?*, 164 RUSI J. 88–96 (2019); ANGELA DALY ET AL., *Artificial Intelligence Governance and Ethics: Global Perspectives* (2019), <https://arxiv.org/ftp/arxiv/papers/1907/1907.03848.pdf> (last visited Jun 28, 2019). Daniel Schiff et al., *What’s Next for AI Ethics, Policy, and Governance? A Global Overview* (2020), <https://econpapers.repec.org/paper/osfsocarx/8jaz4.htm> (last visited Jan 12, 2020).

⁴¹ ITU, *United Nations Activities on Artificial Intelligence (AI)* 66 (2018), https://www.itu.int/dms_pub/itu-s/obp/gen/S-GEN-UNACT-2018-1-PDF-E.pdf; ITU, *United Nations Activities on Artificial Intelligence (AI)* 2019 88 (2019), https://www.itu.int/dms_pub/itu-s/obp/gen/S-GEN-UNACT-2019-1-PDF-E.pdf.

⁴² OECD, *Recommendation of the Council on Artificial Intelligence* (2019), <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> (last visited May 28, 2019).

⁴³ G20, *G20 Ministerial Statement on Trade and Digital Economy* (2019), <https://www.mofa.go.jp/files/000486596.pdf> (last visited May 25, 2020).

⁴⁴ Justin Trudeau, *Mandate for the International Panel on Artificial Intelligence*, PRIME MINISTER OF CANADA (2018), <https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence> (last visited Jul 6, 2019).

⁴⁵ Global Partnership on Artificial Intelligence, *Joint Statement from founding members of the Global Partnership on Artificial Intelligence* (2020), <https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding>. For an overview of developments, see Eugenio V Garcia, *Multilateralism and Artificial Intelligence: What Role for the United Nations?*, in THE GLOBAL POLITICS OF ARTIFICIAL INTELLIGENCE 18 (Maurizio Tinnirello ed., 2020); PHILIPPE LORENZ, *AI Governance through Political Fora and Standards Developing Organizations* 41 (2020), <https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations>. See also the overview in Chapter 2.3.2.

⁴⁶ These issues are discussed in; Jean-Frédéric Morin et al., *How Informality Can Address Emerging Issues: Making the Most of the G7*, 10 GLOB. POLICY 267–273, 2 (2019). See also Anne van Aaken, *Is International Law Conducive To Preventing Looming Disasters?*, 7 GLOB. POLICY 81–96, 84 (2016) (on the WHO’s failure to discuss chemical endocrine disruptors).

⁴⁷ Dimitrov et al. have called this is called a ‘nonregime’ state. Radoslav S. Dimitrov et al., *International nonregimes: a research agenda*, 9 INT. STUD. REV. 230–258 (2007).

⁴⁸ THOMAS HALE & DAVID HELD, BEYOND GRIDLOCK (2017).

has also been subject to ongoing contestation, especially in areas of cybersecurity and cyberwarfare.⁴⁹ In other issue domains such as arms control, past successes and landmark achievements in diplomacy have been steadily fraying in recent years.⁵⁰

Will AI governance be able to overcome these hurdles? If so, what forms or strategies should it aim at? The field may well have entered a sensitive and fleeting window of opportunity to ‘get it right’. History suggests that institutional decisions or (mis)steps that are taken at an early stage in governance may fundamentally impact the viability, operation or effectiveness of governance for decades. For instance, in 1946 the Baruch Plan famously proposed a centralised global approach to regulating nuclear weapons and power. In spite of remarkable contemporary enthusiasm for the ideals of ‘world federalism’, the proposal failed, and in doing so marked—or at least accelerated—the beginning of the Cold War.⁵¹ In 1970, the US diplomat George Kennan proposed the establishment of an “International Environmental Agency”, seeing it as an initial step towards an International Environmental Authority.⁵² The merits of a centralised body for environmental governance have continued to be a vexing subject for the field of international environmental governance since then.⁵³ As such, while early governance choices (or the lack of them) may not be irreversible, they do have long-lasting effects, and therefore merit careful consideration.

To be certain, AI governance has seen plenty of proposals for the road ahead. In recent years, scholars have been examining the reach and limits of diverse governance approaches to regulating disruptive AI challenges. These include existing general norms of public international law,⁵⁴ the human rights regime,⁵⁵ a potential universal ‘AI Ethics and Human Rights’

⁴⁹ Daniel W Drezner, *Technological change and international relations*, 33 INT. RELAT. 286–303 (2019). See also ENEKEN TIKK & MIKA KERTTUNEN, *The Alleged Demise of the UN GGE: An Autopsy and Eulogy* 46 (2017), <https://cpi.ee/wp-content/uploads/2017/12/2017-Tikk-Kerttunen-Demise-of-the-UN-GGE-2017-12-17-ET.pdf>.

⁵⁰ AMY J. NELSON, *Innovation Acceleration, Digitization, and the Arms Control Imperative* (2019), <https://papers.ssrn.com/abstract=3382956> (last visited May 29, 2020). Although see Fehl, who argues arms control and export control regimes have displayed remarkable processes of evolution and adaptation historically Caroline Fehl, *Unequal power and the institutional design of global governance: the case of arms control*, 40 REV. INT. STUD. 505–531 (2014).

⁵¹ John Simpson, *The nuclear non-proliferation regime: back to the future?*, DISARM. FORUM 12 (2004); F. Bartel, *Surviving the Years of Grace: The Atomic Bomb and the Specter of World Government, 1945–1950*, 39 DIPLOMATIC HISTORY 275–302 (2015); See also the seminal work by JOSEPH PRESTON BARATTA, *THE POLITICS OF WORLD FEDERATION: FROM WORLD FEDERALISM TO GLOBAL GOVERNANCE* (2004).

⁵² George F. Kennan, *To Prevent a World Wasteland: A Proposal*, 48 FOREIGN AFFAIRS, 1970, at 401–412.

⁵³ See for instance FRANK BIERMANN, *A WORLD ENVIRONMENT ORGANIZATION: SOLUTION OR THREAT FOR EFFECTIVE INTERNATIONAL ENVIRONMENTAL GOVERNANCE?* (Steffen Bauer ed., 1 edition ed. 2005).

⁵⁴ Martina Kunz & Seán Ó hÉigearaigh, *Artificial Intelligence and Robotization*, in OXFORD HANDBOOK ON THE INTERNATIONAL LAW OF GLOBAL SECURITY (Robin Geiss & Nils Melzer eds., 2020), <https://papers.ssrn.com/abstract=3310421> (last visited Jan 30, 2019); TURNER, *supra* note 2 at 7. Burri, *supra* note 38.

⁵⁵ Eileen Donahoe & Megan MacDuffee Metzger, *Artificial Intelligence and Human Rights*, 30 J. DEMOCRACY 115–126 (2019); Lorna McGregor, Daragh Murray & Vivian Ng, *International Human Rights Law as a Framework for Algorithmic Accountability*, 68 INT. COMP. LAW Q. 309–343 (2019). See also the discussion in section 2.3.1.3.

declaration,⁵⁶ global trade governance,⁵⁷ a diverse constellation of AI ethics principles,⁵⁸ standard-setting bodies,⁵⁹ AI certification schemes,⁶⁰ or other forms of soft law;⁶¹ new centralised international agencies to coordinate national regulatory approaches,⁶² ‘Governance Coordinating Committees’⁶³ or even private regulatory markets or regimes founded on ‘natural law’ theories.⁶⁴ Others have emphasised the changing roles of various actors in the AI governance landscape, exploring how states,⁶⁵ private (technology) companies,⁶⁶ AI researchers and expert ‘epistemic communities’,⁶⁷ or trans-national ‘issue networks’ of ‘norm entrepreneurs’⁶⁸ could or should participate in shaping governance efforts for AI.

However, in all this diversity, the field remains relatively new, and in some ways immature. As noted by Allan Dafoe, we still lack a deep understanding of what ‘ideal AI

⁵⁶ Silja Vöneky, *How Should We Regulate AI? Current Rules and Principles as Basis for “Responsible Artificial Intelligence”* (2020), <https://papers.ssrn.com/abstract=3605440> (last visited Sep 1, 2020).

⁵⁷ Han-Wei Liu & Ching-Fu Lin, *Artificial Intelligence and Global Trade Governance: A Pluralist Agenda*, 61 HARV. INT. LAW J. (2020), <https://papers.ssrn.com/abstract=3675505> (last visited Sep 26, 2020).

⁵⁸ For overviews, see Jobin, Ienca, and Vayena, *supra* note 39; JESSICA FJELD ET AL., *Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches* 1 (2019), <https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf>; Zeng, Lu, and Huangfu, *supra* note 39.

⁵⁹ PETER CIHON, *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development* (2019), https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf (last visited Apr 18, 2019).

⁶⁰ Peter Cihon et al., *AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries* 10 (2020).

⁶¹ Wendell Wallach & Gary E Marchant, *An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics* 7 (2018); Gary Marchant, “Soft Law” Governance Of Artificial Intelligence, AI PULSE (2019), <https://aipulse.org/soft-law-governance-of-artificial-intelligence/> (last visited Feb 26, 2019).

⁶² Olivia J Erdelyi & Judy Goldsmith, *Regulating Artificial Intelligence: Proposal for a Global Solution*, in PROCEEDINGS OF THE 2018 AAAI / ACM CONFERENCE ON ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY 95–101 (2018), http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf; TURNER, *supra* note 2 at 6.

⁶³ Wallach and Marchant, *supra* note 61; See also generally Gary E. Marchant & Wendell Wallach, *Coordinating Technology Governance*, 31 ISSUES SCI. TECHNOL. 43–50 (2015).

⁶⁴ Jack Clark & Gillian K Hadfield, *Regulatory Markets for AI Safety* 9 (2019). Y. Weng & T. Izumo, *Natural Law and its Implications for AI Governance*, 2 DELPHI - INTERDISCIP. REV. EMERG. TECHNOL. 122–128 (2019). See also Joshua Z Tan & Jeffrey Ding, *AI governance through AI markets* 7 (2019).

⁶⁵ Some of this has examined states’ domestic relations to strategic assets, or ‘general-purpose technologies’. See for instance Jeffrey Ding & Allan Dafoe, *The Logic of Strategic Assets: From Oil to Artificial Intelligence*, ARXIV200103246 Cs ECON Q-FIN (2020), <http://arxiv.org/abs/2001.03246> (last visited Jan 15, 2020); Jade Leung, *Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies*, July, 2019, <https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665>.

⁶⁶ JESSICA CUSSINS NEWMAN, *Decision Points in AI Governance* 12–29 (2020), https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf (last visited Sep 3, 2020). (exploring the role of AI ethics advisory committees or shifting AI publication norms).

⁶⁷ Haydn Belfield, *Activism by the AI Community: Analysing Recent Achievements and Future Prospects*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 15–21 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375814> (last visited Feb 12, 2020); Matthijs M. Maas, *How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons*, 40 CONTEMP. SECUR. POLICY 285–311 (2019). For a recent exploration of the adjacent concept of ‘trusted communities’, and their role in regulating military robotic and autonomous systems, see also HUGO KLIJN & MAAIKE OKANO-HEIJMANS, *Managing Robotic and Autonomous Systems (RAS)* (2020), <https://www.clingendael.org/publication/managing-robotic-and-autonomous-systems-ras> (last visited Mar 17, 2020).

⁶⁸ Serif Onur Bahçecik, *Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames*, 0 GLOB. POLICY (2019), <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12671> (last visited Jun 17, 2019); Elvira Rosert & Frank Sauer, *How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies*, 0 CONTEMP. SECUR. POLICY 1–26 (2020); Belfield, *supra* note 67.

governance’ arrangements would look like.⁶⁹ Moreover, there are important outstanding conceptual bottlenecks which hobble AI governance efforts, and which require clarification and exploration.

1.1.3 Technology, Governance, and Three Questions of ‘Change’

Much turns on a core seminal question that arises in the context of the regulation of any new technology: how can we account for change? Specifically, how could or should we (re)organize our AI governance approaches in order to ensure that they track and respond to (1) the relation between technological change and societally relevant change, (2) technology-driven change in our regulatory instruments, and (3) external changes in the governance landscape?

These questions are certainly not new. Indeed, scholars in the field of law, regulation and technology have long wrestled with them. For instance, in introducing an influential volume, Roger Brownsword, Eloise Scotford, and Karen Yeung identify three general themes around the idea of (technological) ‘disruption’, namely:

“(1) technology’s disruption of legal orders; (2) the wider disruption to regulatory frameworks more generally, often provoking concerns about regulatory legitimacy; and (3) the challenges associated with attempts to construct and preserve regulatory environments that are ‘fit for purpose’ in a context of rapid technological development and disruption.”⁷⁰

However, while such questions have received extensive attention in the field of law, regulation and technology, they remain, with occasional exceptions, undertheorised or under-integrated in the field of global AI governance. This is problematic, because it is particularly in the field of AI where governance may encounter three especially strong drivers of change.

In the first place, AI governance needs to consider *when, how, and why technological change produces (global) societal changes that warrant regulatory intervention*. At present, AI governance proposals often focus on specific new applications or visceral edge cases, but elide a comprehensive, systematic or explicit consideration of the cross-sector societal change created or enabled by certain AI capabilities. Beyond the intractable debates over how to define AI, we still have little insight into the systematic features or traits of AI as a governance problem in the first place. When do we need governance, and why? AI policies and laws are often formulated in siloed ways that focus on local problems caused by specific use cases of AI (‘drones’, ‘facial recognition’, ‘autonomous vehicles’), or from the perspective of particular conventional legal subjects (e.g. privacy law, contract law, the law of armed conflict).⁷¹ However, they pay relatively little attention

⁶⁹ DAFOE, *supra* note 16 at 48.

⁷⁰ Roger Brownsword, Eloise Scotford & Karen Yeung, *Law, Regulation, and Technology: The Field, Frame, and Focal Questions*, 1 in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY, 4 (Roger Brownsword, Eloise Scotford, & Karen Yeung eds., 2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-1> (last visited Jan 3, 2019).

⁷¹ Rebecca Crootof & B. J. Ard, *Structuring Techlaw*, 34 HARV. J. LAW TECHNOL., 1 (2021), <https://papers.ssrn.com/abstract=3664124> (last visited Aug 28, 2020). (“Technological breakthroughs challenge core legal assumptions and generate regulatory debates. Practitioners and scholars usually tackle these questions by examining the impacts of a particular technology within conventional legal subjects—say, by considering how drones should be regulated under privacy law, property law, or the law of armed conflict. While individually

to the deep interrelation of these issue areas, the common nature of the underlying questions, nor to what are the underlying types of ‘change’ that we are concerned about.

On the one hand, there is some sense in a piecemeal approach: national expertise bodies and international institutions alike have clear, pre-defined areas of expertise.⁷² Yet on the other hand, this approach limits our ability to understand the sources of AI’s ‘symptomatic’ challenges.⁷³ Lacking an understanding of when, and why, AI technology translates into relevant societal change renders AI governance efforts vulnerable to capture by visceral or particularly symbolic edge use cases, at the cost of addressing more widespread but less visible or visceral societal changes. More generally, it creates the risk that governance responses may at best duplicate one another, and at worst work at cross-purposes.⁷⁴ AI governance proposals could be grounded in a better understanding of the precise vectors of cross-sector change, as well as how new capabilities or developments might shift this ‘problem portfolio’. To achieve this, it is valuable to approach AI governance through the lens of ‘*sociotechnical change*’.⁷⁵ This concept focuses on when, how, and why changes in technologies actually expand human capabilities in ways that give rise to new activities, ways of being, or relationships.⁷⁶ As such, it enables a better examination of when AI-enabled behaviour actually creates new problems or negative externalities and why these provide rationales for governance.

Secondly, it is important to consider *how AI technology changes the tools, processes and assumptions of governance itself*. At present, governance proposals for AI often fail to factor in technology-driven changes to governance itself. To be sure, independent bodies of scholarship exist which explore the use of AI, algorithms and digital technology in legal systems,⁷⁷ including a small but growing body of work exploring this phenomenon at the international law level.⁷⁸

useful, these siloed analyses mask the repetitive nature of the underlying questions and necessitate the regular reinvention of the regulatory wheel.”). Petit calls this the ‘legalistic’ approach; NICOLAS PETIT, *Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications* 2–3 (2017), <https://papers.ssrn.com/abstract=2931339> (last visited May 11, 2020).

⁷² Morin et al., *supra* note 46.

⁷³ Léonard van Rompaey, *Discretionary Robots: Conceptual Challenges in the Legal Regulation of Machine Behaviour*, 2020. See also Margot E Kaminski, *Authorship, Disrupted: AI Authors in Copyright and First Amendment Law*, 51 UC DAVIS LAW REV. 589–616 (2017).

⁷⁴ TURNER, *supra* note 2 at 218–221.

⁷⁵ See Lyria Bennett Moses, *Recurring Dilemmas: The Law’s Race to Keep Up With Technological Change*, 21 UNIV. NEW SOUTH WALES FAC. LAW RES. SER. (2007), <http://www.austlii.edu.au/journals/UNSWLRS/2007/21.html> (last visited Jul 3, 2018); Lyria Bennett Moses, *Regulating in the Face of Sociotechnical Change*, in THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 573–596 (Roger Brownsword, Eloise Scotford, & Karen Yeung eds., 2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49> (last visited May 13, 2017).

⁷⁶ This draws on the account in Lyria Bennett Moses, *Why Have a Theory of Law and Technological Change?*, 8 MINN. J. LAW SCI. TECHNOL. 589–606., 591–592 (2007).

⁷⁷ This literature is extensive. But see: Roger Brownsword, *In the year 2061: from law to technological management*, 7 LAW INNOV. TECHNOL. 1–51 (2015); Yeung, *supra* note 36; Alarie, *supra* note 17; Anthony J Casey & Anthony Niblett, *Self-driving laws*, 66 UNIV. TOR. LAW J. 429–442 (2016); Christopher Markou & Simon Deakin, *Is Law Computable? From Rule of Law to Legal Singularity*, in IS LAW COMPUTABLE? CRITICAL PERSPECTIVES ON LAW + ARTIFICIAL INTELLIGENCE (Christopher Markou & Simon Deakin eds., 2020), <https://papers.ssrn.com/abstract=3589184> (last visited May 15, 2020); Sheppard, *supra* note 17.

⁷⁸ Burri, *supra* note 38; Ashley Deeks, *High-Tech International Law*, 88 GEORGE WASH. LAW REV. 575–653 (2020). See also the much more detailed discussion of this work in Chapter 5.

However, most proposals for the global governance of AI itself often reason in relative isolation from such dynamics.

This is a problem, because, as noted by Colin Picker, technological innovations have driven the “creation, modification, or destruction of international law, or the derailment of the creation of new international law” throughout history.⁷⁹ Indeed, scholars working in various governance areas have already begun to identify the adverse effects which digital technologies may have on the continued viability of regimes in areas such as arms control.⁸⁰ Likewise, AI technology is likely to affect not just the substance of various international regimes, but also their processes and procedures, and potentially even their political scaffolding. It is unlikely that these disruptive effects will remain isolated to all other legal domains. Rather, they may also exert effects on the very AI governance instruments that are being proposed, or on the broader global legal order within which these would be ingrained, in ways that may well render certain governance strategies for AI more or less viable.

Thirdly, this field should consider better *how AI governance regimes are shaped by underlying changes in the broader global governance landscape*. At present, some proposals for AI governance take their cue from past multilateral treaties.⁸¹ Yet it is unclear how closely or well-aligned such designs are with the broader context and trajectory of the global governance system, which is itself undergoing considerable change. While it is valuable to understand ‘ideal’ cooperative governance blueprints for AI,⁸² in practice AI governance will not occur in a vacuum. Indeed, Karen Alter and Kal Raustiala have argued that new cooperative efforts rarely occur on a blank slate, because of the complexity of the existing global governance architecture. Instead, they argue that all too often:

“[g]lobal governance solutions [...] must take one of two approaches: (a) International actors can attempt to create an encompassing regime that can address all dimensions of the problem, or (b) international actors can accept that policy solutions will be crafted, coordinated, and

⁷⁹ Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDOZO LAW REV. 151–219, 156 (2001).

⁸⁰ Amy J Nelson, *The Impact of Emerging Technologies on Arms Control Regimes* (2018), <http://www.isodarco.it/courses/andalo18/paper/iso18-AmyNelson.pdf>; NELSON, *supra* note 50. And see generally, Richard Danzig, *An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order*, 5 LAWFARE RES. PAP. SER. (2017), <https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf> (last visited Sep 1, 2017).

⁸¹ For instance, Microsoft called for a ‘Digital Geneva Convention’. Smith, *The need for a Digital Geneva Convention*, MICROSOFT ON THE ISSUES (2017), <https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/> (last visited Apr 17, 2019). There have been diverse calls for an international ban on killer robots. The ‘Campaign Against Killer Robots’ has taken inspiration from global bans on blinding lasers and anti-personnel mines HUMAN RIGHTS WATCH, *Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition: Memorandum to Convention on Conventional Weapons Delegates* (2015), <https://www.hrw.org/news/2015/11/08/precedent-preemption-ban-blinding-lasers-model-killer-robots-prohibition> (last visited Apr 28, 2017). Though notably, others have questioned the adequacy of these comparisons. Rosert and Sauer, *supra* note 68. Rebecca Crootof, *Why the Prohibition on Permanently Blinding Lasers is Poor Precedent for a Ban on Autonomous Weapon Systems*, LAWFARE (2015), <https://www.lawfareblog.com/why-prohibition-permanently-blinding-lasers-poor-precedent-ban-autonomous-weapon-systems> (last visited Apr 17, 2019). See also Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO LAW REV. 80 (2015).

⁸² DAFOE, *supra* note 16 at 48–51 (emphasizing the need for research into “values and principles”, “institutions and mechanisms”, and ‘positive visions’).

implemented within a larger regime complex. [...] although the first option might be more efficient and effective, it is rarely the solution adopted.”⁸³

To bear this out, the past decades have been marked by significant changes in the global institutional landscape. These include patterns of *institutional proliferation*,⁸⁴ the ongoing *fragmentation* of international law resulting in complex inter-regime impacts and externalities,⁸⁵ and growing patterns of *contested multilateralism*.⁸⁶ Others have identified trends of *legal stagnation*, and have argued that global governance is increasingly marked by a shift towards informality in many issue areas such as international environmental governance or cyberspace.⁸⁷ This suggests that the governance landscape in which AI governance finds itself differs in important ways from the landscape of the past.

All of this is certainly not to suggest that historical lessons can no longer provide value,⁸⁸ nor that governance for AI must do away with all the old, and blithely adopt all the new. It is certainly not the case that the role (or rule) of traditional international law has been eclipsed.⁸⁹ Nonetheless, it has been argued that traditional formal international institutions may be ill suited to address complex, cross-sectoral issues such as AI.⁹⁰ The argument is simply that it is important to consider how well proposed AI governance mechanisms, institutions, or arrangements might slot onto, and interact with the complex, evolving architecture of global governance.

In sum, there are three facets of change to consider. (1) changes in AI capabilities which drive *sociotechnical changes* in international society—and which require governance; (2) changes in AI capabilities which *drive substantive, procedural, or political changes in the instruments of international law and global governance*; (3) *change in the broader governance architecture* around any specific AI governance regime.

⁸³ Karen J. Alter & Kal Raustiala, *The Rise of International Regime Complexity*, 14 ANNU. REV. LAW SOC. SCI. 329–349, 337 (2018).

⁸⁴ Kal Raustiala, *Institutional Proliferation and the International Legal Order*, in INTERDISCIPLINARY PERSPECTIVES ON INTERNATIONAL LAW AND INTERNATIONAL RELATIONS: THE STATE OF THE ART 293–320 (Jeffrey L. Dunoff & Mark A. Editors Pollack eds., 2012).

⁸⁵ MARTTI KOSKENNIEMI & STUDY GROUP OF THE INTERNATIONAL LAW COMMISSION, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law* (2006), http://legal.un.org/ilc/documentation/english/a_cn4_1682.pdf; Frank Biermann et al., *The Fragmentation of Global Governance Architectures: A Framework for Analysis*, 9 GLOB. ENVIRON. POLIT. 14–40 (2009).

⁸⁶ Julia C. Morse & Robert O. Keohane, *Contested multilateralism*, 9 REV. INT. ORGAN. 385–412 (2014); Amitav Acharya, *The Future of Global Governance: Fragmentation May Be Inevitable and Creative* Global Forum, GLOB. GOV. 453–460 (2016); Michael Zürn, *Contested Global Governance*, 9 GLOB. POLICY 138–145 (2018).

⁸⁷ J. Pauwelyn, R. A. Wessel & J. Wouters, *When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking*, 25 EUR. J. INT. LAW 733–763 (2014). On cyberspace governance, see also JOSEPH S. NYE, *The Regime Complex for Managing Global Cyber Activities* (2014), <https://dash.harvard.edu/bitstream/handle/1/12308565/Nye-GlobalCommission.pdf> (last visited Sep 3, 2019).

⁸⁸ Indeed, they can at least provide precedent about the plausibility of achieving governance even on contested, high-stakes technologies. See Maas, *supra* note 67. *Paper [II]*.

⁸⁹ See Karen J. Alter, *The Future of International Law*, 101 ICOURTS WORK. PAP. SER. (2017), <https://papers.ssrn.com/abstract=3015177> (last visited Jun 11, 2020); Eyal Benvenisti & George W. Downs, *Comment on Nico Krisch, “The Decay of Consent: International Law in an Age of Global Public Goods”*, 108 AJIL UNBOUND 1–3 (2014).

⁹⁰ Morin et al., *supra* note 46.

Individually, each of these three facets of change ought to be an important consideration in any discussion of AI global governance. Taken together, they arguably provide a set of foundational considerations with which AI governance scholarship must engage, if it is to reckon with change.

1.2 The Research and the Argument

My *main aim* in the present work is to explore how we can govern this changing technology, in a changing world, using regulatory tools that may themselves be left altered. The hope is that this can help ground improved global governance strategies for AI, given a regulatory target, regulatory tools, and a regulatory environment that are in various and increasing states of flux.

Concretely, in articulating the three analytical perspectives on ‘change’ in AI governance, this work builds on four published papers,⁹¹ and draws connections between three bodies of work—on regulating for sociotechnological change; technology-driven legal ‘disruption’, and regime complex theory—and applies them to extant problems and questions in AI governance. By reading these frameworks through illustrative examples and cases, I aim to add conceptual clarity as well as highlight potentially fruitful avenues of further research.

1.2.1 Research Questions

The central research question explored below is:

- “*How should global governance for artificial intelligence account for change?*”

This core question entails a set of thematic sub-questions. Drawing on the work in the core papers, these questions are explored in detail in the individual chapters of this work.

- A. *Why do we require governance strategies for artificial intelligence? Why do these require new strategies for change?*
 - a. Why would AI drive extensive change? Does the technology actually matter?
 - i. What is the current state of AI technology? What are its uses and limitations?
 - ii. What are possible and plausible future trajectories in AI progress? How do we engage with uncertainty around these changes?
 - iii. What are possible trends in the proliferation and application of existing AI capabilities? Why or how would we expect these to create disruptive change?
 - b. Why would AI’s changes require global governance?
 - i. What global policy challenges do or could AI applications create?

⁹¹ The arguments of these papers, and their relation to the analysis in this document, will be introduced and discussed shortly, in Section 1.3.1.

- ii. Why and on what grounds would such challenges warrant or require global cooperation?
- iii. Would global governance for AI technology be viable? What barriers does it face? What levers are available?
- c. How do current governance solutions to these AI challenges fare?
 - i. How might we apply existing norms or instruments under existing international law? What are their assumptions, strengths and limits?
 - ii. What are the currently proposed global governance approaches to AI? What are their strengths and limits?

Sub-question A. will be explored in Part I (*'Foundations of Change'*), especially in Chapter 2, which discusses the background of AI, whether or when or why it matters, whether or how it might be governed, and what is the state and adequacy of existing governance approaches, initiatives and proposals.

The next three sub-questions relate to the three facets of change that affect AI governance. These are explored in the Chapters in Part II (*'Facets of Change'*).

B. *Why, when, and how should governance systems approach and respond to AI-driven sociotechnical change?*

- a. How should AI governance consider concepts of ‘technology’ and ‘technological change’? How do these concepts relate to the concept of ‘sociotechnical change’?
- b. When, where, and why could AI applications produce relevant ‘sociotechnical’ changes?
 - i. How can we characterize and understand *types* of sociotechnical change?
 - ii. When and on what grounds can sociotechnical change create a governance *rationale*?
 - iii. How can the resulting governance responses engage with AI as governance *target*?
 - iv. How can these types of AI-driven sociotechnical changes (and their rationales and governance) be mapped onto distinct ‘problem logics’?
- c. What are the analytical strengths and shortfalls of this lens to AI governance?

Sub-question B will be explored in Chapter 4 (*‘Sociotechnical Change: AI as Governance Rationale and Target’*).

C. *Why, when, and how might AI applications disrupt global governance, by driving or necessitating changes to its substance and norms, its processes and workings, or its political scaffolding?*

- a. How, and in what ways have technologies historically affected and shaped international law?

- b. How, when and why might AI applications produce conceptual or doctrinal uncertainty in existing laws, necessitating processes of ‘development’ in law and governance?
- c. How, when and why might AI applications support the creation, enforcement, or even ‘replacement’ or international regimes, driving processes of governance ‘displacement’?
- d. How, when and why might AI applications produce situations of conceptual or political friction in existing regimes, or contribute to outright contestation, in ways that drive processes of governance ‘destruction’?

Sub-question C will be explored in Chapter 5 (‘Technology-Driven Legal Change: AI as Governance Disruptor’).

D. *Why and how might changes in the broader global governance architecture, as well as amongst individual AI regimes, affect the prospects, development and efficacy of the ‘regime complex’ for AI?*

- a. What is the constitution of the global governance architecture?
- b. What are governance regimes? What are regime complexes?
- c. What are the effects or consequences of regime complexity? What are the anticipated outcomes of fragmentation or integration?
- d. How can regime complex theory be used to structure analysis of an AI regime complex? What stages of development or analytical questions does it highlight?

Sub-question D will be explored in Chapter 6 (‘Changes in Global Governance Architecture: Regime Complexity and AI’).

Finally, in Part III (*Frameworks for Change*), I work out the implications of these three lenses for a range of selected AI governance issues and contexts. As such, I delve into the following sub-questions:

E. *What insights can these three conceptual frameworks provide in exploring the prospects and dynamics of the emerging AI governance regime complex?*

- a. How can we examine the *origin* of a governance regime for a given AI issue? What are its foundations?
 - i. What is the regime’s *purpose*? What ‘sociotechnical change’ does it respond to, and how does this create a governance rationale?
 - ii. What is the regime’s *viability*? How is this shaped by interests and norms?
 - iii. What are the regime’s optimal and adequate *designs* in the face of change?
- b. How can we chart the *topology* of the global AI governance ecology at a certain moment in time? How can we analyse its demographics, its organisation (in terms of normative and institutional linkages), and interactions, at various levels (micro, meso, macro) of the governance architecture?

- c. How can we study potential *evolution* or development in AI regime complexes, given both general trends and drivers of fragmentation, as well as the effects of governance disruption?
- d. What might be the *consequences* of various regime complex configurations for the efficacy, resilience and coherence of AI governance? What trade-offs would the AI governance system face, depending on whether they are centralised or decentralised?
- e. What *strategies* might AI governance scholars and actors adopt, in terms of conceptual approach, instrument choice, or instrument design, in order for AI governance regimes to remain fit for change?
 - i. Which strategies could allow AI regimes to ensure *efficacy* in responding to AI-driven sociotechnical changes?
 - ii. Which strategies could allow AI regimes to ensure *resilience* against further AI-driven governance disruption?
 - iii. Which strategies could allow AI regimes to ensure *coherence* of the regime complex, in order to avoid or manage conflicts amongst institutions or regimes?

These questions will be explored in Chapter 7 ('AI Governance in 5 Parts').

1.2.2 Thesis

The core argument of this dissertation will be that:

Main thesis: to remain fit for change, global governance for artificial intelligence will have to adopt new strategies in order to effectively track the technology's sociotechnical impacts, remain resilient to further AI-driven disruption in the tools, norms or broader conditions for governance, and coherently manage the interactions of the AI governance regime complex.

A rough outline of my reasoning in support of this argument, where each step requires much more elaboration and defence, is sketched below.

1. *AI technology will drive significant changes which will need global governance:*
 - a. While AI still faces important limitations as of this writing, the technology will prove an important driver of global change. It is already seeing significant and widespread application. Continued progress in AI capabilities, or even simply the proliferation of existing algorithmic techniques to more domains, will more than suffice for AI to give rise to diverse global challenges.
 - b. While not all of these issues demand international cooperation to be successfully addressed, many will gain from it, and some challenges will require it, because they create classic global public good problems.

- c. Unfortunately, the emerging AI global governance regime remains incipient and fragmented, and may not prove adequate. Therefore, new strategies are required to improve the efficacy, resilience and coherence of global governance for AI in the face of change.
2. *To remain effective, AI governance should track sociotechnical change:*
- a. Analyses of AI governance regimes should not just focus on isolated AI use cases or legal domains; rather, they should also include or be informed by an investigation into when and how shifts in AI capabilities drive cross-domain patterns of sociotechnical change. Specifically, new AI applications can create new forms of conduct or ways of being, or facilitate old forms of conduct.
 - b. Under some conditions, these resulting sociotechnical changes can give rise to one or more of the following governance *rationales*: (1) possible market failures; (2) new risks to human health or safety; (3) new risks to moral interests, rights, or values; (4) new threats to social solidarity; (5) new threats to democratic processes; (6) new and direct threats to the coherence, efficacy or integrity of the existing governance system charged with mitigating these prior risks.
 - c. In designing regulatory responses, one should consider the regulatory texture of the AI application as governance *target*. This includes some minimum consideration of the technology's material or artefactual characteristics. It should also consider the 'problem logic' that characterises the AI-driven change. We can parse problem logics into 'ethical challenges', 'security threats', 'safety risks', 'structural shifts', 'common benefits', and 'governance disruption', each of which can have distinct origins or barriers, and which invoke different governance logics or responses.
3. *To remain resilient, AI governance should anticipate governance disruption:*
- a. AI can produce patterns of 'governance disruption': here, certain uses of AI can (1) result in legal uncertainty in existing laws or regimes, creating a need for governance *development*; (2) contribute to processes of 'legal automation', resulting in the *displacement* of certain governance practices; (3) result in patterns of governance *destruction* which could further erode areas of international law.
 - b. The ability of AI to produce regular pressures towards governance disruption may erode the resilience of certain traditional approaches in international law, such as multilateral treaties. In other cases, however, AI technologies can speed up or support global governance.
4. *To remain coherent, AI governance should expect regime complexity:*
- a. General dynamics in the global governance architecture, as well as 'governance disruption' by AI applications, weakly suggest there are diverse pressures towards continued fragmentation of the AI regime complex. For some domains this may be adequate; but not for all. Whichever trajectory the regime takes, strategies will be necessary to ensure coherence and coordination within the regime. In a decentralised regime, active efforts must be undertaken to achieve regime harmonisation and modernisation. If a centralised institution is viable, its efficacy

and resilience would depend sensitively on a set of institutional design choices, as well as ways to manage external relations.

5. *AI governance requires new strategies to account for change:*

- a. Given the above, policymakers and scholars should consider shifts in conceptual approach, instrument choice, and instrument design, in order to ensure the efficacy, resilience and coherence of AI governance regimes going forward.

1.2.3 Research Objectives and Original Contributions

As noted, a key aim of this project is to highlight the three facets of change and their interactions. This approach is motivated by the hope that further conceptual clarity and disaggregation can help the emerging field of AI governance explore and design improved governance architectures for AI, which are more effective, coherent, legitimate and resilient. Indeed, such conceptual clarification has been fundamental to much past scholarship at the intersection of ‘law’, ‘regulation’ and ‘technology’.⁹² For instance, Colin Picker has noted that a high-level view of strategic dynamics amongst technology and international law can be critical:

“[e]ven though policy makers must be closely concerned with the “nitty gritty” of their international regimes and negotiations, [...] [they] have much to gain from taking a macro or holistic view of the issues raised by technology. Macro-examinations can provide larger theoretical understandings and can reveal previously hidden characteristics that are simply not discernable from the “trenches.” Viewing technology from “40,000 feet up” reveals certain patterns, pitfalls, and lessons for policy.”⁹³

In this spirit, the present work aims to introduce and develop three lenses which may better capture major drivers or surfaces of ‘change’ in AI governance. Specifically, it draws from three bodies of scholarship—on law, technology and society; technology-driven legal disruption, and regime complex theory—in order to weave together and situate these concepts amidst the new and developing field of research on AI governance.

I will briefly introduce the background of each of these three lenses, explain the way I apply them, and outline the resulting theoretical contributions.⁹⁴

First, in terms of (a) the complexity of *AI as a changing governance rationale- and target*, I draw on scholarship on Sociotechnical change, in particular as articulated in the work of Lyria Bennett Moses.⁹⁵ I use this lens to illustrate and characterise the differences between focusing governance debates on underlying AI technologies, specific applications or use cases, or on considering which socio-technical changes warrant governance intervention. This lens aims at clarifying the interlinkages between these levels, in order to better understand whether new behaviours enabled or produced by AI systems in fact give rise to governance rationales and, if

⁹² For instance, in reflecting on these topics, Roger Brownsword and others have noted that “debates over these terms, and about the conceptualization of the field or some parts of it, can significantly contribute to our understanding.” Brownsword, Scotford, and Yeung, *supra* note 70 at 6.

⁹³ Picker, *supra* note 79 at 151–152.

⁹⁴ More detailed reflections on the choice for these three lenses, their origins, and their strengths and limitations to the context of AI governance, is provided in Chapter 4.1-4.2.

⁹⁵ Bennett Moses, *supra* note 76; Bennett Moses, *supra* note 75; Bennett Moses, *supra* note 75.

so, how such interventions should be tailored to the governance ‘target’ in question. Such a systematic approach can contribute valuable insight for both research and practice. In terms of scholarship, it enables systematic and comparative analysis of AI governance strategies. In practical terms, while this lens is applied to analyse the emerging regime complex for AI, it is in principle also applicable to other domains of global technology governance.

Secondly, in terms of (b) *AI technology’s ability to change the norms, instruments, or processes of governance*, I elaborate a ‘governance disruption’ framework.⁹⁶ This framework departs from the existing scholarship on ‘law and technology’ and ‘TechLaw’, which has explored at length how new technologies can alter the concepts, texture and dynamics of legal systems. However, while most of this work has to date focused on a domestic law context, the current work will also aim to extend such analysis to the international level, by exploring how the use of AI may affect the future form and viability of international law and global governance.

The idea that new technologies are not just objects for regulation, but can also change the operation and processes of legal systems, and even the goals of regulators, is hardly new. Reflections on what a new technology reveals about the changing face of law has extensive precedent in technology law scholarship. For instance, in early debates in the field of cyberlaw, Lawrence Lessig famously examined several legal questions involving the new technology of cyberspace, not merely in order to discuss the relative efficacy of different approaches to regulating certain topics (e.g. zoning or copyright) on the internet, but also to ground and illustrate systemic reflections on the changing nature and workings of the different ‘regulatory modalities’ of laws, norms, markets and architectures (‘code’).⁹⁷ Likewise, Roger Brownsword has used studies of behaviour-shaping technologies and geoengineering in order to reflect respectively, upon the rising role of non-normative ‘technological management’ and ‘regulatory responsibilities’ for the core ‘global commons’.⁹⁸

Moreover, this comparison with earlier scholarship on cyberlaw is illuminating in both directions. The internet ‘merely’ led to the ‘informatisation’ of infrastructure and public space—and already proved pivotal in altering the ways that regulation operates. In their turn, AI systems may enable the increasing ‘intelligentisation’ or ‘cognitisation’ of these infrastructures,⁹⁹ suggesting that this technology may in time have an impact on the practices and dynamics of law and governance that is at least as far-reaching. The ‘governance disruption’ framework is therefore apt, since exploring the dynamics of AI governance can reveal interesting and important

⁹⁶ Previously set out in Paper [III]: Matthijs M. Maas, *International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order*, 20 MELB. J. INT. LAW 29–56 (2019). A version of this is also developed at greater length in Hin-Yan Liu et al., *Artificial intelligence and legal disruption: a new model for analysis*, 0 LAW INNOV. TECHNOL. 1–54 (2020).

⁹⁷ Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. LAW REV. 501 (1999); LAWRENCE LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE, VERSION 2.0 (2nd Revised ed. edition ed. 2006), <http://codev2.cc/download+remix/Lessig-Codev2.pdf>.

⁹⁸ ROGER BROWNSWORD, LAW, TECHNOLOGY AND SOCIETY: RE-IMAGINING THE REGULATORY ENVIRONMENT (1 edition ed. 2019); Roger Brownsword, *Law and Technology: Two Modes of Disruption, Three Legal Mind-Sets, and the Big Picture of Regulatory Responsibilities*, 14 INDIAN J. LAW TECHNOL. 1–40 (2018).

⁹⁹ This draws on the terminology used by the Chinese PLA. Elsa Kania, *AlphaGo and Beyond: The Chinese Military Looks to Future “Intelligentized” Warfare*, LAWFARE (2017), <https://www.lawfareblog.com/alphago-and-beyond-chinese-military-looks-future-intelligentized-warfare> (last visited Jun 10, 2017).

lessons concerning the changing nature of technology governance specifically, and of 21st-century international law more broadly.

Thirdly, in terms of (c) the broader trends, trajectories, and *changes in the global governance architecture*, this work aims to draw on the extensive scholarship on ‘regime complexity’, and to draw this framework into debates for AI governance. Theoretically, this enables more refined discussion over the development and prospects of AI governance regimes, in terms of the *origin* of individual institutions or regimes; the *topology* of the regime complex, potential drivers of *evolution* in the regime architecture development towards fragmentation or integration, the *consequences* of either trajectory, and *strategies* to mitigate adverse consequences, and improve efficacy, coherence, and resilience.

In return, it also aims to contribute to the existing scholarship on regime complexity, by expanding the foundations for a case study in this new domain. Moreover, by considering the effects of the governance disruption framework, this analysis aims to contribute new insights to theoretical debates over the relative role of rationalist (interest-based), ideational (norms-based), and ‘material’ (that is, artefactual or architectural) factors in the constitution and maintenance of global governance systems.

Practically, the regime complexity framework provides a valuable tool to debates around AI governance. It is at present still unclear if AI will receive effective governance, or if recent initiatives might fizzle, leaving the topic lingering in a non-regime or gridlock state. If some lasting governance system does emerge these coming years, it is still unclear whether this will continue to take the shape of decentralised or fragmented regimes, or whether a centralised treaty or institution may eventually emerge. To be clear, this is not meant to provide predictions about the direction of governance, but rather to sketch a series of conditional scenarios.¹⁰⁰ However, present uncertainty over the future of AI governance only renders it all the more important to explore various scenarios and trajectories as well as their strategic and normative implications.

It is, of course, impossible to be exhaustive of all the distinct scenarios. Nonetheless, general themes can and should be discussed. For example, if AI remains largely into a non-regime state, the arguments here highlight the urgency of resolving and addressing it. Moreover, even if one considers it unlikely that a centralised regime will emerge for AI, should circumstances shift in the coming years, such that the creation of such an institution does become viable, then it will be of particular importance to ensure that this regime is well designed and fit for purpose, since poor institutional design could lock in catastrophic outcomes. Conversely, in a fragmented AI ‘regime complex’ consisting of many parallel and overlapping institutions, one should expect to see clusters and various (conflictive or cooperative) interactions between different institutions. This analysis is intended to help navigate such questions, by exploring dependencies and implications today.

In sum, the theoretical approach utilised in this work results in three conceptual lenses which focus on how distinct trends of change can have direct and indirect implications for AI governance regimes (see Figure 1.1).

¹⁰⁰ For further reflections on methodology and epistemology, see also Chapter 3.3.

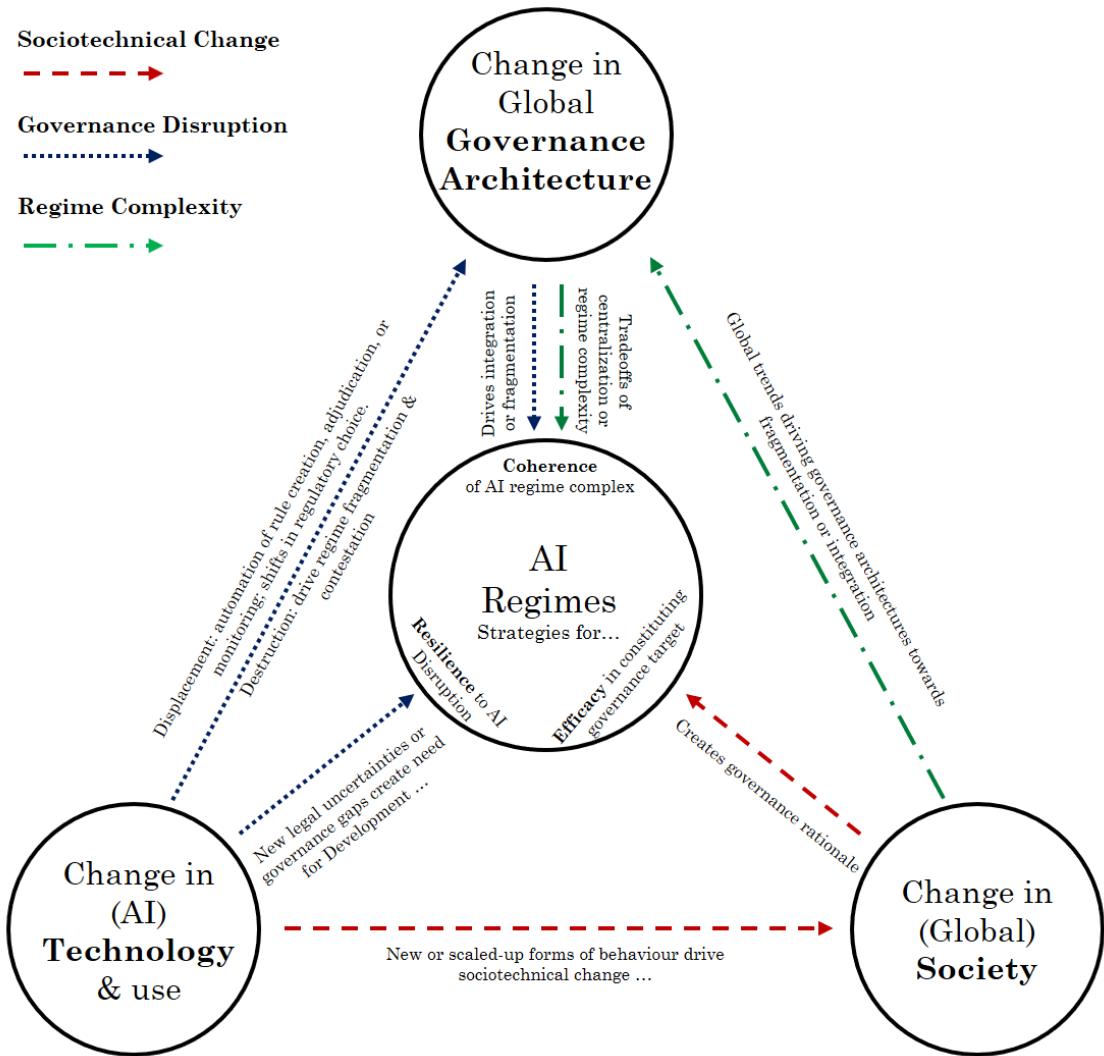


Figure 1.1. Conceptual sketch: AI Governance and Three Facets of Change

As such, the work presented below is aimed at the benefit of several audiences. The three perspectives therein presented—‘sociotechnical change’, ‘governance disruption’, and ‘regime complexity’—are, it is argued, potentially useful conceptual tools for understanding challenges and opportunities for AI governance regimes in a range of contexts. Moreover, they are useful for various global policy actors seeking to understand AI governance in its broader landscape. They help such actors identify points of intersection or leverage onto the trajectory of governance, and to articulate or refine various substantive, procedural or functional design desiderata for AI governance regimes and institutions.

Having outlined these contributions, it is also important at this stage to clarify what this analysis does *not* aim to do.¹⁰¹ Firstly, while the three conceptual lenses have many contact points, and are commensurable, the aim is not to provide a comprehensive or unified framework that

¹⁰¹ For further discussion, see also Chapter 4.1.

must be adopted in its entirety. Rather, the goal of this work is to provide a set of exploratory perspectives and tools on AI governance which, taken individually, can help inform more nuanced, targeted and resilient governance strategies for AI.¹⁰² One need not accept or adopt all three lenses at the same time, but can simply focus productive analysis on one amongst them. Secondly, this is not meant to be a detailed or exhaustive application of each of these lenses. Rather, it is intended as an initial exploration to highlight their promises and limits, as a demonstration of their promise, and as an indication that they warrant further exploration and application.

1.3 Context and approach

As a paper-based PhD dissertation, this document constitutes the theoretical framework which encapsulates, supplements, and builds upon the findings of my research conducted over the course of my PhD fellowship. It should therefore be read as more explorative than a traditional monograph. In particular, the following chapters draw thematic links between four published papers.¹⁰³

To provide some background and context to what follows, the below section briefly summarizes the four papers¹⁰⁴ and provides an overview of the methodological choices and theoretical assumptions informing this project.¹⁰⁵

1.3.1 On the papers composing the present dissertation

In chronological order of publication, the four papers that make up the core of this contribution are:

Paper [I]: Maas, Matthijs M. "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons." *Contemporary Security Policy* 40, no. 3 (February 6, 2019): 285–311. <https://doi.org/10.1080/13523260.2019.1576464>.

In the face of growing alarmism or pessimism in recent years around the imminence of the 'AI arms race', Paper [I] explores the political viability of international arms control for military AI systems of particular concern on the grounds of legality, ethics, safety or stability. The paper draws a comparison between the historical experience of the (non)proliferation and control of another pivotal, strategically high-stakes technology—nuclear weapons. The article

¹⁰² These methodological choices are discussed and defended in greater detail in Chapter 4.

¹⁰³ As of the time of submission (September 2020), three of these papers have been published in peer-reviewed domain journals; the last has been accepted (after double-blind peer review through a selective process) in the proceedings of the *2020 AAAI/ACM Conference on AI, Ethics, and Society*, with a considerably altered version of the paper currently under review at a highly ranked policy journal. These papers have been reproduced in their original layout in the Appendix.

¹⁰⁴ In terms of reading order, the chapters in this dissertation cover and refer to the content of the four papers throughout, and therefore can be read as a standalone argument. Alternatively, a recommended reading order would be to read Part I (Chapters 2-3); Papers I-IV (in the Appendix), and then to proceed to Chapters 4-7.

¹⁰⁵ These are discussed and defended at greater length in Chapter 3.

first discussed the strengths and limits of this analogy, before proceeding to review three lessons that could be derived for the arms control of military AI.

By reviewing the differential roles of security interests, bureaucratic politics, and norms in driving or constraining nuclear proliferation decisions, the paper argues that (1) international arms control for military AI might be more politically viable than is anticipated ahead, and might be achieved through various forms of engagement with domestic policy groups, as well as norm institutionalisation.

By studying the role of ‘epistemic communities’ in laying the foundations for the 1972 Anti-Ballistic Missile (ABM) Treaty, the paper suggests that (2) small communities of experts, appropriately organised and mobilised, can play a disproportionate role in shifting state norms and perceptions in order to frame global AI arms control cooperation.

Finally, by drawing on a ‘normal accident theory’ (NAT) analysis of (near)-failures in the operation of complex nuclear forces, the paper argues that (3) many military AI systems could meet the criteria for ‘normal accidents’, and that this susceptibility could inhibit the viability or sufficiency of governance arrangements grounded on (narrow formulations of) the concept of ‘meaningful human control’ (MHC).

Paper [III]: Maas, Matthijs M. “Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot.” *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57. <https://doi.org/10.1163/18781527-01001006>.

Confronted with the possibility that ongoing innovation in military AI or robotics systems might bypass LAWS, Paper [II] explores the question of whether, or how, we might formulate ‘innovation-proof’ governance approaches which might be able to leapfrog or anticipate future disruptive innovations in military AI, or which would be more resilient or adaptive to it.

The article first discussed various trends and advances in military applications of AI systems, arguing that such ongoing innovation may give rise to a severe case of the ‘Collingridge Dilemma’ (whereby regulatory initiatives at an early stage of the technology’s development face an *information* problem of not knowing what or why to regulate, whereas later regulation faces a *power* problem because the technology has matured, such that interests and stakes have been established), and could challenge the efficacy of extant governance approaches. Accordingly, the paper developed a typology for understanding the distinct ways in which different innovations in military AI may challenge or disrupt existing international legal frameworks.

This framework differentiated between ‘direct’ governance disruption—whereby new types of military AI systems or capabilities appear to elude inclusion under existing regimes—and ‘indirect’ governance disruption, whereby new military AI systems instead shift the technology’s risk landscape (‘problem portfolio’), or change the incentives for states in ways that erode compliance with regimes. In the latter case, innovation in military AI could erode compliance, by increasing the perceived military *benefits* of the technology, by reducing the *barriers* to access, by reducing the political *costs* of noncompliance; or even by shifting state *values*.

Finally, the article explores ‘innovation-proof’ strategies for military AI. It discusses and rebuts some potential objections to attempting to make governance innovation-proof: it argued

that innovation-proof governance need not require the ability to make accurate technological predictions (and indeed, that this might sometimes be politically counterproductive, by clarifying states' differential expected stakes), but can be grounded in adaptive regimes that anticipate change; and that while such 'innovation-proof' regimes might be more politically difficult to negotiate than narrow governance aimed at direct problems, they are less intractable than attempting to amend or resolve too-narrow regimes at a later stage.

Accordingly, the paper sketches the advantages and shortcomings of three possible approaches to tackling direct governance disruption ('firefighting', 'technology neutrality', or 'adaptive governance'), arguing that the first strategy may often be insufficient, and that the second and third strategy are more promising but may require some reconfiguration or clarification in governing (military) AI. It also discussed ways that epistemic communities might shape norms and shift policy to address AI's indirect disruption, by working to challenge policymakers' perceptions of the unambiguous military *benefits* of using unproven AI systems, by increasing the *barriers* to access, by improving the capabilities for detecting treaty compliance violations or for increasing the reputational *costs*; and by *countering value shifts* towards AI-enabled unilateralism.

Paper [III]: Maas, Matthijs M. "International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order." *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56.

Paper [III] transfers the focus from governance regimes for (military) AI, to a broader analysis of how AI might affect the integrity, viability, form or relevance of the international law system itself. The paper reviews scholarship on the long history of technological innovation driving, shaping or destroying international law. Drawing on a range of historical cases, as well as a series of previous theoretical frameworks charting the relation between law and technology, the paper argues that the deployment and use of AI systems might produce three types of global legal impacts: 'legal development', 'legal displacement', and 'legal destruction'.

The paper discusses the potential impact of AI in all three modalities. It argues that AI uses might result in many situations of legal uncertainty, ambiguity, over- or under-inclusiveness, or obsolescence of various norms or regimes under international law, driving a need for these to be accommodated or addressed through legal *development*. However, it also notes that in principle the international law system is able to carry out such developments.

In terms of legal *displacement*, the paper argues that while AI technologies have promise in the monitoring of compliance, the prospects for a shift towards an 'automated international law' might be relatively slim.

Finally, the paper argues that technical and political features of the technology may in practice render AI destructive to key areas of international law: the legal gaps it creates may be conceptually or politically hard to patch through 'development', and the strategic capabilities it offers chip away at the rationales for powerful states to engage fully in, or comply with, international law regimes.

Paper [IV]: Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History." In *Proceedings of the 2020*

AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), 228-34. New York, NY, USA: ACM, 2020. <https://doi.org/10.1145/3375627.3375857>.

Noting that the global governance landscape for AI remains relatively incipient and fragmented, the co-authored Paper [IV] departs from the question, ‘can effective international governance for artificial intelligence remain fragmented, or is there a need for a centralised international organisation for AI?’ To answer this question, we refer to theoretical scholarship on the ‘fragmentation’ of global governance architectures, and the emergence of dense ‘regime complexes’ of parallel and overlapping institutions and norms. We then draw on the history of various other international regimes—especially in the areas of trade, environment and security—in order to identify potential advantages, disadvantages, or trade-offs in the centralisation of a regime.

We argue that some considerations, such as ensuring centralised authority and power as well as averting the inefficiency costs of a fragmented regime, speak in favour of centralised institutions. Conversely, we suggest that the slow process of establishing centralised AI institutions or treaties, their relative ‘brittleness’ once created, and the ‘breadth vs. depth dilemma’ of securing wide-enough support for meaningful rules, all speak against centralisation. Other considerations, such as the regime system’s ability to deter ‘forum shopping’ by actors, or the greater relative ability of a centralised institution to ensure adequate policy coordination, are more ambiguous, and depend on detailed trade-offs or contexts.

We conclude by offering two recommendations. First, we argued that overall, the above trade-offs might depend on the exact design of a centralised AI institution: a well-designed centralised regime covering a set of coherent AI issues could be beneficial, yet locking-in an inadequate or ill-designed integrated regime could proffer a fate worse than fragmentation. Second, we argued that since fragmentation will likely persist in AI governance for the time being, steps should be taken to improve the monitoring of this regime complex, in order to assess whether interactions between the emerging AI institutions and norms are resulting in conflict, coordination, or proactive ‘catalyst’ and self-organisation by the regime complex to fill governance gaps.

Through the following Chapters, the current document weaves together, revises, elaborates, and explores the implications of the arguments, themes, theories, and cases explored in these four papers.

1.3.2 Summary of Methodology

While these choices are discussed and defended at greater length below,¹⁰⁶ it is valuable to briefly review some of the main methodological choices and epistemological and ontological assumptions involved.

This contribution primarily consists of conceptual and theoretical analyses of the concept and implications of ‘change’ for AI governance. Locating itself at the intersection of various strands of scholarship on AI governance, on ‘law, regulation and technology’, and on regime

¹⁰⁶ See Chapter 3.

complexity and the broader discipline of ‘International Law and International Relations’, this project draws on an interdisciplinary range of theoretical frameworks, concepts and cases.¹⁰⁷

Epistemologically,¹⁰⁸ as a theoretical exploration, this project explores the three lenses with reference to various examples or historical cases. However, these are not presented as structured case studies, but simply as clarifying cases or ‘existence proofs’ of certain technology governance dynamics, illustrating the operation of the processes (e.g. governance disruption) under examination. In addition, while this project will broadly remain rooted in- and focused on near-term questions for AI governance, it does at a few points discuss potential future trajectories in both AI capabilities as well as in AI governance. Where it does so, the aim here is not to provide predictions, but rather to sketch implications of different scenarios under modest assumptions. The aim is to illustrate the utility and implications of the lenses across these scenarios.

Ontologically,¹⁰⁹ in terms of my model of world order, I adopt relatively broad definitions of the relevant actors, instruments, and trends active in the global governance architecture. In discussing potential foundations and dynamics of international regimes within this system (especially in the analysis of regime complexity), I adopt an integrated perspective that combines rationalist and constructivist arguments, assuming international regimes reflect both (state) interests but also norms which can be affected and shifted by various actors.

This argument also adopts a mixed ontology with regards to the interaction of *technologies with society*. It does not assume hard determinist effects, but it makes three modest assumptions, namely, that: (1) technical systems can have socio-political goals or effects embedded into them (a ‘*technological politics*’ assumption); (2) that many of the second- or third-order consequences of technological choices are hard to foresee, demonstrating how sociotechnical changes can be hard to predict (an assumption of ‘*unintended consequences*’); that (3) even if it is ‘*sociotechnical change*’ (and not ‘*technology*’) which creates a rationale for governance, a technology’s material features or qualities may still play a role in determining the ‘*texture*’ of the technology as governance target, and in tailoring appropriate regulation in response (*‘asymmetric governance rationale/target materiality’*).

1.4 Structure of dissertation

The argument of the current contribution unfolds according to the below schema (See Figure 1.2).

¹⁰⁷ See also Chapter 3.1.

¹⁰⁸ See also Chapter 3.3

¹⁰⁹ See also Chapter 3.4.

Part

		<u>Chapter 7</u>	
III Frameworks		AI Governance in 5 Parts <i>(Papers I – IV)</i>	
II Facets	<u>Chapter 4</u> Sociotechnical Change <i>(Paper I)</i>	<u>Chapter 5</u> Governance Disruption <i>(Papers II, III)</i>	<u>Chapter 6</u> Regime Complexity <i>(Paper IV)</i>
I Foundations	<u>Chapter 2</u> 'Global Governance for AI' Rationale, background and context (‘What?’ & ‘Why?’)	<u>Chapter 3</u> 'Methodology, Theory, Assumptions' Methodology and theory (‘How?’)	

Figure 1.2. Structure and build-up of chapters

In Part I (*Foundations of Change* – Chapters 2-3), I set out the background of this project—why it is needed, how it conceptualizes ‘AI’; how it situates itself in existing scholarship, and what methodological, epistemic, and ontological assumptions it makes.

In Chapter 2 (*Global Governance for AI: Stakes, Challenges, State*), I provide the background and rationale for this work, and ground the need for new concepts and strategies for AI governance to account for change. As such, the chapter discusses various developments in AI technology, its challenges, and developments and gaps in the global governance landscape. It first (2.1) examines the *importance* of AI technology. After reviewing three types of definitions of AI—and highlighting this project’s understandings of AI from each—it examines arguments pro and contra AI’s importance today. By taking stock of AI systems’ achievements as well as underappreciated limitations, it is argued that rather than a panacea or complete hype, AI is best considered as a ‘high-variance’ technology comprising many applications that are limited, but where a small share of peak applications may produce considerable impact.

I then explore arguments and uncertainties around how this importance could develop within the coming years, considering whether, how, or under what conditions we could expect change to come from either ‘further progress’ or from ‘algorithmic proliferation’. On this basis, the chapter argues that even under conservative assumptions about further fundamental progress, AI’s impact on the world the coming years will likely be far-reaching. This chapter then (2.2) explores the global challenges of- and to AI governance. It argues that some AI issues may require global cooperation, drawing on typologies of global public goods. It also argues that while AI

governance faces specific barriers, there are also distinct levers that can make such governance viable. Finally, this chapter (2.3) surveys developments in the field of AI governance. It reviews the strengths and limits of customary international law, the human rights framework, as well as various existing regimes and frameworks in addressing AI challenges, before surveying recent developments in AI ethics and AI governance. It concludes that in spite of promising steps, the current AI governance landscape remains incipient and fragmented, such that there is a need for new lenses, approaches and solutions to help inform debate.

In Chapter 3 (*Methodology, Theory, and Assumptions*), I (3.1.) introduce the overall methodology and eclectic approach of this project, and discuss some of its limits. I (3.2.) sketch the three core theoretical lenses of sociotechnical change, governance disruption, and regime complexity, and justify their analytical relevance to AI governance. I (3.3.) clarify my epistemological approach, in terms of the use of cases and examples, as well as the ‘anticipatory’ orientation to future technological- and governance trajectories. I (3.4.) briefly set out the project’s ontological assumptions: I detail the actors, instruments, and processes that make up the global governance architecture. I adopt and defend a hybrid positions about the role of interests and norms in establishing international regimes, and defend three assumptions about the interaction of technology with human agency (technological politics; unintended consequences; asymmetric governance rationale/target materiality). I (3.5.) conclude by briefly reviewing the methods by which this research was carried out.

In Part II (*Facets of Change – Chapters 4-6*), I set out the three theoretical perspectives on AI governance.

In Chapter 4 (*Sociotechnical Change: AI as Governance Rationale and Target*), I introduce the ‘Sociotechnical change’ framework to global AI governance questions. I first (4.1) provide a review and definition of the concept of ‘technology’, and explore how this relates to the theory of sociotechnical change. I then explore the two facets of this theory. I (4.2) discuss sociotechnical change as governance *rationale*, by discussing different types of societal change that can be produced by a technology, and the five distinct grounds on which these can produce a need for new regulation, and the epistemic limits of attempting to forecast or predict sociotechnical changes. I then (4.3) discuss sociotechnical change as governance *target*, considering the role of both material factors, as well as ‘problem logics’. To chart these logics, I (4.4) articulate a taxonomy of six types of problem logics (ethical challenges; security threats; safety risks; structural shifts; common benefits; governance disruption), and their corresponding governance rationales and regulatory logics. Finally, I (4.5) review the analytical uses and limits of this model for AI governance. Finally, (4.6) I conclude.

In Chapter 5 (*Technology-Driven Legal Change: AI as Governance Disruptor*), I explore how and why the use of AI capabilities might produce intended or unintended change in the global legal order or governance system itself. This discussion proceeds as follows: after defending the importance of exploring governance disruption at the global level, I (5.1) provide an abridged review of the historical connection between technological innovation and global legal change, and (5.2) sketch the governance disruption framework. I then explore the three varieties of governance disruption. Under (5.3) *Development*, the sociotechnical change produced by AI capabilities results in situations of substantive legal uncertainty or ambiguity. Their use (a) creates new legal

gaps, (b) reveals uncertainties or ambiguities in existing laws, (c) blurs the scope of application of regimes, (d) renders obsolete core justifying assumptions, or (e) rebalances the sociotechnical problem landscape facing the regime, possibly beyond extant institutional competencies. Any of these situations create a need for change in governance in order to resolve these tensions, and to ensure the international legal system can adequately address these challenges. Under (5.4) *Displacement*, AI technologies could be used to support—or to substitute for—key processes or practices of the global legal order. They could see use in the (partial or even full) automation of various tasks of international rule creation or adjudication; in support of compliance monitoring; or in the wholesale shifting of ‘modalities’ of international relations. Finally, in cases of (5.5) *Destruction*, AI affects and erodes the effectiveness and coherence of the global legal order, either because existing regimes or instruments prove conceptually or politically incapable of carrying through certain urgent developments, or because AI applications directly or indirectly erode the political scaffolding of international regimes. Finally, (5.6) I conclude.

In Chapter 6 (*Change in Governance Architecture: Regime Complexity and AI*), I turn to the question of overarching changes in the institutional ecologies and governance architectures within which AI will be embedded. The regime complexity lens focuses on patterns and drivers of institutional and normative change in the broader institutional ecologies surrounding AI governance regimes, as well as interactions within the emerging AI governance architecture. I first (6.1.) provide a brief background on the development of regime theory, and (6.2) the articulation of the concept of a ‘regime complex’. I then (6.3) examine some of the debates around the consequences and desirability of fragmentation or centralisation in a regime complex. Finally, I (6.4) explore how the regime complexity lens allows one to explore a governance regime at five levels—origin, topology, evolution, consequences, and strategies. This will lay out the set of questions that structure the analysis in Part III.

In Part III (*Frameworks for Change*), I aim to show how the three lenses can shed light on particular aspects or issues in the emerging AI governance regime complex. Throughout the sub-sections of Chapter 7 (*AI Governance in Five Parts*), the sociotechnical change, governance disruption, and regime complexity lenses are each used to reframe various conceptual, strategic, and political questions or choices in AI governance, in order to respond or reckon better with these facets of change.

I show how these lenses can help offer insight into five questions that can be asked of diverse existing, emerging, or proposed AI governance systems in distinct domains. In terms of (7.1) *origins*, these lenses allow consideration of an AI regime’s purpose, viability, and design considerations. In terms of (7.2) *topology*, they allow an examination of the state and degree of normative or institutional fragmentation of an AI governance architecture at a certain time. In terms of (7.3) *evolution*, these lenses permit the examination of potential trajectories of the AI regime complex, via consideration of how external governance trends as well as AI-driven patterns of governance disruption can serve as drivers of regime complex integration or fragmentation. In terms of (7.4) *consequences*, these lenses enable us to consider the political, institutional and normative challenges and trade-offs that the AI governance architecture may face, conditional on whether it remains fragmented, or whether it is integrated. Finally, in terms of (7.5) *strategies*, these three lenses highlight potential shifts in conceptual approach, instrument

choice, or instrument design, which can help ensure the efficacy, resilience, and coherence of the AI regime complex in the face of change.

Finally, in the *Conclusion*, I review the project, reflect on the strengths and limits of the frameworks presented and the arguments made, and work out implications of this both for practical questions of AI global governance, as well as for future research.

Part I: FOUNDATIONS OF CHANGE

Chapter 2. Global Governance for AI: stakes, challenges, state

This chapter provides a detailed background and rationale for this present contribution, by discussing developments in AI technology, its challenges, and gaps in the global governance landscape. I first (2.1) investigate the ‘stakes’ of AI. I briefly survey three ways of defining AI technology, before turning to an examination of the importance of AI technology, taking stock of both achievements as well as important limits. It is argued that rather than categorically effective or categorically over-hyped, AI is instead a high-variance technology portfolio that comprises many techniques and use cases that are limited, but where a subset of peak applications can nonetheless enable considerable impact. I also explore the changing near-term importance of AI, by evaluating arguments around potential future progress or dissemination rates, to suggest that even under conservative assumptions, AI’s impact on the world the coming years will be far-reaching. I then (2.2) survey the range and diversity of AI’s global challenges; argue that some of these will need forms of global cooperation or coordination, and that while such governance faces barriers, there are levers that can make it enforceable, providing that global agreements are reached. Finally, I (2.3) review existing international law approaches to governing AI, as well as developments in AI ethics and AI governance, to argue that there remain gaps, and that new solutions or approaches may be needed.

2.1 Stakes: AI matters and will drive change

Recent years have seen extraordinary attention for AI technology amongst industry, regulators, scientists, and the general public. There is a growing perception that the technology will give rise to a wide range of new challenges that require some form of governance or regulation. Yet what is AI? How important will it be, really? What kind of change will it create? To fully appreciate the trends, developments, and stakes around AI development, we must first briefly understand what the technology is.

2.1.1 Defining AI Definitions: Three Purposes

What is AI? This question is deceptively simple to ask, but remarkably difficult to answer. To date, even AI researchers themselves still disagree over the precise definition of both ‘intelligence’ and ‘AI’.¹ This conceptual ambiguity has also contributed to challenges in public debate.

¹ STUART RUSSELL & PETER NORVIG, *ARTIFICIAL INTELLIGENCE: A MODERN APPROACH* (3rd ed. 2016); See also the collection of (70) definitions of “intelligence”, in Shane Legg & Marcus Hutter, *A Collection of Definitions of Intelligence*, ARXIV07063639 Cs (2007), <http://arxiv.org/abs/0706.3639> (last visited Jan 28, 2017). For a recent discussion and debate, see also Dagmar Monett et al., *Special Issue “On Defining Artificial Intelligence”—Commentaries and Author’s Response*, 11 J. ARTIF. GEN. INTELL. 1–100 (2020).

This should not be surprising: as a discursive topic, ‘artificial intelligence’ combines a number of aspects that make it a perfect storm for conceptual confusion. Firstly, and in contrast to, say ‘nuclear engineering’, the very phrase ‘artificial intelligence’ evokes a term—‘intelligence’—which is in widespread and everyday use, and which for many people has strong (evaluative or normative) connotations. Yet the term is a suitcase word that packages together many competing meanings,² even while it hides deep and perhaps even intractable scientific and philosophical disagreement.³ Secondly, and in contrast to, say, ‘blockchain ledgers’, AI technology has the baggage of decades of depictions in popular culture, resulting in a whole genre of tropes that can colour public perceptions and policymaker debates. Thirdly, and in contrast to, say, ‘quantum computing’, AI is an evocative general-purpose technology that sees use in a wide variety of domains, and accordingly has provoked commentary from virtually every disciplinary angle, including neuroscience, philosophy, psychology, law, politics, and ethics.

As a result of this, a persistent challenge in work and discourse on AI governance—and indeed, in the broader public debates around AI—has been that different people use the word ‘AI’ to refer to widely different artefacts, practices or systems, or operate on the basis of definitions or understandings which package together a range of assumptions.⁴ As such, it is of relevance to reflect briefly on the diverse ways AI is theorised, why this matters, and how this project consequently approaches AI technology.

Roughly speaking, we might distinguish at least three goals or rationales for defining AI:⁵ as AI practitioner or scientist may seek to (1) define AI from the ‘inside’, as a *science*; scholars in other fields may seek to (2) understand AI from the ‘outside’ as a *sociotechnical system in practice*; regulators may seek to (3) pragmatically delineate AI definition as *handles for law*, regulation or governance.⁶

² Marvin Minsky (on terms like ‘consciousness’, ‘learning’ or ‘memory’). Marvin Minsky, *Consciousness is a Big Suitcase* (1998), https://www.edge.org/conversation/marvin_minsky-consciousness-is-a-big-suitcase (last visited Sep 10, 2020).

³ In this way, ‘intelligence’ might be considered an ‘essentially contested concept’. For the classic account of this concept, see W. B. Gallie, *Essentially Contested Concepts*, 56 PROC. ARISTOT. SOC. 167–198 (1955).

⁴ “By far the greatest danger of Artificial Intelligence”, decision theorist Eliezer Yudkowsky has quipped, “is that people conclude too early that they understand it” Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk.*, in GLOBAL CATASTROPHIC RISKS 308–345, 308 (Nick Bostrom & Milan M. Cirkovic eds., 2008).

⁵ This is just one possible typology, and may be too narrow. More generally speaking, many perspectives or narratives around AI might be classified into 5 classes which define AI by different major ‘aspects’: the (1) *nature* of AI (‘what AI is’ – e.g. a field of science, information technology, robots, software, mind); (2) the *operation* of AI (‘how AI works’ – e.g. autonomous system, complex adaptive system, agent, optimizing process, ...); (3) our *relation to AI* (e.g. AI as tool, moral patient, moral agent, cultural object, ...); (4) AI’s *function* or use (e.g. companion, advisor, weapon, tool for misuse, labour enhancer or substitute; tool for power concentration or social control; tool for empowerment or resistance; ...); (5) AI’s *unintended impact* or side-effects (e.g. AI as gradual usurper of human decision-making authority; driver of societal value drifts; vector of eventual catastrophic risk). Each of these have sub-framings, and emphasize different features, societal problems, and policy responses. This is explored further in other unpublished draft work.

⁶ This anticipates some of our later discussion, in Chapter 4.3., on AI and sociotechnical change as ‘governance target’.

2.1.1.1 Defining AI as science

Firstly, definitions of AI as science hew close to the ways in which AI researchers see their own field. There are a range of influential definitions; for instance, Nilsson has defined the field of AI as being concerned with “making machines intelligent, [where] intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.”⁷ However, even amongst AI researchers, there is no set consensus on this definition,⁸ and there are many other scientific definitions in currency.⁹ Indeed, some AI researchers themselves have been happy to shelve definitional questions—and to ‘get on with it’—even from a scientific perspective.¹⁰ Nonetheless, a broad and encapsulating ‘scientific’ definition of AI is provided by William Rapaport;

“AI is a branch of computer science (CS), which is the scientific study of what problems can be solved, what tasks can be accomplished, and what features of the world can be understood computationally (i.e., using the language of Turing Machines), and then to provide algorithms to show how this can be done efficiently, practically, physically, and ethically.”¹¹

Definitions of AI as science provide an important baseline or touchstone for analysis in many other disciplines. Nonetheless, they are not fully sufficient or analytically enlightening to many fields of study dealing with the societal *consequences* of the technology’s application, or with avenues for governing these. As such, this project will take this definition as a starting point, but will be more concerned with this other types of definitions.

2.1.1.2 Defining AI as sociotechnical system in practice

Secondly, definitions of AI can be pursued with the goal of understanding the technology as a *sociotechnical system in practice*. Such definitions accordingly consider the functional inputs and effects of AI technology as a sociotechnical system.¹² Taking such an approach can aid scholars seeking to study the societal impacts and effects of the technology, in diagnosing or

⁷ NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE: A HISTORY OF IDEAS AND ACHIEVEMENTS xiii (2010).

⁸ For instance, in one survey of more than 70 definitions of ‘intelligence’, Legg and Hutter concluded that while it is hard to settle on a single ‘correct’ term, many of the most concise and precise definitions share a number of features: (1) intelligence is a property of some agent that interacts with an environment; (2) intelligence is generally indicative of that agent’s ability to succeed at a particular task or stated goal; (3) there is an emphasis on learning, adaptation, and flexibility within a wide range of environments and scenarios. Legg and Hutter, *supra* note 1 at 9. See also the more recent broad overview in Sankalp Bhatnagar et al., *Mapping Intelligence: Requirements and Possibilities*, in PHILOSOPHY AND THEORY OF ARTIFICIAL INTELLIGENCE 2017 117–135 (Vincent C. Müller ed., 2018).

⁹ For a classic taxonomy of ways in which AI researchers have defined ‘intelligence’—distinguishing between systems that achieve ‘thinking humanly’, ‘thinking rationally’, ‘acting humanly’, or ‘acting rationally’—see also the influential distinction offered in RUSSELL AND NORVIG, *supra* note 1 at 2. See also the versions of this in STEPHAN DE SPIEGELEIRE, MATTHIJS MAAS & TIM SWELJS, ARTIFICIAL INTELLIGENCE AND THE FUTURE OF DEFENSE: STRATEGIC IMPLICATIONS FOR A SMALL FORCE PROVIDER 29 (2017); KELLEY M SAYLER & DANIEL S. HOADLEY, *Artificial Intelligence and National Security* 43 3 (2020).

¹⁰ PETER STONE ET AL., *Artificial Intelligence and Life in 2030* 12 (2016), <http://ai100.stanford.edu/2016-report> (last visited Feb 26, 2017).

¹¹ William J. Rapaport, *What Is Artificial Intelligence?*, 11 J. ARTIF. GEN. INTELL. 52–56, 54 (2020).

¹² Various concepts of ‘technology’ and ‘sociotechnical systems’ will be discussed in more detail in Chapter 4.1.

analysing the potential dynamics of AI development, diffusion, and application, as well as the socio-political problems and opportunities these trends create.

This is the type of definition that informs my approach in this project. I functionally understand AI as a sociotechnical system in practice, which consists of four linked layers:

- (a). a varied suite of computational *techniques*¹³ which can improve the accuracy, speed, or scale of machine decision-making across diverse information-processing or decision-making contexts; yielding...
- (b) a set of *capabilities*¹⁴ which can be used to support, substitute for-, and/or improve upon human performance in diverse tasks; which in turn enables...
- (c) a spectrum of useful *applications*¹⁵ across diverse domains; which in turn produce or support...
- (d) diverse forms of *new behaviour* resulting in sociotechnical changes.¹⁶

¹³ ‘Techniques’ encompass a variety of paradigms and approaches, such as symbolic approaches; supervised learning, unsupervised learning; reinforcement learning; Generative Adversarial Networks (GANs), as well as other approaches including ‘evolutionary programming’, ‘causal reasoning’, ‘Bayesian inference’, etc. PAUL SCHARRE & MICHAEL C HOROWITZ, *Artificial Intelligence: What Every Policymaker Needs to Know* 23 (2018), <https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know>. There are, naturally, other ways of distinguishing between these. At a high level, Pedro Domingos has identified 5 ‘tribes’ of AI. PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD (1 edition ed. 2015). In another recent mapping, Hernandez-Orallo and colleagues identify 14 categories of AI techniques, with distinct subcategories and techniques. These include: cognitive approaches; declarative machine learning, evolutionary & nature-inspired methods; general machine learning; heuristics & combinatorial optimization; information retrieval; knowledge representation and reasoning; multiagent systems & game theory; natural language processing; neural networks; parametric machine learning; planning & scheduling; probabilistic & Bayesian approaches; reinforcement learning & MDPs. Jose Hernandez-Orallo et al., *AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues* 8, 3 (2020), http://ecai2020.eu/papers/1364_paper.pdf. I also thank James Fox for discussing and clarifying some of these approaches.

¹⁴ These ‘Capabilities’ include a range of (presently) narrow but domain-agnostic functions that are of use in diverse contexts. They include (i) *data classification* (e.g. detecting if a skin photo shows cancer or not; facial recognition); (ii) *data generation* (e.g. synthesizing ‘DeepFake’ images, video, audio or other media); (iii) anomaly or pattern detection (e.g. detecting cybercrime or fraudulent financial transactions); (iv) *prediction* (e.g. of consumer preferences; weather patterns; loan performance; judicial rulings); (v) *optimization* of complex systems and tasks (e.g. optimizing energy usage in data centers); or (vi) *autonomous operation* of cyber-physical platforms (e.g. robots such as Roomba vacuum cleaners, social robots, or military drones). See also SCHARRE AND HOROWITZ, *supra* note 13. Naturally, there are other ways to segment these functions. See for instance Francesco Corea, *AI Knowledge Map: how to classify AI technologies*, MEDIUM (2020), https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020 (last visited Jul 8, 2020). See also the discussion of AI capabilities’ broad usability in section 2.1.7.2.

¹⁵ ‘Applications’ are an extremely diverse class of use cases. It is where AI *Techniques* that enable certain *Capabilities* refract into the full range of specific ‘useful’ tasks that can be carried out for different principals: from email spam detection to facial recognition, from self-driving cars to criminal re-offense predictions; from chatbots to DeepFakes, and from cybersecurity to Lethal Autonomous Weapons Systems. In a legal context, this category underpins what Schuett discusses as a ‘use case’-centric approach. Jonas Schuett, *A Legal Definition of AI*, ARXIV190901095 Cs, 5–6 (2019), <http://arxiv.org/abs/1909.01095> (last visited Jan 6, 2020). Petit calls this the ‘technical’ approach (contrasted to the ‘legalistic’ approach, which departs from the specific effects of pre-sorted legal systems and doctrine. NICOLAS PETIT, *Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications* 2 (2017), <https://papers.ssrn.com/abstract=2931339> (last visited May 11, 2020).

¹⁶ The layer of ‘sociotechnical change’ involves a very wide array of situations—Involving new behaviours, ways of being, or relations—which will be explored at length in Chapter 4.

I suggest this definition of AI as sociotechnical system in practice is analytically useful, because it helps draw the connections between the levels of AI technological approaches and artefacts, to systems, applications and societal impacts out in the world. In this way, this definition helps appreciate the diversity of AI techniques, the versatility of its capabilities, and the potential ubiquity and importance of its applications. It will be this definition that will underpin or inform much of the analysis of AI's effects as driver or enabler of change.

2.1.1.3 Defining AI as a handle or target for governance

Finally, a third goal in defining AI is to define it for the purposes of *governance*. How should regulators or policymakers understand or define AI? How should legal texts or treaties refer to it? Certainly, such definitions need to take stock of both other types of AI definitions, both 'insider' definitions of AI as scientific field, as well as analytical perspectives of AI as sociotechnical system in practice. Nonetheless, when choosing working definitions of technology that are fit for (and in) law and regulation, we may and must also consider additional practical features.

One key question at this point is whether we even need a single definition of AI for law and regulation? Some suggest we should. For instance, Jacob Turner has argued that a single workable definition for AI would be necessary to ensure that a legal system uses specific and workable definitions when describing which conduct around and of AI is or is not permitted.¹⁷ Without a baseline definition, the concern is that we risk being over- or under-inclusive in our laws.¹⁸

Others have argued that we do not need such a single legal definition, because AI—as highlighted by the second type of definition—is 'not a single thing' or a specific technology, but rather a disparate set of techniques, which cannot be cohesively regulated.¹⁹ Indeed, while overarching definitions can be valuable in anchoring regimes, Jonas Schuett has articulated a compelling critique of pursuing a single legal definition of 'AI' for use across all regulation, and instead favours a more pluralistic and pragmatic approach.²⁰ In doing so, he first derives a set of general requirements for (any) legal definitions. These are: 'inclusiveness', 'precision', 'comprehensiveness', 'practicability', and 'permanence'.²¹ Schuett argues that the umbrella term 'AI' does not meet many of these criteria, and that it is broadly over- or under-inclusive in most cases. Instead, he recommends that regulators use a risk-based approach, in which they:

¹⁷ JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE 8–9 (2018). He also argues such definitions are necessary for laws to meet Fuller's Principles of Legality, as set out in LON L. FULLER, THE MORALITY OF LAW (1969).

¹⁸ Accordingly, Turner settles on a definition of 'Artificial Intelligence' as "the Ability of a Non-natural Entity to Make Choices by an Evaluative Process." TURNER, *supra* note 17 at 16.

¹⁹ Urs Gasser & Virgilio A.F. Almeida, *A Layered Model for AI Governance*, 21 IEEE INTERNET COMPUT. 58–62 (2017).

²⁰ Schuett, *supra* note 15.

²¹ *Id.* at 2. Schuett notes that some of these requirements are legally binding in at least some jurisdictions (for instance, 'inclusiveness' can be derived from the principle of proportionality in EU law; 'precision' and 'comprehensiveness' are based on the EU law principle of legal certainty, or the vagueness doctrine in U.S. law), others are good legislative practice. *Id.* at 2.

“(1) [...] decide which specific risk they want to address, (2) identify which property of the system is responsible for that risk and (3) precisely define that property. In other words, the starting point should be the underlying risk, not the term AI.”²²

He argues that, rather than tailor regulation around ‘AI’, governance aimed at mitigating specific risks from AI might instead be centred around any one or a combination of the following properties of AI systems: their ‘design’ (e.g. reinforcement learning; supervised learning); their ‘use case’ (e.g. self-driving cars; facial recognition; medical diagnosis), or their ‘capability’ (what the system can do—e.g. physical interaction; automated decision-making without human intervention, or creating legally relevant effect on a person’s legal status or legal rights.).²³ In this argument, specific laws or regulations can be framed around one of these elements, or around combinations of them.²⁴ Schuett’s framework is valuable, because it pursues a ‘risk-based approach’, rather than one that is focused *prima facie* on ‘AI’ itself. This therefore can put the emphasis more on patterns of sociotechnical change.

Schuett’s risk-focused approach to ‘generating governance definitions’ of AI is effective at covering many AI risks, whether or not the overarching governance regime is focused on AI applications across diverse sectors, or on specific narrowly constrained domains. Ultimately, this project will only indirectly touch upon the question of which specific AI definitions governance regimes should adopt in legal texts. That is, while it adopts a fairly detailed definition of AI as a sociotechnical system in practice (involving techniques, capabilities, applications, and sociotechnical changes), it leaves open the question of which specific definitions regulators should use.²⁵ Instead, it will address additional considerations around the nature of the governance rationale created by AI applications, as well as the ‘texture’ of the AI as governance target, which can inform more granular considerations as whether to focus regimes or treaty texts on AI ‘design’ (*techniques*), ‘use cases’ (*applications*), or ‘capabilities’.

2.1.2 Does AI actually matter? Reasons for caution

Having briefly sketched these three definitions of AI provides the background necessary to address a second and more salient question: how important is AI? Does it ‘matter’—in the sense that this technology should be the subject of particular or even unusual political and regulatory attention?

It is important to raise this question, because a technology’s transformative societal impact, novelty, or legal ‘exceptionality’ can at times be too easily taken for granted.²⁶ Indeed, as

²² Schuett, *supra* note 15 at 4.

²³ *Id.* at 7–8. It should be noted that these categories of ‘design’, ‘use case’, and ‘capability’ correspond somewhat to the categories of ‘technique’, ‘application’, and ‘capability’, respectively, although they do not fully overlap.

²⁴ Schuett does argue for multi-element definitions—for example, “This regulation applies to reinforcement learning agents which can physically interact with their environment and make automated decisions.” (i.e. a regulation scoped by use case and two capabilities). *Id.* at 7–8.

²⁵ This is also discussed in Chapter 4.5.1 and 4.6, where it is argued that the ‘sociotechnical change’ lens is not meant to provide blueprints for regulation, but simply a thinking tool for approaching and framing governance initiatives in the first place.

²⁶ On the limits of the ‘exceptionalist’ legal approach, see also Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. LAW REV. CIRCUIT 17 (2015); Rebecca Crootof & B. J. Ard, *Structuring Techlaw*, 34 HARV. J. LAW TECHNOL. (2021), <https://papers.ssrn.com/abstract=3664124> (last visited Aug 28, 2020).

AI sceptics often recount, the field of AI has been subject to multiple previous cycles of hype and disillusionment throughout its 70-year history.²⁷ These latter periods resulted in so-called ‘AI winters’: periods, in the 1970s and 1990s, when once-lavish research funding dried up, after initially high expectations were not met.²⁸

Could the same be true today? Some have argued that the core breakthroughs necessary for a sustained ‘AI revolution’ have not yet occurred.²⁹ Is recent AI progress ‘for real’, or will this all result in another bust and a new winter?³⁰ Even if progress is real, will it translate into widespread changes for society? Even if it does, why would the regulation of the resulting problems pose a particularly new challenge for existing approaches or regimes in international law or global governance?

It is important to be cautious about the prospects of AI disruption. After all, while legal scholarship and action has at times been held to ‘lag behind’ technological innovation,³¹ legal scholars have also on occasion jumped the gun on attempting to regulate new technologies which they did not yet fully grasp. For instance, Gregory Mandel has discussed how legal scholars and courts alike have often been too easily ‘wowed’ by new technologies, at times being too eager to admit into court fingerprinting or DNA identification techniques when these were still in a premature stage of development.³² At the international legal level, Colin Picker has charted how during the 1960’s and 1970’s legal scholars jumped the gun on regulating then-imminently-anticipated advances in deep terrestrial mining, weather control technology or deep seabed mining, of which the latter took many more decades to realize, and the former remain out of reach

²⁷ For histories, see PAMELA MCCORDUCK, MACHINES WHO THINK: A PERSONAL INQUIRY INTO THE HISTORY AND PROSPECTS OF ARTIFICIAL INTELLIGENCE (25th anniversary update ed. 2004); DANIEL CREVIER, AI: THE TUMULTUOUS SEARCH FOR ARTIFICIAL INTELLIGENCE (1993); NILSSON, *supra* note 7.

²⁸ RUSSELL AND NORVIG, *supra* note 1 at 24–28. See also the study on the DARPA strategic computing initiative; ALEX ROLAND & PHILIP SHIMAN, STRATEGIC COMPUTING: DARPA AND THE QUEST FOR MACHINE INTELLIGENCE, 1983–1993 (2002). See also the brief historical review in DE SPIEGELEIRE, MAAS, AND SWEIJS, *supra* note 9 at 31–34.

²⁹ See also Michael I. Jordan, *Artificial Intelligence—The Revolution Hasn’t Happened Yet*, 1 HARV. DATA SCI. REV. (2019), <https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/8> (last visited Jun 23, 2020). Though see also some replies, Emmanuel Candès, John Duchi & Chiara Sabatti, *Comments on Michael Jordan’s Essay “The AI Revolution Hasn’t Happened Yet”*, 1 HARV. DATA SCI. REV. (2019), <https://hdsr.mitpress.mit.edu/pub/djb16hzl/release/4> (last visited Jun 23, 2020).

³⁰ See for instance Luciano Floridi, *AI and Its New Winter: from Myths to Realities*, 33 PHILOS. TECHNOL. 1–3 (2020). See also Will Knight, *Facebook’s Head of AI Says the Field Will Soon ‘Hit the Wall’*, WIRED, 2019, <https://www.wired.com/story/facebook-ai-says-field-hit-wall/> (last visited Aug 17, 2020).

³¹ For an influential discussion of this regulatory ‘pacing problem’, see Gary E. Marchant, *The Growing Gap Between Emerging Technologies and the Law*, in THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM 19–33 (Gary E. Marchant, Braden R. Allenby, & Joseph R. Herkert eds., 2011), https://doi.org/10.1007/978-94-007-1356-7_2 (last visited Jan 28, 2019). This assumption is shared by many, see also Ryan Hagemann, Jennifer Huddleston & Adam D. Thierer, *Soft Law for Hard Problems: The Governance of Emerging Technologies in an Uncertain Future*, 17 COLO. TECHNOL. LAW J. 94 (2018). Others, however, argue that this commonly alleged gap between technology and law may be overstated. Cf. Lyria Bennett Moses, *Agents of Change: How the Law ‘Copes’ with Technological Change*, 20 GRIFFITH LAW REV. 763–794, 763–764 (2011) (discussing and critiquing the metaphor of the “hare and the tortoise”, common in debates around the regulation of technology, arguing that it too quickly assumes that law will always be the loser in the race against technology). See also See also Crootof and Ard, *supra* note 26 at 14. (“The facile but persistent claim that “law cannot keep up with new technologies” ignores the myriad ways in which law shapes most technological developments.”).

³² Gregory N. Mandel, *Legal Evolution in Response to Technological Change*, OXF. HANDB. LAW REGUL. TECHNOL., 235–239 (2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-45> (last visited Sep 26, 2018).

even today.³³ Such cases recall the warnings of Judge Frank Easterbrook, who in a now-famous critique of the early ‘cyberlaw’ project cautioned that;

“[w]e are at risk of multidisciplinary dilettantism, or, as one of my mentors called it, the cross-sterilization of ideas. Put together two fields about which you know little and get the worst of both worlds. [...] Beliefs lawyers hold about computers, and predictions they make about new technology, are highly likely to be false. This should make us hesitate to prescribe legal adaptations for cyberspace. The blind are not good trailblazers.”³⁴

Nonetheless, that does not mean that investigation of the implications of emerging technologies is not warranted—but rather that we should be cautious. As with any emerging technology, there remains considerable confusion and uncertainty around AI. It is necessary, then, to address some of these uncertainties and to explore and demystify the state of the art in relative detail and nuance. The importance of AI should not be assumed, but critically examined; the point may be less about whether AI does or does not matter than specifically about when, how, and why it matters.

As such, the following sections will explore the debate over how AI may matter—and whether, or what kind of change we might expect. I will review both the (2.1.3.) advances and (2.1.4.) limits of current AI approaches and applications. Reviewing these debates, I (2.1.5) suggest that rather than categorical optimism or pessimism, it is best to conceive of AI as a high-variance technology portfolio. I then discuss how AI could drive increasing change through either of two paths. I will (2.1.6) discuss the potential of ‘*change from future progress*’, by offering an outside- and inside-view analysis of the uncertainties and debates around potential future AI advances, ceilings, or capability trajectories. I then discuss (2.1.7) the complementary scenario of ‘*change from algorithmic proliferation*’, reviewing models of technology dissemination and proliferation, I argue that even if underlying AI progress would slow considerably, the rollout and deployment of existing capabilities to ever more applications will likely more than suffice to make the technology ubiquitous and highly societally disruptive, even in the near term. On this basis, I argue, AI will matter and create considerable change and challenges for governance.

2.1.3 Progress and promise: a lasting AI Summer?

The last decade has seen no shortage of excitement around AI’s promise. Of course, computing and robotic technologies—even the very discipline of AI—are hardly new. The very term ‘artificial intelligence’ dates all the way back to the 1956 Dartmouth Conference.³⁵ Likewise, although the field of AI went through funding ‘winters’ after initial high expectations were not met, it is also untrue to suggest that earlier AI technologies entirely failed to see

³³ C. Wilfred Jenks, *The New Science and the Law of Nations*, 17 INT. COMP. LAW Q. 327–345, 338–339 (1968) (“we may... within the next few years need a Sonic Boom Treaty, a Center of the Earth Treaty, a Cybernetics Treaty and a Molecular Biology Treaty.”); As quoted in Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDOZO LAW REV. 151–219, 194 (2001). On international regulation for weather control, see also Edith Brown Weiss, *International responses to weather modification*, 29 INT. ORGAN. 805–826 (1975).

³⁴ Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. LAW REV. 501, 207 (1999).

³⁵ J. McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (1955), <http://robotics.cs.tamu.edu/dshell/cs625/2014-09-02.pdf>.

successes or achieve adoption in society. Indeed, amongst others, rule-based ‘expert systems’ have been in use in diverse sectors for decades.³⁶ In many cases, this highlights the role of the so-called ‘AI effect’—the problem whereby, as John McCarthy famously lamented, “[a]s soon as it works, no one calls it AI anymore”.³⁷ As a result, it has often been the case that successful advances and applications of AI soon were considered mere software, leaving the ultimate goal ever receding—and leaving AI researchers to appear to deal in perpetual ‘failures’.³⁸

Nonetheless, whatever the field’s prior history, there can be no doubt that over the last decade the field of AI has undergone rapid and sustained progress, particularly in a subfield of machine learning called ‘deep learning’.³⁹ This success has been driven by concurrent improvements in algorithms, steady advances in computing power, and a sharply increasing ability to capture, store, and bring to bear large amounts of training data from diverse contexts.⁴⁰ These three primary drivers have been further complemented and enabled by broader advances, including, as some argue, “the decades-long accumulation of software in general as a cultural-technological heritage; and [...] the rapid cost decreases and hence widespread availability of complementary technologies, such as ubiquitous connectivity, ubiquitous computing, and the internet of things (IoT).”⁴¹

As a result, the field of AI has advanced at an extraordinary—if not uniform—pace, with algorithms gradually managing to meet or mimic human performance across diverse tasks. Progress and breakthroughs in the use of deep learning have been particularly pronounced in the

³⁶ GREG ALLEN, *Understanding AI Technology* 20 (2020).

³⁷ Attributed in Moshe Y. Vardi, *Artificial Intelligence: Past and Future*, 55 COMMUN. ACM 5 (2012).

³⁸ MCCORDUCK, *supra* note 27 at 423. (describing the “odd paradox” that “practical AI successes, computational programs that actually achieved intelligent behaviour, were soon assimilated into whatever application domain they were found to be useful in, and became silent partners alongside other problem-solving approaches, which left AI researchers to deal only with the “failures”, the tough nuts that couldn’t yet be cracked”).

³⁹ For overviews of recent trends, see YOAV SHOHAM ET AL., *AI Index 2018 Annual Report* (2018), <http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf> (last visited Dec 17, 2018); RAYMOND PERRAULT ET AL., *The AI Index 2019 Annual Report* (2019), https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf. For attempts to track performance metrics, see Peter Eckersley & Yomna Nasser, *EFF AI Progress Measurement Project* (2017-), ELECTRONIC FRONTIER FOUNDATION (2017), <https://www.eff.org/ai/metrics> (last visited Jun 22, 2020).

⁴⁰ ALEX CAMPOLO ET AL., *AI Now 2017 Report* 3 (2017), https://assets.contentful.com/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf (last visited Oct 20, 2017). Note, these three factors are closely intertwined, and their relative importance may depend on which area or application of AI is under consideration. For instance, for areas such as reinforcement learning in simulation, only self-generated data may matter. I thank Miles Brundage for this point.

⁴¹ Claudio Feijóo et al., *Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy*, TELECOMMUN. POLICY 101988, 2 (2020).

areas of image recognition,⁴² various games,⁴³ and natural language processing.⁴⁴ However, progress has extended beyond this to diverse applications. To provide an incomplete list: AI systems have proven that they can match human performance in image and object recognition,⁴⁵ speech transcription, and prosaic translation in major languages;⁴⁶ AI systems have shown they can drive vehicles,⁴⁷ parse paragraphs to answer questions,⁴⁸ create new encryption schemes and detect malware,⁴⁹ or find and correct errors in legal contracts with similar accuracy to professional

⁴² Notably beginning in late 2012, after its use in the ImageNet competition. Alex Krizhevsky, Ilya Sutskever & Geoffrey E Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 25 1097–1105 (F. Pereira et al. eds., 2012), <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf> (last visited Jun 8, 2020).

⁴³ Including: large numbers of Atari games; Volodymyr Mnih et al., *Playing Atari with Deep Reinforcement Learning*, ARXIV13125602 Cs 9 (2013). The board game of Go; David Silver et al., *Mastering the game of Go with deep neural networks and tree search*, 529 NATURE 484–489 (2016); David Silver et al., *Mastering the game of Go without human knowledge*, 550 NATURE nature24270 (2017); David Silver et al., *A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play*, 362 SCIENCE 1140–1144 (2018) (expanding the same algorithm to chess and shogi). The real-time strategy videogame ‘Starcraft’; DeepMind, *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*, DEEPMIND (2019), <https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/> (last visited Jan 25, 2019).

⁴⁴ Alec Radford et al., *Language Models are Unsupervised Multitask Learners* 24 (2019); Tom B. Brown et al., *Language Models are Few-Shot Learners*, ARXIV200514165 Cs (2020), <http://arxiv.org/abs/2005.14165> (last visited May 29, 2020).

⁴⁵ Allison Linn, *Microsoft researchers win ImageNet computer vision challenge*, NEXT AT MICROSOFT (2015), <https://blogs.microsoft.com/next/2015/12/10/microsoft-researchers-win-imagenet-computer-vision-challenge/> (last visited Feb 25, 2017). For research on object recognition algorithms outperforming humans in standard tests, see Kaiming He et al., *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, ARXIV150201852 Cs (2015), <http://arxiv.org/abs/1502.01852> (last visited Jun 22, 2020).

⁴⁶ Wayne Xiong et al., *Achieving human parity in conversational speech recognition*, ARXIV PREPR. ARXIV161005256 (2016), <https://arxiv.org/abs/1610.05256> (last visited Feb 25, 2017); Gideon Lewis-kraus, *The Great A.I. Awakening*, THE NEW YORK TIMES, December 14, 2016, <https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html> (last visited Feb 26, 2017); Yonghui Wu et al., *Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, ARXIV160908144 Cs (2016), <http://arxiv.org/abs/1609.08144> (last visited Jan 20, 2020). Note, while such systems may have matched or exceeded average human performance at translating, there are still critics who argue that such systems make literary blunders or miss subtleties, and on this ground argue that they are not ‘real’ translations. See Douglas Hofstadter, *The Shallowness of Google Translate*, THE ATLANTIC, 2018, <https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/> (last visited Jun 22, 2020). I thank Gavin Leech for clarifying some of the debates on this point.

⁴⁷ Ross Bryant, *Google’s AI becomes first non-human to qualify as a driver*, DEZEEN (2016), <https://www.dezeen.com/2016/02/12/google-self-driving-car-artificial-intelligence-system-recognised-as-driver-usa/> (last visited Feb 25, 2017). These autonomous vehicles tend to display lower accident rates than human drivers; Eric R. Teoh & David G. Kidd, *Rage against the machine? Google’s self-driving cars versus human drivers*, 63 J. SAFETY RES. 57–60 (2017). At the same time, recent years have seen new challenges and limits to autonomous vehicle technology. See P. A. Hancock, *Some pitfalls in the promises of automated and autonomous vehicles*, 62 ERGONOMICS 479–495 (2019).

⁴⁸ Alec Radford et al., *Better Language Models and Their Implications*, OPENAI BLOG (2019), <https://blog.openai.com/better-language-models/> (last visited Feb 17, 2019).

⁴⁹ Martin Abadi & David G. Andersen, *Learning to Protect Communications with Adversarial Neural Cryptography* (2016), <https://arxiv.org/pdf/1610.06918v1.pdf>; Linda Musthaler, *How to use deep learning AI to detect and prevent malware and APTs in real-time*, NETWORK WORLD (2016), <http://www.networkworld.com/article/3043202/security/how-to-use-deep-learning-ai-to-detect-and-prevent-malware-and-apts-in-real-time.html> (last visited Feb 25, 2017).

lawyers.⁵⁰ These systems have produced cookbooks and published poetry;⁵¹ they can generate pieces of ‘neural fake news’ that fool up to half of all readers,⁵² compose music and songs in various genres,⁵³ and generate (admittedly passable) jokes and puns.⁵⁴

Moreover, beyond merely matching or mimicking human performance, AI systems also critically outperform humans at some narrow tasks. Most visibly, algorithms have begun to comprehensively defeat human champions at a range of board games from chess to go,⁵⁵ and can perform at an internationally competitive level in complex real-time strategy videogames.⁵⁶ Remarkably, in some of these cases, the AI systems in question achieved dominion not by learning from human performance, but rather by discovering and deploying their own strategies—strategies which in some cases are so inscrutable and ‘alien’ that they would never pass for ‘human’, but which outperform human players nonetheless.⁵⁷ Moreover, such ‘un-human’, ‘narrow but high’ performance extends beyond cyberspace and game environments. Machine learning systems have been able to predict the structure of proteins;⁵⁸ can detect some skin cancers earlier

⁵⁰ LAWGEEX, *Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts* (2018), <https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf> (last visited Aug 27, 2020).

⁵¹ Hyacinth Mascarenhas, *Associated Press to expand its sports coverage by using AI to write Minor League Baseball articles*, INTERNATIONAL BUSINESS TIMES UK (2016), <http://www.ibtimes.co.uk/associated-press-expands-its-sports-coverage-by-using-ai-write-minor-league-baseball-articles-1568804> (last visited Feb 25, 2017); Alex Marshall, *From Jingles to Pop Hits, A.I. Is Music to Some Ears*, THE NEW YORK TIMES, January 22, 2017, <https://www.nytimes.com/2017/01/22/arts/music/jukedeck-artificial-intelligence-songwriting.html> (last visited Feb 18, 2017); Zackary Scholl, *Turing Test: Passed, using computer-generated poetry*, RASPBERRY PI AI (2015), <https://rpiai.wordpress.com/2015/01/24/turing-test-passed-using-computer-generated-poetry/> (last visited Feb 25, 2017); Alexandra Kleeman, *Cooking with Chef Watson, I.B.M.’s Artificial-Intelligence App*, THE NEW YORKER (2016), <http://www.newyorker.com/magazine/2016/11/28/cooking-with-chef-watson-ibms-artificial-intelligence-app> (last visited Feb 25, 2017).

⁵² Rowan Zellers et al., *Defending Against Neural Fake News*, ARXIV190512616 CS (2019), <http://arxiv.org/abs/1905.12616> (last visited Jun 3, 2019); Radford et al., *supra* note 48. Likewise, studies of OpenAI’s GPT-2 text generation AI found that humans found its output convincing, and that it could be tailored to produce propaganda in support of diverse ideological positions. Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models*, ARXIV190809203 CS (2019), <http://arxiv.org/abs/1908.09203> (last visited Nov 18, 2019).

⁵³ Prafulla Dhariwal et al., *Jukebox: A Generative Model for Music*, ARXIV200500341 CS EESS STAT (2020), <http://arxiv.org/abs/2005.00341> (last visited May 18, 2020). In another application, a system generated classical music in the style of Bach, which fooled around half of listeners. Gaëtan Hadjeres, François Pachet & Frank Nielsen, *DeepBach: a Steerable Model for Bach Chorales Generation*, ARXIV161201010 CS (2016), <http://arxiv.org/abs/1612.01010> (last visited May 21, 2018).

⁵⁴ He He, Nanyun Peng & Percy Liang, *Pun Generation with Surprise*, ARXIV190406828 CS (2019), <http://arxiv.org/abs/1904.06828> (last visited May 4, 2019).

⁵⁵ Most notably AlphaGo and its successor algorithms, including AlphaZero: Silver et al., *supra* note 43; David Silver et al., *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*, ARXIV171201815 CS (2017), <http://arxiv.org/abs/1712.01815> (last visited Dec 6, 2017); Silver et al., *supra* note 43; Silver et al., *supra* note 43.

⁵⁶ DeepMind, *supra* note 43. OpenAI, *OpenAI Five*, OPENAI (2018), <https://openai.com/blog/openai-five/> (last visited Sep 3, 2020).

⁵⁷ Will Knight, *Alpha Zero’s “Alien” Chess Shows the Power, and the Peculiarity, of AI*, MIT TECHNOLOGY REVIEW, 2017, <https://www.technologyreview.com/s/609736/alpha-zeros-alien-chess-shows-the-power-and-the-peculiarity-of-ai/> (last visited Feb 12, 2019).

⁵⁸ R. Evans et al., *De novo structure prediction with deep-learning based scoring*, in THIRTEENTH CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION (ABSTRACTS) (2018), <https://deepmind.com/blog/alphafold/>. For commentary on how unexpected and shocking this breakthrough was to those in the field, see Mohammed AlQuraishi, *AlphaFold @ CASP13: “What just happened?”*, SOME THOUGHTS

and with lower error rates than human doctors;⁵⁹ can predict the outcomes of complex chemical reactions better than trained chemists;⁶⁰ and are able to predict earthquake aftershocks⁶¹ and pandemic outbreaks.⁶² This is a necessarily incomplete survey that will soon be out of date.

The point is simply that AI algorithms have come to pass a series of performance benchmarks that are either salient and eye-catching for symbolic reasons (e.g. AlphaGo), or ‘boring-but-functional’ or profitable for various principals. Accordingly, the technology is now being used across diverse domains—from finance to healthcare, policing to art, science to government.⁶³ Moreover, contrary to previous ‘futuristic technologies’ such as nuclear power or gene editing, which were cloistered in ‘far-away’ industrial stations or labs, many of us interact directly (if not always visibly) with simple AI systems in our daily lives—from online product recommendation algorithms to social media content management systems, and from translation services to cleaner robots.⁶⁴

As a result of such high-profile progress and use cases, the promise of AI has hardly gone unnoticed. Accordingly, over the past years, private-sector investment in AI has been considerable;⁶⁵ dozens of states have articulated national AI strategies,⁶⁶ and many have begun investing considerable sums in both research and application. Given these successes, many have perceived that we are now at long last in the midst of the ‘AI revolution’, and AI will soon transform almost every domain of human activity.

ON A MYSTERIOUS UNIVERSE (2018), <https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/> (last visited Dec 13, 2018)..

⁵⁹ For work on a system that was able to classify skin cancer, see Andre Esteva et al., *Dermatologist-level classification of skin cancer with deep neural networks*, 542 NATURE 115–118 (2017).; for a recent breakthrough of a system for breast cancer screening, see Scott Mayer McKinney et al., *International evaluation of an AI system for breast cancer screening*, 577 NATURE 89–94 (2020). However, it should be noted that there remains controversy and critique over the accuracy of such systems. See Adewole S. Adamson & H. Gilbert Welch, *Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard*, 381 N. ENGL. J. MED. 2285–2287 (2019).

⁶⁰ Philippe Schwaller et al., *Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction*, 5 ACS CENT. SCI. 1572–1583 (2019).

⁶¹ Phoebe M. R. DeVries et al., *Deep learning of aftershock patterns following large earthquakes*, 560 NATURE 632–634 (2018).

⁶² Eric Niiler, *An AI Epidemiologist Sent the First Warnings of the Wuhan Virus*, WIRED, 2020, <https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/> (last visited Jan 27, 2020). However, in the context of the COVID-19 pandemic, there have also been critiques of many rushed and misspecified use cases. See also Nathan Benach, *AI has disappointed on Covid*, FINANCIAL TIMES, September 20, 2020, <https://www.ft.com/content/0aaafc2de-f46d-4646-acfd-4ed7a7f6fea> (last visited Sep 26, 2020).

⁶³ For a discussion of trends in automation in diverse sectors and industries, see JOHN DANAHER, AUTOMATION AND UTOPIA: HUMAN FLOURISHING IN A WORLD WITHOUT WORK 1 (2019).

⁶⁴ In this way, it is arguably distinct from previous technologies that featured prominently in e.g. bioethics debates. For instance, human cloning saw extensive debate long before it affected (m)any people directly. Paul Root Wolpe, *We Have Met AI, and It Is Not Us*, 11 AJOB NEUROSCI. 75–76 (2020).

⁶⁵ In 2019, global private AI investment exceeded \$70 billion. PERRAULT ET AL., *supra* note 39 at 6.

⁶⁶ See for overviews: Jessica Cussins, *National and International AI Strategies*, FUTURE OF LIFE INSTITUTE (2020), <https://futureoflife.org/national-international-ai-strategies/> (last visited Jun 22, 2020).; TIM DUTTON, BRENT BARRON & GAGA BOSKOVIC, *Building an AI World: Report on National and Regional AI Strategies* 32 (2018), https://www.cifar.ca/docs/default-source/ai-society/buildinganaiworld_eng.pdf?sfvrsn=fb18d129_4.

2.1.4 Limits and problems: a coming AI autumn?

But are we then, at long last, in a sustained AI ‘summer’? While progress in AI is certainly remarkable and has exceeded expectations, this does not mean that the technology is without limits, shortfalls or caveats. As noted, AI has been subject to previous hype cycles. It is important to engage with sceptical perspectives, because for all its successes, there remain at present under-recognised limits to the technology, its application, or the associated scientific fields of study, all of which must be reckoned with. Specifically, we can distinguish between challenges in (2.1.4.1) industry and AI application; (2.1.4.2) general ‘input’ preconditions or thresholds; (2.1.4.3) underlying challenges in today’s AI approaches; and (2.1.4.4) problems in the scientific field of AI.

2.1.4.1 Challenges in industry and application

For one, it is important to recognize that many prospective and implemented AI use cases are surrounded by unreflexively high expectations in AI industry or amongst governments—expectations that are in some cases not warranted, or should at least be subject to caution.

In the first place, there is a *mis-branding problem*. The lack of a clear definition of ‘AI’ means that many products are presented and sold as ‘AI’ when in fact they are merely rebranded applications of well-established statistical techniques or decades-old algorithms, in order to afford these solutions a greater veneer of objectivity or modernity.⁶⁷ For instance, one 2019 survey found that 40% of European firms classified as an ‘AI start-up’ did not actually use AI technology in any material way in their business.⁶⁸ In other cases, alleged ‘AI’ products are even operated by humans, a practice which has been called ‘fauxtomation’.⁶⁹

In the second place, there is a *hidden labour challenge*. Even where an application does involve more sophisticated AI algorithms, the collation of the requisite training data for many of these approaches—especially supervised learning systems—often still relies on tacit human labour practices, such as poorly paid ‘clickwork’ and ‘ghost work’, or by scraping user information

⁶⁷ VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018). This seems to be an ironic reversal of the aforementioned ‘AI effect’, where things are now called ‘AI’ that have worked for years.

⁶⁸ Parmy Olson, *Nearly Half Of All ‘AI Startups’ Are Cashing In On Hype*, FORBES (2019), <https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/> (last visited Jan 21, 2020). Also MMC VENTURES, *The State of AI 2019: Divergence 90* (2019), <https://www.mmcentures.com/wp-content/uploads/2019/02/The-State-of-AI-2019-Divergence.pdf>. On the other hand, AI’s impact may be ‘high-variance’: a 2019 global survey suggests that most companies that did deploy AI had seen measurable benefits, but that only a few ‘high performer’ managed to scale up that impact. MCKINSEY & COMPANY, *Global AI Survey: AI proves its worth, but few scale impact* 11 (2019), <https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact>.

⁶⁹ Astra Taylor, *The Automation Charade*, LOGIC MAGAZINE, 2018, <https://logicmag.io/failure/the-automation-charade/> (last visited May 12, 2020). And in a specific sector, see Maaike Verbruggen, *AI & Military Procurement: What Computers Still Can’t Do*, WAR ON THE ROCKS (2020), <https://warontherocks.com/2020/05/ai-military-procurement-what-computers-still-cant-do/> (last visited May 12, 2020). See also Maureen Ebben, *Automation and Augmentation: Human Labor as Essential Complement to Machines*, in MAINTAINING SOCIAL WELL-BEING AND MEANINGFUL WORK IN A HIGHLY AUTOMATED JOB MARKET (2020), www.igi-global.com/chapter/automation-and-augmentation/252995 (last visited Jun 22, 2020).

from social media or dating websites.⁷⁰ Even in cases where datasets are available, cleaning, segmenting, and structuring the data still requires extensive human tinkering.⁷¹ That is not to say that these algorithms cannot be useful, but merely that some do still require detailed human input, which creates an ‘economies of scale’ investment threshold that only bigger principals might be able to cross.⁷² This challenge also illustrates the (more) limited degree to which some algorithmic decision-making can truly be considered ‘autonomous’ today.

In the third place, diverse applications of algorithms are vulnerable to a *garbage-in-garbage-out* problem. In such cases, algorithms do not perform as well as anticipated, or even misperform in dangerous ways. This can often (but not always) be because they have been trained on insufficient, limited or flawed datasets. For instance, the controversial ‘COMPAS’ algorithm for predicting recidivism has been critiqued as being no more accurate than predictions made by people with no criminal justice expertise.⁷³ Many stakeholders do not always fully appreciate the limitations of current machine learning (ML) approaches, or why their tasks are not suited for those techniques. In other cases, private companies frequently face challenges over obtaining sufficient well-structured data,⁷⁴ which may lead them to settle prematurely for deploying the first systems that appear to perform well on tests, creating significant risks that these systems display biases or errors once deployed to a messy real-world context. Such situations can therefore result in sub-par algorithmic performance which, while perhaps merely ‘disappointing’ or amusing in relatively innocuous contexts (e.g. book recommendation algorithms), can be positively alarming in sensitive contexts where the rights or welfare of large classes of people are at stake. In some cases, this problem even derives from fundamental flaws in the underlying data collection methodologies, as with predictive policing algorithms built or trained on data created or documented during periods of flawed, biased, or even unlawful policing, resulting in the use of ‘dirty data’.⁷⁵

In the fourth place, AI applications occasionally manifest a *snake oil* problem. Certain proposed uses of AI are categorically flawed, because they are based on the presumed existence of certain patterns or correlations that do not exist, or at least cannot be sufficiently predicted or validated. The problem here is that the very versatility of AI technology lends it all too easily to

⁷⁰ MARY L. GRAY & SIDDHARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS (2019). See also KATE CRAWFORD & VLADAN JOLER, *Anatomy of an AI System* (2018), <http://www.anatomyof.ai> (last visited Jun 22, 2020).

⁷¹ For a lucid discussion of the many steps in the training process of a supervised learning algorithm, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS LAW REV. 653 (2017).

⁷² I thank James Fox for this observation.

⁷³ Julia Dressel & Hany Farid, *The accuracy, fairness, and limits of predicting recidivism*, 4 SCI. ADV. eaao5580 (2018). However, others have countered that on other datasets, and under more realistic conditions where human respondents were not provided with immediate feedback on their accuracy, algorithms did outperform humans. Zhiyuan “Jerry” Lin et al., *The limits of human predictions of recidivism*, 6 SCI. ADV. eaaz0652 (2020).

⁷⁴ One review suggests that data preparation still represents over 80% of all time in most AI projects. COGNILYTICA, *Data Engineering, Preparation, and Labeling for AI 2019* (2019), <https://www.cognilytica.com/2019/03/06/report-data-engineering-preparation-and-labeling-for-ai-2019/> (last visited Jun 18, 2020).

⁷⁵ Rashida Richardson, Jason Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 192 94 NYU LAW REV. (2019), <https://papers.ssrn.com/abstract=3333423> (last visited Jun 6, 2020).

what Arvind Narayanan has called ‘AI snake oil’.⁷⁶ In such cases, the problem derives from the fact that AI is deployed on the basis of assumptions or theories that remain unproven or flawed—and where ML algorithms are being deployed without demonstrated predictive accuracy, or where the ‘black-box’ nature of algorithms occludes the fact that seeming accuracy in tests is derived not from the existence of the presumed connection, but on spurious other correlations.⁷⁷ This, many have argued, has been the case with AI systems that were promoted as being capable of detecting ‘micro-expressions’;⁷⁸ recruitment firms that claimed to predict job performance solely from an applicant’s photo or short video;⁷⁹ or, most controversially, applications that claimed to predict criminality based only off photos of a face.⁸⁰ The problem here lies in the fact that such deployments can result in systems that are intrinsically limited in their performance, but where these flaws can be hard to spot by outsiders or unsuspecting users. More fundamentally, this illustrates how if the basic assumptions are flawed, there might be certain applications where no AI system will ever be adequate, regardless of progress. To be certain, AI systems can at times pick up on subtle patterns that humans cannot recognize. Yet if there simply is no underlying correlation (e.g. no meaningful ‘signal’ to be found), then there are information-theoretic limits that no AI system, no matter how capable, would be able to overcome.

In the fifth place, some AI applications display a *wrong tool* problem. This occurs when the use to which some algorithm is put meets some perceived narrow organisational objective, but is in fact at odds with the broader needs, functions, or purposes of the principal deploying them, or the broader society. For instance, in the context of legal systems, predictive algorithms—even if (or precisely when) they are accurate in individual cases, may create damaging or arbitrary incentives for citizens, or can end up affecting and corrupting their own training data in ‘runaway feedback loops’ of predictive policing.⁸¹ Some have argued that in such highly sensitive contexts, the use of predictive algorithms may be fundamentally misaligned with the purposes of judicial systems.⁸²

⁷⁶ Arvind Narayanan, *How to recognize AI snake oil* (2019), <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>.

⁷⁷ This is also a considerable problem in the field of so-called ‘black box medicine’. See W. NICHOLSON PRICE, *Black-Box Medicine* (2014), <https://papers.ssrn.com/abstract=2499885> (last visited Jun 23, 2018).

⁷⁸ Louise Marie Jupe & David Adam Keatley, *Airport artificial intelligence can detect deception: or am I lying?*, SECUR. J. (2019), <https://doi.org/10.1057/s41284-019-00204-7> (last visited Oct 8, 2019). That does not mean however that such systems cannot infer sensitive (and even tacit) psychological or mental information by observing samples of our online behaviour. For a broad discussion, see Christopher Burr & Nello Cristianini, *Can Machines Read our Minds?*, MINDS MACH. (2019), <https://doi.org/10.1007/s11023-019-09497-4> (last visited Jun 27, 2019).

⁷⁹ Manish Raghavan et al., *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, ARXIV190609208 CS (2019), <http://arxiv.org/abs/1906.09208> (last visited Jul 1, 2020).

⁸⁰ Coalition for Critical Technology, *Abolish the #TechToPrisonPipeline*, MEDIUM (2020), <https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techttoprisonpipeline-9b5b14366b16> (last visited Sep 3, 2020).

⁸¹ Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, ARXIV170609847 CS STAT (2017), <http://arxiv.org/abs/1706.09847> (last visited May 13, 2019).

⁸² See for instance Pasha Kamyshev, *Machine Learning In The Judicial System Is Mostly Just Hype*, PALLADIUM MAGAZINE (2019), <https://palladiummag.com/2019/03/29/machine-learning-in-the-judicial-system-is-mostly-just-hype/> (last visited Apr 23, 2019). (arguing that a judicial system balances four key purposes—provide incentives against infractions, provide safety for other citizens; pursue rehabilitation of convicted criminals, and provide closure for victims—none of which are adequately or well served by the use of crime prediction algorithms).

In the sixth place, AI has many *underappreciated barriers to adoption* and further penetration, which prevent easy use by many enterprises. For instance, a 2018 US survey suggested that only 2.8% of American businesses had implemented machine learning.⁸³ Likewise, a 2020 European survey found that companies experienced considerable barriers to adoption.⁸⁴ Other surveys have found time lags of at least several years between cutting-edge AI techniques being developed in pure research settings, and these approaches seeing application to grounded, real world use cases.⁸⁵ Indeed, even for well-resourced governments, the process of procuring AI is rarely frictionless and pristine. For instance, Maaike Verbruggen has investigated differences in innovation culture, business practices, and values amongst civilian tech companies and militaries, which provide underappreciated hurdles to military procurement of AI.⁸⁶ More generally, public resistance and activism in AI industry may hinder government procurement or deployment of certain AI applications, as illustrated by high-profile withdrawals by some US tech companies from military contracts,⁸⁷ or the decision by various US companies to refrain from offering facial recognition services to police departments.⁸⁸

That is not to say, of course, that governments will find it impossible to find private vendors for certain categories of AI applications.⁸⁹ However, it does suggest how, especially in a

⁸³ Kristina McElheran et al., *Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey* 72 (2020); Will Knight, *AI Is All the Rage. So Why Aren't More Businesses Using It?*, WIRED, 2020, <https://www.wired.com/story/ai-why-not-more-businesses-use/> (last visited Aug 11, 2020).

⁸⁴ Specifically, the European companies surveyed reported difficulties in hiring new staff with the right skills (57%), the cost of adoption (52%), or the cost of adapting operational processes to AI (49%). SNEZHA KAZAKOVA ET AL., *European enterprise survey on the use of technologies based on artificial intelligence* (2020), <https://ec.europa.eu/digital-single-market/en/news/european-enterprise-survey-use-technologies-based-artificial-intelligence>.

⁸⁵ For instance, a recent survey of applied neural network-based vision systems to identify plant diseases found that the most prominent framework was TensorFlow (developed in 2015), even though many researchers in pure research fields are already shifting to new frameworks such as PyTorch. See Andre S. Abade, Paulo Afonso Ferreira & Flavio de Barros Vidal, *Plant Diseases recognition on images using Convolutional Neural Networks: A Systematic Review*, ARXIV200904365 Cs (2020), <http://arxiv.org/abs/2009.04365> (last visited Sep 16, 2020); For commentary, see Jack Clark, *Import AI 214: NVIDIA+ARM; a 57-subject NLP test; and AI for plant disease identification* (2020), <https://us13.campaign-archive.com/?u=67bd06787e84d73db24fb0aa5&id=a88c355f6a> (last visited Sep 16, 2020).

⁸⁶ Challenging the common perception that there is little to constrain the proliferation of Lethal Autonomous Weapons Systems, she suggests that civil-military cooperation (or ‘spin-in’) on producing LAWS will be far more hindered and inhibited than is often anticipated. Maaike Verbruggen, *The Role of Civilian Innovation in the Development of Lethal Autonomous Weapon Systems*, 10 GLOB. POLICY 338–342 (2019).

⁸⁷ Haydn Belfield, *Activism by the AI Community: Analysing Recent Achievements and Future Prospects*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 15–21 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375814> (last visited Feb 12, 2020).

⁸⁸ Jay Peters, *IBM will no longer offer, develop, or research facial recognition technology*, THE VERGE (2020), <https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-recognition-analysis-software> (last visited Jun 22, 2020); AXON AI & POLICING TECHNOLOGY ETHICS BOARD, *First Report of the Axon AI & Policing Technology Ethics Board* (2019), https://static1.squarespace.com/static/58a33e881b631bc60d4f8b31/t/5d13d7e1990c4f00014c0aeb/1561581540954/Axon_Ethics_Board_First_Report.pdf (last visited Jun 28, 2019).

⁸⁹ Notably, surveys have found only mixed willingness amongst tech companies to rule out participation in the production of lethal autonomous weapons systems; Sliper and colleagues have surveyed 50 tech companies from 12 countries, and on the basis of the responses, classified 7 companies as showing ‘best practice’ (meaning the company “clearly committed to ensuring its technology will not be used to develop or produce autonomous weapons”), 22 as companies of ‘medium concern’, and 21 as ‘high concern’. FRANK SLIJPER ET AL., *Don't be evil? A survey of the tech sector's stance on autonomous weapons* 5 (2019), <https://www.paxforpeace.nl/publications/all-publications/dont-be-evil>.

defence context, AI procurement challenges and more broadly relations between defence departments and industry can slow down or prove a considerable bottleneck.⁹⁰ It highlights, more broadly, the degree to which actors' internal organisational culture, or fits or mismatches amongst key actors' respective cultures, can variably strengthen or inhibit their ability to effectively procure and deploy a given AI application.

2.1.4.2 Challenges from input preconditions to AI deployment

Moreover, the deployment of AI tools is at times constrained by pragmatic issues relating to the availability of key inputs—computing hardware, training data, and skilled human expertise.⁹¹

One notable constraint is *data*: training up AI systems to a level where they can perform at an adequate level currently requires access to large—and in the case of supervised learning, pre-hand-labelled—datasets, often of thousands to hundreds of thousands of instances. While some large datasets are relatively public and open-source (e.g. website text corpora), in other application domains, only some actors may have access to the requisite datasets.⁹² In other domains, obtaining sufficiently clean or reliable training data may remain non-viable for a long time.⁹³ Different parties' ability to marshal sufficient training data is therefore dependent on a range of both industrial and legal factors, from their access to smart infrastructure to data privacy regulations.

A second requirement is *computational power*.⁹⁴ Training a leading AI system requires considerable computing power, and these requirements have been growing steeply. Between 2012 and 2018, the amount of 'compute' used in top AI training runs increased exponentially,⁹⁵ with a

⁹⁰ Gavin Pearson, Phil Jolley & Geraint Evans, *A Systems Approach to Achieving the Benefits of Artificial Intelligence in UK Defence*6 (2018), <https://arxiv.org/abs/1809.11089>; Trevor Taylor, *Artificial Intelligence in Defence: When AI Meets Defence Acquisition Processes and Behaviours*, 164 RUSI J. 72–81 (2019).

⁹¹ See BEN BUCHANAN, *The AI Triad and What It Means for National Security Strategy* (2020), <https://cset.georgetown.edu/research/the-ai-triad-and-what-it-means-for-national-security-strategy/> (last visited Aug 19, 2020).

⁹² For instance, social media companies can rely on user-labelled social media posts or photos; governments are able to leverage records of judicial decisions, etc.

⁹³ Some have argued that machine learning cannot be implemented in nuclear command and control, for lack of adequate or sufficiently reliable datasets of nuclear targets. Rafael Loss & Joseph Johnson, *Will Artificial Intelligence Imperil Nuclear Deterrence?*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/> (last visited Dec 4, 2019). However, see also Zachary Kallenborn, *AI Risks to Nuclear Deterrence Are Real*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/> (last visited Oct 16, 2019).

⁹⁴ Tim Hwang, *Computational Power and the Social Impact of Artificial Intelligence* (2018), <https://arxiv.org/abs/1803.08971> (last visited Mar 27, 2018). For an introduction to AI computing hardware chips, see SAIF M. KHAN & ALEXANDER MANN, *AI Chips: What They Are and Why They Matter* (2020), <https://cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/> (last visited Aug 19, 2020).

⁹⁵ Dario Amodei & Danny Hernandez, *AI and Compute*, OPENAI BLOG (2018), <https://blog.openai.com/ai-and-compute/> (last visited May 22, 2018).

3.5-month doubling-time compared to Moore's Law's 18-month doubling period.⁹⁶ This equates to an increase of more than 300.000x over a six-year period.⁹⁷

A third requirement comes in the form of human *expertise*. At present, 'tacit knowledge' about the configuration and training of ML systems remains one limit on the unrestrained proliferation of AI applications, even if it is at times underestimated by policymakers or the public.⁹⁸ These are important and at times underappreciated constraints on AI.⁹⁹

2.1.4.3 Challenges to current AI techniques and approaches

Moreover, underlying many of AI's pragmatic challenges is the fact that, for all of its real and significant successes, the current generation of AI techniques still has limits.¹⁰⁰ There are outstanding challenges, both for the algorithms, as well as broader scientific field. Today's AI systems, being narrow, do not possess adequate 'common sense',¹⁰¹ and suffer from a cluster of problems sometimes described collectively as 'artificial stupidity'.¹⁰² This has a number of operational implications.

For one, many ML algorithms are *brittle*. If an AI image-recognition system runs into an apparent anomaly (e.g. classifying a new object on the highway as a shark), it will not by itself recognize the absurdity of this. As such, deep neural networks can fail to generalize to 'out-of-

⁹⁶ Gordon E. Moore, *Cramming more components onto integrated circuits*, 38 ELECTRONICS 82–85 (1965). Note that in this initial formulation, Moore predicted an annual doubling, which he revised to a 2-year doubling period in 1975, Gordon E. Moore, *Progress in Digital Integrated Electronics*, TECHNICAL DIGEST, 1975, at 11–13. The frequently cited prediction of a '18-month' doubling time was instead made by Intel executive David House, by considering not just the number of transistors, but also improvements in transistor speed; see Michael Kanellos, *Moore's Law to roll on for another decade*, CNET, 2003, <https://www.cnet.com/news/moores-law-to-roll-on-for-another-decade/> (last visited Jul 25, 2019).

⁹⁷ Amodei and Hernandez, *supra* note 95. However, for commentary, see also Ryan Carey, *Interpreting AI Compute Trends*, AI IMPACTS (2018), <https://aiimpacts.org/interpreting-ai-compute-trends/> (last visited Jan 16, 2020); Ben Garfinkel, *Reinterpreting "AI and Compute"*, AI IMPACTS (2018), <https://aiimpacts.org/reinterpreting-ai-and-compute/> (last visited Jan 16, 2020).

⁹⁸ Significantly, this is not the only domain in which human talent has proven a restrictive precondition. Indeed, in the domain of nuclear weapons, Mackenzie & Spinardi have influentially chronicled the ways in which constrained scientific 'tacit' knowledge served as an important inhibitor (and even arguably, force of 'reversal') on nuclear proliferation; Donald MacKenzie & Graham Spinardi, *Tacit Knowledge, Weapons Design, and the Uninvention of Nuclear Weapons*, 101 AM. J. SOCIOL. 44–99 (1995); Donald Mackenzie, *Uninventing the bomb?*, 12 MED. WAR 202–211 (1996).

⁹⁹ Though all three thresholds may be falling, as will be discussed in section 2.1.7.1.

¹⁰⁰ See for instance the technological and practical barriers enumerated in: Thilo Hagendorff & Katharina Wezel, *15 challenges for AI: or what AI (currently) can't do*, AI Soc. (2019), <https://doi.org/10.1007/s00146-019-00886-y> (last visited Apr 12, 2019).

¹⁰¹ Indeed, the lack of common sense is one of the oldest identified shortfalls in AI. John McCarthy, *Programs with Common Sense*15 (1959), <http://www-formal.stanford.edu/jmc/mcc59.html>. This remains a problem in modern ML systems. Ernest Davis & Gary Marcus, *Commonsense reasoning and commonsense knowledge in artificial intelligence*, 58 COMMUN. ACM 92–103 (2015). However, there has also been recent progress in combining text transformers such as GPT-2 with older approaches in order to produce some measure of common-sense. See Antoine Bosselut et al., *COMET: Commonsense Transformers for Automatic Knowledge Graph Construction*, ARXIV190605317 CS (2019), <http://arxiv.org/abs/1906.05317> (last visited Aug 26, 2020); John Pavlus, *Common Sense Comes to Computers*, QUANTA MAGAZINE, 2020, <https://www.quantamagazine.org/common-sense-comes-to-computers-20200430/> (last visited Aug 26, 2020).

¹⁰² DOMINGOS, *supra* note 13.

distribution’ (OoD) inputs, such as objects presented at strange angles.¹⁰³ This has proven one problem to many self-driving car systems.¹⁰⁴ Relatedly, AI is also susceptible to ‘*adversarial input*’—data (e.g. visual or auditory patterns) designed to alter the way the system processes stimuli, making the system ‘hallucinate’ inputs (e.g. objects or sounds) that are not there.¹⁰⁵ AI systems are also prone to ‘*catastrophic forgetting*’, or the tendency of some AI systems to lose previous knowledge as they learn new information.¹⁰⁶ As a result of this, the performance of AI systems often degrades in contexts that do not resemble their training data.¹⁰⁷

As already alluded to, many ML algorithms can also suffer from *data bias*.¹⁰⁸ The AI system’s lack of ‘common sense’ means that if a dataset is biased, or unrepresentative of the real world, an AI system will not recognize this, reject it as inadequate, or find a way to clean or correct the data. Instead, it will simply reproduce or amplify existing bias.¹⁰⁹ In this way, AI systems can end up perpetuating past patterns or biases.¹¹⁰

Finally, certain AI approaches have problems with the *lack of transparency* or (*un*)*explainability* of their decisions. This opacity occurs for three reasons: (1) intentional corporate or state secrecy, (2) technical illiteracy of users, and (3) an opacity that arises from the characteristics of machine learning algorithms applied at scale.¹¹¹ This third type of opacity derives from the fact that the same functionalities that make neural networks useful—the way they can encode information—also makes it hard for even their designers to understand their behaviour or reactions.¹¹²

2.1.4.4 Challenges in the field of AI

Finally, there are outstanding challenges in the *scientific field of AI research*. Despite—or perhaps because—of its high-profile success, AI researchers have at times struggled with high expectations, and this has led to a series of ‘growing pains’.

¹⁰³ Michael A. Alcorn et al., *Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects*, ARXIV181111553 Cs (2019), <http://arxiv.org/abs/1811.11553> (last visited May 19, 2020).

¹⁰⁴ An underlying problem is ‘state-ambiguity’, where a certain input (e.g. a row of cars) could be interpreted in different ways (‘row of parked cars’; ‘cars in a traffic jam’) that would recommend different actions. Humans find it easy to distinguish between those, but the system finds it challenging. I thank James Fox for this clarification.

¹⁰⁵ Ian J. Goodfellow et al., *Attacking machine learning with adversarial examples*, OPENAI (2017), <https://openai.com/blog/adversarial-example-research/> (last visited Feb 18, 2017).

¹⁰⁶ SCHARRE AND HOROWITZ, *supra* note 13 at 6.

¹⁰⁷ The so-called problem of ‘distributional shift’. Dario Amodei et al., *Concrete Problems in AI Safety*, ARXIV160606565 Cs (2016), <http://arxiv.org/abs/1606.06565> (last visited May 13, 2017).

¹⁰⁸ Lehr and Ohm, *supra* note 71.

¹⁰⁹ Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 671 CALIF. LAW REV. (2016), <https://papers.ssrn.com/abstract=2477899> (last visited Jan 28, 2019); Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOCIOl. METHODS RES. 004912411878253 (2018).

¹¹⁰ Hagendorff and Wezel, *supra* note 100.

¹¹¹ Burrell Jenna Burrell, *How the machine ‘thinks’: Understanding opacity in machine learning algorithms*, 3 BIG DATA SOC. 2053951715622512 (2016).

¹¹² Although there are also counter-arguments; see for instance Joshua A. Kroll, *The fallacy of inscrutability*, 376 PHIL TRANS R SOC A 20180084 (2018). There has also been work in increasing the interpretability of AI. Leilani H. Gilpin et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning*, ARXIV180600069 Cs STAT (2019), <http://arxiv.org/abs/1806.00069> (last visited Jun 16, 2020).

Some scholars have suggested that machine learning research has suffered from a ‘replication crisis’, as researchers in some areas have been unable to replicate one another’s results because of different experimental and publication practices.¹¹³ It has been argued that differences in empirical rigor have slowed progress,¹¹⁴ or have even made the field of machine learning akin to ‘alchemy’.¹¹⁵ Underlying this is the problem that the immense computational demands of leading AI projects render it difficult or impossible for many academic AI scientists to replicate the results of top private-sector AI projects.¹¹⁶ More generally, in other cases the replicability of research is challenged by the underlying opacity of particular neural network architectures.

Others suggest that apparent advances in AI may not always be representative or hold up under scrutiny.¹¹⁷ In some cases, newer machine learning systems failed to outperform the performance of older, non-neural algorithms, revealing ‘phantom progress’.¹¹⁸ This is exacerbated by the fact that the AI research community at times lacks standardised benchmarks and metrics, making comparisons difficult.¹¹⁹ In those areas where there are standardised metrics, we often see only small improvements that may not fully capture underlying progress at the level we care about.¹²⁰ Others have noted an increasing ‘narrowing’ of AI research, through the growing dominance of deep learning methods and private labs in AI research, which may be foreclosing

¹¹³ Matthew Hutson, *Artificial intelligence faces reproducibility crisis*, 359 SCIENCE 725–726 (2018).

¹¹⁴ D. Sculley et al., *Winner’s Curse? On Pace, Progress, and Empirical Rigor* (2018), <https://openreview.net/forum?id=rJWF0Fywf> (last visited May 29, 2020).

¹¹⁵ Matthew Hutson, *AI researchers allege that machine learning is alchemy*, SCIENCE | AAAS (2018), <https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy> (last visited May 29, 2020). Note, this is not the first time that the field of AI has been characterised as ‘alchemy’; see famously: HUBERT L. DREYFUS, *Alchemy and Artificial Intelligence* (1965), <https://www.rand.org/pubs/papers/P3244.html>.

¹¹⁶ Accordingly, Brundage, Avin, Wang, Belfield, Krueger and others have argued for the provision of computing subsidies to academics, in order to maintain the ability of independent actors to verify discoveries and reliability assurances made by private actors. Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, ARXIV200407213 Cs (2020), <http://arxiv.org/abs/2004.07213> (last visited Apr 16, 2020).

¹¹⁷ Matthew Hutson, *Eye-catching advances in some AI fields are not real*, SCIENCE | AAAS (2020), <https://www.sciencemag.org/news/2020/05/eye-catching-advances-some-ai-fields-are-not-real> (last visited May 29, 2020). For instance, it is also claimed that advances in algorithmic accuracy over the past years reflected flaws in the experimental setup of those papers. Kevin Musgrave, Serge Belongie & Ser-Nam Lim, *A Metric Learning Reality Check*, ARXIV200308505 Cs (2020), <http://arxiv.org/abs/2003.08505> (last visited May 29, 2020).

¹¹⁸ For instance, one 2019 study suggested that neural network recommendation systems, such as those used by media streaming services, failed to outperform much older and simpler non-neural algorithms. Maurizio Ferrari Dacrema, Paolo Cremonesi & Dietmar Jannach, *Are we really making much progress? A worrying analysis of recent neural recommendation approaches*, in PROCEEDINGS OF THE 13TH ACM CONFERENCE ON RECOMMENDER SYSTEMS 101–109 (2019), <https://doi.org/10.1145/3298689.3347058> (last visited May 28, 2020). Likewise, a 2019 review of information retrieval algorithms in search engines concluded that the “high-water mark ... was actually set in 2009.” Wei Yang et al., *Critically Examining the “Neural Hype”: Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models*, in PROCEEDINGS OF THE 42ND INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 1129–1132 (2019), <https://doi.org/10.1145/3331184.3331340> (last visited May 28, 2020).

¹¹⁹ This is the case both for inter-algorithm comparison: Davis Blalock et al., *What is the State of Neural Network Pruning?*, in PROCEEDINGS OF MACHINE LEARNING AND SYSTEMS 2020 (MLSYS 2020) 18 (2020), <https://proceedings.mlsys.org/book/296.pdf>. More generally, in many areas such as image recognition tasks, it can be hard to meaningfully compare the performance of algorithms with humans. Christina M. Funke et al., *The Notorious Difficulty of Comparing Human and Machine Perception*, ARXIV200409406 Cs Q-BIO STAT (2020), <http://arxiv.org/abs/2004.09406> (last visited Sep 27, 2020).

¹²⁰ I thank Jaime Sevilla for this point.

the exploration of other avenues.¹²¹ Finally, others point not to scientific challenges but to physical limits, arguing that the current ‘wave’ in AI will slow down, because of the computational limits of scaling up current techniques.¹²²

As a result of this, some researchers have even begun to predict a new ‘AI winter’,¹²³ or at least are arguing that current deep learning approaches are insufficient to sustain further progress for much longer, and would need to be cross-fertilised with other approaches if any progress is to be made towards more general AI systems capable of ‘common-sense’ reasoning.¹²⁴

2.1.5 AI as high-variance technology

At a glance, the above extensive discussion might be taken as strong evidence that, notwithstanding a few vivid successes, AI technology has been mostly or entirely overblown, that performance is plateauing, and that a new ‘AI Winter’ may shortly be upon us.

Yet while an injection of sobriety would certainly be welcome in many debates around AI, pervasive or excessive scepticism might be a wrong lesson to draw at this point. The reality may be messier, and reflect something more like ‘Amara’s Law’, which holds that “we tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run”.¹²⁵ As such, while this project focuses primarily on governance trajectories for existing and near-term AI capabilities, given its emphasis on ‘governance for change’, the question of possible future AI progress remains of relevance to explore the scope (and ceiling) of pressures that governance systems may be put under.

With respect to the widespread challenges of AI applications in industry, while we should certainly identify and call out high-profile cases where the use of AI systems has been rushed and fumbled by some actors, it is unclear in which direction this should direct our overall expectations. Such high-profile cases, after all, may be more a symptom of the technology’s perceived low threshold for usage,¹²⁶ than that they are a fundamental indictment of AI’s utility for all actors in all domains. To be sure, the hidden requirements, shortfalls, and risks of deploying immature machine learning to various complex practical contexts ought to make many principals (as well as AI researchers) more cautious or reflexive about when to pursue the rapid deployment of these tools. However, it is also worth asking whether there has ever been a new technology that offered so many use cases at such a (seemingly) low usage threshold, which did not likewise see a wave

¹²¹ Joel Klinger, Juan Mateos-Garcia & Konstantinos Stathopoulos, *A narrowing of AI research?*, ARXIV200910385 Cs (2020), <http://arxiv.org/abs/2009.10385> (last visited Sep 26, 2020).

¹²² Neil C. Thompson et al., *The Computational Limits of Deep Learning*, ARXIV200705558 Cs STAT (2020), <http://arxiv.org/abs/2007.05558> (last visited Jul 15, 2020).

¹²³ Sam Shead, *Are we on the cusp of an ‘AI winter?’*, BBC NEWS, January 12, 2020, <https://www.bbc.com/news/technology-51064369> (last visited May 29, 2020); Floridi, *supra* note 30.

¹²⁴ For instance, Gary Marcus has identified 11 ‘computational primitives’, the achievement of which he argues will require the reintroduction of older approaches to AI. Gary Marcus, *Deep Learning: A Critical Appraisal*, ARXIV180100631 Cs STAT (2018), <http://arxiv.org/abs/1801.00631> (last visited May 19, 2020).

¹²⁵ The adage is named after American futurologist Roy Amara. Roy Amara 1925–2007: American futurologist, *in* OXFORD ESSENTIAL QUOTATIONS (Susan Ratcliffe ed., 2016), <http://www.oxfordreference.com/view/10.1093/acref/9780191826719.001.0001/acref-9780191826719> (last visited Sep 13, 2018).

¹²⁶ Exacerbated by its relatively cheap cost and easy scaling, combined with the widespread misconception that AI algorithms are objective and free from error.

of sub-par applications in its early days, as a result of being pressed into service by inexperienced and overeager users.¹²⁷

With regards to the state of the scientific field, difficulties in reliably measuring and comparing AI algorithmic progress are ground for concern, and might slow ongoing progress relative to what it might otherwise be. However, it is unclear that these should be considered fundamental challenges rather than growing pains. Moreover, given the speed and ease with which software can be copied and diffused, it may be the case that the peak (in either AI science, or AI system performance) may be more significant than the average performance.

As such, a better lesson could be: AI is *diverse, heterogeneous, and high-variance*. In the usage of public and policymakers, ‘AI’ is often still treated as a single label. This can easily prime such observers to expect or ascribe to the technology a uniform level of performance (whether high or low). Yet the ‘AI’ label may be far too coarse to capture the granularity or heterogeneity of both AI science and AI applications. The point might be that rather than displaying even quality—rather than being a monolithic technology that either ‘works flawlessly everywhere’ or ‘is completely over-hyped’—both AI-systems-in-application and AI-as-scientific-field are high-variance. Even if some use cases do not perform well (or perform actively badly, in high-profile ways), many other AI use cases will perform adequately enough in the background, and a few will perform particularly well.¹²⁸

There is no single neat judgment to be made about AI. None of its successes mean that AI is ‘magic’: while it can be opaque, it is not mysterious or supernatural; it rarely if ever works flawlessly; and it will not solve all our problems.¹²⁹ Yet with all that, this does not mean that the technology cannot prove highly transformative over time—and in some sectors, sooner than we expect. More interesting and productive than the binary question of whether AI matters or is entirely overblown, therefore, may be the more heterogeneous question of *where* and *how* AI matters—and how the answers to these questions might change over time.¹³⁰ To what extent should we expect either further AI capability progress, or rapid proliferation? What kinds of change would these produce? These questions are important not just when talking about the prospects for AI today, but also when debating its importance going forward.

¹²⁷ This is an interesting question on itself, though an in-depth investigation is beyond the scope of this dissertation.

¹²⁸ Moreover, while some issues derive from machine learning systems falling short, others derive from them performing all too well. This is reflected in the debates over facial recognition systems; following demonstrations that such systems performed less well on gender or ethnic minorities, several companies worked to ‘debias’ these systems. This sparked reflection over whether ‘unbiased’ AI surveillance was in fact a good thing—i.e. whether to ‘mend’ such applications, or rather to ‘end’ them entirely. Frank Pasquale, *The Second Wave of Algorithmic Accountability*, LPE PROJECT (2019), <https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/> (last visited Sep 3, 2020).

¹²⁹ Although in fairness, the same objection might be levelled at many other solutions that are proposed to diverse global problems, including various regulatory or political solutions that are at times held up as the salutary ‘human’ or ‘social’ alternative to ‘technofixes’. The point is less to adjudicate which types of solutions (whether social or technological) almost always or never work, but rather to work towards better principles for better evaluating the assumptions, uncertainties, resilience, ‘corrigibility’ (that is, the ability to get feedback and adapt) and brittleness (that is, their tendency to fail in ‘graceful’ or in catastrophic ways) of any proposed solutions or interventions, whether their means are social (human or institutional) or technological.

¹³⁰ See also ALLAN DAFOE, *AI Governance: A Research Agenda* 52 15–24 (2018), <https://www.fhi.ox.ac.uk/govaiagenda/> (sketching components of mapping the technical landscape, and progress across it).

As such, the next two sections will explore two distinct (though commensurable) arguments for why we ought to expect considerable societal impact and change from AI. I first (2.1.6) discuss the possibility of ‘Change from Further AI Progress’, with the aim of clearing up some of the confusion around this contested topic, and to argue that while there is pervasive uncertainty around anticipating or forecasting such breakthroughs, this might be grounds not for comfort but for caution and further study. I then (2.1.7) turn to the parallel if complementary scenario of ‘Change from Algorithmic Proliferation’, to argue that even if one assumes AI progress may slow or hit unexpected barriers, and granting all the technology’s present-day limits, we may nonetheless expect the rapid dissemination and usage of various already-existing AI applications within the coming years, with disruptive global effects.

2.1.6 Change from Further AI Progress: AI between hype and counterhype

In the first place, we should explore whether, how, or under what conditions we might expect further societal impact and change from AI on the basis of ongoing and future progress in its underlying capabilities.

Could we expect a continuation of the current ‘AI summer’? After a decade of progress that has been surprisingly rapid (if not spotless), how far- and how fast might we expect AI technology to advance within the coming years? Indeed, while there is some disagreement amongst AI researchers, many expect significant breakthroughs over the coming decades. Surveys of experts show that, on aggregate, they give a 50% chance to some form of ‘High-Level Machine Intelligence’ (HLMI) being achieved by the 2060s, with many disruptive performance levels being reached throughout the intermediate period.¹³¹ This would suggest that, however important or disruptive we consider AI to be today, this importance will only increase. Nonetheless, the question of future progress in AI remains surrounded by deep uncertainties, as well as pervasive public misunderstandings and misperceptions. As such, it is worthwhile briefly delving into it, to clarify the assumptions, uncertainties, and parameters.

2.1.6.1 How ‘transformative’ could advanced AI become?

AI systems today are ‘narrow’, in that they perform well only within the specific tasks that they have been trained for. Accordingly, they cannot be transferred well to new contexts, and they lack broader human sense of context. Nonetheless, many AI researchers assume that whether

¹³¹ Defined as defined as “unaided machines [that] can accomplish every task better and more cheaply than human workers” Katja Grace et al., *When Will AI Exceed Human Performance? Evidence from AI Experts*, 62 J. ARTIF. INTELL. RES. 729–754, 731 (2018). For previous surveys, see also Vincent C. Müller & Nick Bostrom, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, in FUNDAMENTAL ISSUES OF ARTIFICIAL INTELLIGENCE, 553 (Müller, Vincent C. ed., 2016), <http://www.nickbostrom.com/papers/survey.pdf> (“[a median estimate of] a one in two chance that high-level machine intelligence will be developed around 2040-2050, rising to a 9 in 10 chance by 2075.”). See also S.D. Baum, Ben Goertzel & Ted G. Goertzel, *How long until human-level AI? Results from an Expert Assessment*, 78 TECHNOL. FORECAST. SOC. CHANGE 185–195 (2011). More recent surveys find slightly closer timelines: Ross Gruetzmacher, David Paradice & Kang Bok Lee, *Forecasting Transformative AI: An Expert Survey*, ARXIV190108579 Cs (2019), <http://arxiv.org/abs/1901.08579> (last visited Apr 17, 2019). Interestingly, even the most ‘optimistic’ expert estimates appear further out than those given by the general public, which, in a recent (US) survey assigned a 54% likelihood of developing HLMI within the next 10 (!) years. BAOBAB ZHANG & ALLAN DAFOE, *Artificial Intelligence: American Attitudes and Trends* 111 (2019), <https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/>.

progress comes fast or slow, further development may yield increasingly advanced AI systems that can precipitate considerable or even transformative social changes.¹³²

What would such systems look like? Notions of advanced AI come under a wide diversity of headers and terms: a well-known if relatively older distinction has been drawn between ‘weak’ AI and ‘strong’ AI.¹³³ Others have focused on benchmarks such as “human-level AI” (HLAI)¹³⁴ or “high-level machine intelligence” (HLMI).¹³⁵ One common term in recent years has been reference to “artificial general intelligence” (AGI).¹³⁶ As contrasted with present-day ‘narrow’ AI systems, AGI denotes hypothetical future AI systems that would achieve “the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments”,¹³⁷ achieving performance as well as a human in many or all domains.

However, an interesting common thread in these definitions is that they make broad assumptions about what AI systems would need to look like (and which technological breakthroughs would therefore be necessary) in order to produce considerable societal change. That is, they imply that “the only types of systems we should be concerned about are those which are human-level or sufficiently general in their capabilities.”¹³⁸ Yet from the perspective of assessing the technology’s societal impact, that benchmark may be overtly strict. Various scholars have increasingly noted how even if AGI were to be proven impossible or far away, the coming decades may plausibly see a broad spectrum of AI systems that could drive far-reaching societal change.¹³⁹ This has resulted in the concept of ‘transformative AI’ (TAI),¹⁴⁰ a class which would, at its ceiling, include ‘AGI’ like systems, but which would also cover many simpler systems.¹⁴¹

There remains considerable debate over advanced AI, and especially the idea of AGI. Some have rejected out of hand the very possibility of AGI, or have objected that the concept of ‘AGI’ is meaningless, because even human intelligence is not properly ‘general’, but rather performs differently across many dimensions, and is indeed outperformed on some of these dimensions by

¹³² Edward Parson et al., *Artificial Intelligence in Strategic Context: An Introduction*, AI PULSE (2019), <https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/> (last visited Feb 26, 2019). For a survey, see also Grace et al., *supra* note 131.

¹³³ RAY KURZWEIL, THE SINGULARITY IS NEAR 260 (2005).

¹³⁴ John McCarthy, *From here to human-level AI*, in PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING 640–646 (1996).

¹³⁵ Grace et al., *supra* note 131.

¹³⁶ ARTIFICIAL GENERAL INTELLIGENCE, (Ben Goertzel & Cassio Pennachin eds., 2007), http://link.springer.com/10.1007/978-3-540-68677-4_5 (last visited Jun 8, 2020).

¹³⁷ Ben Goertzel, *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, 5 J. ARTIF. GEN. INTELL. 1–48, 2 (2014). See also Shane Legg & Marcus Hutter, *Universal Intelligence: A Definition of Machine Intelligence*, 17 MINDS MACH. 391–444 (2007).

¹³⁸ Ross Gruetzmacher & Jess Whittlestone, *Defining and Unpacking Transformative AI*, ARXIV191200747 CS, 1 (2019), <http://arxiv.org/abs/1912.00747> (last visited Jan 6, 2020).

¹³⁹ This anticipates our discussion, in section 2.1.7, on ‘disruptive AI’.

¹⁴⁰ Holden Karnofsky, *Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity*, OPEN PHILANTHROPY PROJECT (2016), <http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity> (last visited Mar 4, 2017).

¹⁴¹ Gruetzmacher and Whittlestone furthermore distinguish between *Transformative AI (TAI)* (“AI capabilities or products which lead to societal change comparable to that precipitated by previous individual [General-Purpose-Technologies], such as nuclear power, the internal combustion engine and electricity.”), and *Radically Transformative AI (RTAI)* (“AI capabilities or products which lead to societal change comparable to that precipitated by the agricultural or industrial revolutions”) Gruetzmacher and Whittlestone, *supra* note 138 at 4.

some animals.¹⁴² Yet putting semantics aside, there is a surprising amount of agreement amongst AI researchers that it would be *theoretically* possible to build some forms of ‘high-level machine intelligence’.¹⁴³ There is far less consensus as to when this could happen, however, with estimates ranging from the technology being a decade out to centuries away.¹⁴⁴

Nonetheless, there are active research efforts pursuing advanced AI. One 2017 survey found dozens of projects worldwide pursuing the development of AGI or AGI-like systems.¹⁴⁵ Of course, their prospects for success are deeply unclear. Even amongst those who expect or pursue such technologies, there remain many live debates over the possible or plausible pathways;¹⁴⁶ over expected timelines; over whether we could expect a sudden ‘discontinuous’ leap in AI capabilities, or whether we should expect continued gradual and incremental progress;¹⁴⁷ and over the risks that might be involved in creating such advanced systems before we have

¹⁴² Kevin Kelly, *The Myth of a Superhuman AI*, BACKCHANNEL (2017), <https://backchannel.com/the-myth-of-a-superhuman-ai-59282b686c62> (last visited May 13, 2017). Though for a response, see STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL 166 (2019), <https://www.amazon.com/Human-Compatible-Artificial-Intelligence-Problem-ebook/dp/B07N5J5FTS> (last visited Dec 4, 2019). For a discussion of the versatility of ‘intelligence’ (given various animals’ performance on such tests), see also Section 2.1.6.3. For a satirical take on some often-raised but underexamined claims made to argue that ‘smarter-than human AI is impossible’, see Ben Garfinkel et al., *On the Impossibility of Supersized Machines*, ARXIV170310987 PHYS. (2017), <http://arxiv.org/abs/1703.10987> (last visited Nov 12, 2018).

¹⁴³ Grace et al., *supra* note 131.

¹⁴⁴ *Id.* at 729.

¹⁴⁵ SETH D. BAUM, *A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy* (2017), <https://papers.ssrn.com/abstract=3070741> (last visited Jan 17, 2018).

¹⁴⁶ For an older survey, see: Sam Adams et al., *Mapping the Landscape of Human-Level Artificial General Intelligence*, 33 AI MAG. 25–42 (2012). See also Daniel Eth, *The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes*, 41 INFORMATICA (2017), <http://www.informatica.si/index.php/informatica/article/view/1874> (last visited Jan 2, 2018). Allan Dafoe has distinguished between scenarios that focus on (1) a (single agent-like) ‘superintelligence’ perspective, which can be contrasted or complemented with an (2) ‘AI ecology’ perspective, which anticipates “a diverse, global, ecology of AI systems [some of which] may be like agents, but others may be more like complex services, systems, or corporations.” These can in turn be contrasted with the (3) ‘GPT perspective’, which sees AI as a General-Purpose Technology (see also section 2.1.7.2), and which emphasizes the potential disruptive effects of not only agent-like AI or powerful AI systems, but also many ways in which even mundane AI could “transform fundamental parameters in our social, military, economic, and political systems, from developments in sensor technology, digitally mediated behaviour, and robotics.” Allan Dafoe, *AI Governance: Opportunity and Theory of Impact*, EA FORUM (2020), <https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact> (last visited Sep 20, 2020). Perspective (1) is represented by: NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (2014); RUSSELL, *supra* note 142; MAX TEGMARK, LIFE 3.0: BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE (2017). The (2) AI ecology perspective is instead represented by ROBIN HANSON, THE AGE OF EM: WORK, LOVE, AND LIFE WHEN ROBOTS RULE THE EARTH (2016); K ERIC DREXLER, *Reframing Superintelligence: Comprehensive AI Services as General Intelligence* 210 (2019), https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.

¹⁴⁷ The notion of sudden, discontinuous capability improvements (an ‘intelligence explosion’) has a long history; cf. I.J. Good, *Speculations Concerning the First Ultraintelligent Machine*, 6 in ADVANCES IN COMPUTERS 31–88 (Franz L. Alt & Morris Rubinoff eds., 1964); David J. Chalmers, *The Singularity: A Philosophical Analysis*, 17 J. CONSCIOUS. STUD. 7–65 (2010); MURRAY SHANAHAN, THE TECHNOLOGICAL SINGULARITY (1 edition ed. 2015). However, it also has received critiques, see AI Impacts, *Evidence against current methods leading to human level artificial intelligence*, AI IMPACTS (2019), <https://aiimpacts.org/evidence-against-current-methods-leading-to-human-level-artificial-intelligence/> (last visited Oct 1, 2019); Paul Christiano, *Takeoff speeds*, THE SIDEWAYS VIEW (2018), <https://sideways-view.com/2018/02/24/takeoff-speeds/> (last visited Sep 3, 2020); Alessio Plebe & Pietro Perconti, *The Slowdown Hypothesis*, in SINGULARITY HYPOTHESES 349–365 (Amnon H. Eden et al. eds., 2012), http://link.springer.com/chapter/10.1007/978-3-642-32560-1_17 (last visited Sep 20, 2016).

adequately come to terms with the challenge of ‘aligning’ them with widely shared human values or interests.¹⁴⁸

This project is relatively agnostic about the possibility of- or timelines towards ‘AGI’. Instead, from the perspective of exploring potential ‘change from further AI progress’, it takes a greater interest in a broader set of scenarios where the availability and use of more capable AI systems could drive far-reaching disruption. As such, this line of argument is mostly interested in further progress towards ‘Advanced AI’, which Allan Dafoe has loosely defined as “systems substantially more capable (and dangerous) than existing systems, without necessarily invoking specific generality capabilities or otherwise as implied by concepts such as “Artificial General Intelligence”.”¹⁴⁹

As such, despite far-reaching uncertainties, and whatever one’s assumptions or expectations about AGI, the more direct questions over the magnitude and rate of future AI progress may be critical in setting an approximate ceiling to capabilities and impact, but also in considerations of the adaptability we want from our institutions and governance structures.¹⁵⁰ As such, it can be useful to briefly touch on the prospects of future progress, from both an ‘outside’ view and an ‘inside’ view.¹⁵¹

2.1.6.2 Outside view: barriers, rapids, and vagaries in scientific progress

The ‘outside view’ on this question aims to step back from the object-level considerations about AI and AI progress, to ask more general or comparative questions around the topology of

¹⁴⁸ Iason Gabriel, *Artificial Intelligence, Values and Alignment*, DEEP. RES. (2020), https://deepmind.com/research/publications/Artificial-Intelligence-Values-and-Alignment?fbclid=IwAR1gmUGQSnHD-ly56XDKde4Cvo3OTwMddtHOX3aBxo_xTJiBgGEOOSe_mlk (last visited Feb 3, 2020). See also informal overviews of recent developments in these debates Tom Adamczewski, *A shift in arguments for AI risk*, FRAGILE CREDENCES (2019), <https://fragile-credences.github.io/prioritising-ai/> (last visited Feb 21, 2020); Richard Ngo, *Disentangling arguments for the importance of AI safety*, THINKING COMPLETE (2019), <https://thinkingcomplete.blogspot.com/2019/01/disentangling-arguments-for-importance.html> (last visited Jan 22, 2019). Howie Lempel, Robert Wiblin & Keiran Harris, *Ben Garfinkel on scrutinising classic AI risk arguments*, <https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/> (last visited Sep 26, 2020).

¹⁴⁹ DAFOE, *supra* note 130 at 5.

¹⁵⁰ As is discussed in Peter Cihon, Matthijs M. Maas & Luke Kemp, *Should Artificial Intelligence Governance be Centralised?: Design Lessons from History*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 228–234, 231, 233 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375857> (last visited Feb 12, 2020). *Paper [IV]*. See also the discussion in Chapter 7.4.

¹⁵¹ The distinction between an ‘inside view’ and an ‘outside view’ analysis used here, derives from the classical psychological work by Kahnemann, Tverskey, and Lovallo on planning and forecasting biases. In this model, “[a]n inside view forecast is generated by focusing on the case at hand, by considering the plan and the obstacles to its completion, by constructing scenarios of future progress, and by extrapolating current trends. The outside view [...] essentially ignores the details of the case at hand, and involves no attempt at detailed forecasting of the future history of the project. Instead, it focuses on the statistics of a class of cases chosen to be similar in relevant respects to the present one. The case at hand is also compared to other members of the class, in an attempt to assess its position in the distribution of outcomes for the class.” Daniel Kahneman & Dan Lovallo, *Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking*, 39 MANAG. SCI. 17–31, 25 (1993). The insight that outside view forecasts are often more accurate underlies the methodology called ‘reference class forecasting’. In more recent work on ‘superforecasting’ of various events, there is an emphasis on first taking the outside view, and only then to modify the conclusion using the inside view. PHILIP E. TETLOCK & DAN GARDNER, *SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION* (Reprint edition ed. 2016). I thank Olle Häggström for highlighting this link, and reminding me of the appropriate sequence.

scientific progress. How often (and under what circumstances) have high hopes been disappointed? How often have scientists been blindsided by sudden advances? From this perspective, one might consider distinct external observations both against and in favour of expecting considerable further scientific progress.

From a sceptical perspective, we might expect advanced AI (such as AGI) to be even further away than it appears to the average AI researcher. Such a position could find root in the observed difficulty of technological forecasting, and the common and perhaps intuitive assumption that, if or when technology forecasters err, they will usually err on the side of optimism. There is evidence to support this idea. Nuclear fusion is proverbially always ‘just twenty years away’.¹⁵² As noted, during the 1960’s and 1970’s, legal scholars jumped the gun on trying to regulate then-anticipated advances in weather control technology or deep seabed mining, which in the event did not come to pass, or at least took many more decades, respectively.¹⁵³ Indeed, from a pessimistic perspective, there are a range of unexpected barriers that could emerge, which would slow down AI progress in the near- to medium-term, or at least keep very transformative (or certainly AGI-type) capabilities out of reach.

For one, there may be both specific and general *scientific barriers*. Scientific progress has historically been hard-won at best.¹⁵⁴ Beyond certain fields that were held back for some time until appropriate conceptual breakthroughs were made, there are in fact suggestions that the general rate of scientific progress has slowed down over the past decades, because a lot of scientific ‘low-hanging fruit’ has been plucked, and because the task of training larger cohorts of junior researchers up to the expanding edge of knowledge is placing increasing time demands on older generations of senior scientists.¹⁵⁵ In principle, such factors would be expected to slow down science across the board. For instance, a recent survey argues that “the number of researchers required today to achieve the famous doubling every two years of the density of computer chips is more than 18 times larger than the number required in the early 1970s.”¹⁵⁶ As such, whether one expects that further AI advances are (at present, or ‘from here on out’) more bottlenecked by continued increases in computing hardware power, or by peak insights from a few researchers, one might find reasons in either case to expect further friction on fundamental AI progress.

Of course, from an optimist’s perspective, even if critical scientific and conceptual barriers remain to further AI progress, it is unclear why we should expect such scientific insights to remain

¹⁵² Michael Brooks, *Forever 20 years away: will we ever have a working nuclear fusion reactor?*, NEWSTATESMAN, 2014, <https://www.newstatesman.com/sci-tech/2014/11/forever-20-years-away-will-we-ever-have-working-nuclear-fusion-reactor> (last visited Sep 4, 2020).

¹⁵³ Picker, *supra* note 33 at 194. See also section 2.1.1.

¹⁵⁴ Of course, from an ‘inside view’ perspective, this trend is derived from the history of human science, and may not fully account for the possibility that AI systems (including some early present-day systems) could be brought to bear on resolving scientific questions. On the other hand, scientists throughout history have arguably had some benefit from advancing (artefactual and conceptual) tools. I thank Olle Häggström for raising this point.

¹⁵⁵ Nicholas Bloom, *Are Ideas Getting Harder to Find?*, 110 AM. ECON. REV. 1104–1144 (2020). For an exploration, see also Tyler Cowen & Ben Southwood, *Is the rate of scientific progress slowing down?* 43 (2019). Previously, Benjamin Jones has argued that “would-be innovators require more training and specialization to reach the frontier of a given field. Research teams are also getting bigger, and the number of patents per researcher has declined.” Benjamin F. Jones, *The Burden of Knowledge and the “Death of the Renaissance Man”: Is Innovation Getting Harder?*, 76 REV. ECON. STUD. 283–317 (2009).

¹⁵⁶ Bloom, *supra* note 155.

fundamentally intractable for many more years and decades, given that some of the world's wealthiest companies and most powerful countries have now begun to funnel extensive funding and top talent into the pursuit of AI technology. Yet while large investments may perhaps be necessary or at least conducive to scientific progress, they are of course no guarantee of it. Even large state or corporate investments are no guarantee of 'brute-forcing' fundamental conceptual breakthroughs. One need only consider the meagre results delivered by the Human Brain Project (HBP), a 1bn Euro flagship project which has fallen far short of its 2009 claim to be able to simulate a human brain within a decade.¹⁵⁷ The ITER fusion reactor program is projected to average just under \$2 billion per year during 12 years of construction,¹⁵⁸ yet many argue success remains elusive.¹⁵⁹ Indeed, just prior to the second AI winter, DARPA's 1983-1993 Strategic Computing Initiative spent about \$1 billion in pursuit of what we would now call AGI, but failed.¹⁶⁰ That does not mean that scientific problems are impervious to large sums of funding, of course; and some major projects have managed to achieve breakthroughs even under budget.¹⁶¹ Nonetheless, if the fundamental 'surface area' of the scientific problem remains small, extensive resources may produce sharply diminishing returns in the absence of a more informed theoretical understanding of where or how to apply them productively.¹⁶²

Secondly, there may be unexpected *political barriers*, because no scientific field or technology is insulated from its external political and social context. High-profile industrial 'failures' or 'accidents' involving AI systems could sour the public or governments. For example, the Three-Mile Island nuclear power plant meltdown has been argued to have significantly slowed down the development of nuclear power in the US.¹⁶³ A similar public souring on nuclear energy resulted after the 2011 Fukushima nuclear disaster.¹⁶⁴ More broadly, recent years have seen a

¹⁵⁷ See Tim Requarth, *The Big Problem With "Big Science" Ventures—Like the Human Brain Project*, NAUTILUS (2015), <http://nautil.us/blog/the-big-problem-with-big-science-ventureslike-the-human-brain-project> (last visited Jan 21, 2020); Ed Yong, *The Human Brain Project Hasn't Lived Up to Its Promise*, THE ATLANTIC (2019), <https://www.theatlantic.com/science/archive/2019/07/ten-years-human-brain-project-simulation-markram-ted-talk/594493/> (last visited Jan 21, 2020).

¹⁵⁸ Henry Fountain, *A Dream of Clean Energy at a Very High Price*, THE NEW YORK TIMES, March 27, 2017, <https://www.nytimes.com/2017/03/27/science/fusion-power-plant-iter-france.html> (last visited Jun 26, 2020).

¹⁵⁹ Daniel Jassby, *Fusion reactors: Not what they're cracked up to be*, BULLETIN OF THE ATOMIC SCIENTISTS (2017), <https://thebulletin.org/2017/04/fusion-reactors-not-what-theyre-cracked-up-to-be/> (last visited Sep 1, 2020).

¹⁶⁰ ROLAND AND SHIMAN, *supra* note 28.

¹⁶¹ National Human Genome Research Institute, *Human Genome Project FAQ*, GENOME.GOV (2020), <https://www.genome.gov/human-genome-project/Completion-FAQ> (last visited Jun 26, 2020).

¹⁶² See also Levin & Maas, discussing the concept of a scientific 'problem surface', and comparing the present prospects of 'AGI' to the Apollo Program or the Manhattan Project, which were both at a state where they were able to roadmap engineering sub-problems much better. John-Clark Levin & Matthijs M. Maas, *Roadmap to a Roadmap: How Could We Tell When AGI is a 'Manhattan Project' Away?* (2020), http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf.

¹⁶³ Although for a sceptical account, which argues that the role of the Three-Mile Island meltdown in spurring opposition to US nuclear industry is generally overstated, see Nathan Hultman & Jonathan Koomey, *Three Mile Island: The driver of US nuclear power's decline?*, 69 BULL. AT. SCI. 63–70 (2013).

¹⁶⁴ Younghwan Kim, Minki Kim & Wonjoon Kim, *Effect of the Fukushima nuclear disaster on global public acceptance of nuclear energy*, 61 ENERGY POLICY 822–828 (2013). However, see also NUCLEAR ENERGY AGENCY & OECD, *IMPACTS OF THE FUKUSHIMA DAIICHI ACCIDENT ON NUCLEAR DEVELOPMENT POLICIES* 3 (2017), https://www.oecd-ilibrary.org/nuclear-energy/impacts-of-the-fukushima-daiichi-accident-on-nuclear-development-policies_9789264276192-en (last visited Sep 17, 2020). (arguing that "policy changes driven by the Fukushima Daiichi accident have slowed the development of nuclear energy, but countries' policy re-evaluations of nuclear power linked to the accident generally appear to have subsided").

‘techlash’ to the conduct of digital technology companies in a range of sectors.¹⁶⁵ A high-profile and sustained public techlash against the technology or its principals could occur. If this focuses primarily on certain controversial downstream applications (e.g. facial recognition), this might not extend to underlying AI research. But if the backlash is broader, it could conceivably still slow or even stall AI progress, at least in some regions or in some domains.¹⁶⁶

Yet at the same time, while technological forecasting may be coloured by an optimism bias, this does not mean that categorical scepticism is the better heuristic. Indeed, as much as futurist predictions might be skewed, there are also many cognitive biases that could produce undue pessimism. These include an overtly linear and ‘irreducibly complex’ model of the steps that are necessary or sufficient to achieving a particular innovation,¹⁶⁷ along with a tendency, even by subject experts, to underestimate the compounding effects of improvements in scientific tools.¹⁶⁸ As a result, subject matters in various domains have on various occasions perceived—and provided plausible rationales for—the existence of ‘categorical’ or ‘intrinsic’ barriers to further progress, which in the event proved ethereal or merely modest.¹⁶⁹

As such, while from an outside view we must consider unexpected barriers, we also cannot rule out unexpected rapids. Historically, there have been diverse cases of sudden, even

¹⁶⁵ cf. Mark Scott, *In 2019, the ‘techlash’ will go from strength to strength*, POLITICO (2018), <https://www.politico.eu/article/tech-predictions-2019-facebook-techclash-europe-united-states-data-misinformation-fake-news/> (last visited Dec 9, 2019).

¹⁶⁶ Andrew Grotto has drawn a comparison with the historical differences in US and EU attitudes to GMOs, suggesting that, if mishandled, AI technology may experience a similar backlash and subsequent strict regulatory response as GMO crops in the EU. Andrew Grotto, *Genetically Modified Organisms: A Precautionary Tale For AI Governance*, AI PULSE (2019), <https://aipulse.org/genetically-modified-organisms-a-precautionary-tale-for-ai-governance-2/> (last visited Feb 26, 2019). Another compariosn could be drawn with the landscape of geo-engineering research. See Adrian Currie, *Geoengineering Tensions*, FUTURES (2018), <http://www.sciencedirect.com/science/article/pii/S0016328717301696> (last visited Mar 7, 2018).

¹⁶⁷ Gwern Branwen, *Complexity no Bar to AI* (2014), <https://www.gwern.net/Complexity-vs-AI> (last visited Jun 9, 2020) Appendix: “TECHNOLOGY FORECASTING ERRORS: FUNCTIONAL FIXEDNESS IN ASSUMING DEPENDENCIES.” (“Many people use a mental model of technologies in which they proceed in a serial sequential fashion and assume every step is necessary and only all together are they sufficient, and note that some particular step is difficult or unlikely to succeed and thus as a whole it will fail & never happen. But in reality, few steps are truly required. Progress is predictably unpredictable: A technology only needs to succeed in one way to succeed, and to fail it must fail in all ways. There may be many ways to work around, approximate, brute force, reduce the need for, or skip entirely a step, or redefine the problem to no longer involve that step at all”).

¹⁶⁸ Eliezer Yudkowsky, *There’s No Fire Alarm for Artificial General Intelligence*, MACHINE INTELLIGENCE RESEARCH INSTITUTE (2017), <https://intelligence.org/2017/10/13/fire-alarm/> (last visited Oct 24, 2018); Alex Irpan, *My AI Timelines Have Sped Up* (2020), <http://www.alexirpan.com/2020/08/18/ai-timelines.html> (last visited Aug 24, 2020).

¹⁶⁹ For instance, in the first edition of his 2001 book *On the Internet*, Hubert Dreyfus built on his previous critique of AI to argue against the very possibility of searching the internet, claiming that without embodied knowledge, online search would hit an intractable wall. HUBERT DREYFUS, *ON THE INTERNET* 21–25 (1st ed. 2001), <https://www.abebooks.com/first-edition/Internet-Thinking-Action-HUBERT-DREYFUS-Routledge/30612233740/bd> (last visited Sep 20, 2020). (“All search techniques on the Web are crippled in advance by having to approximate a human sense of meaning and relevance grounded in the body, without a body and therefore without commonsense. It follows that search engines are not on a continuum at the far end of which is located the holy grail of computerized retrieval of just that information that is relevant given the interests of the user; they are not even on the relevance dimension.” *Id.* at 25–26.). These sections were quietly dropped from the book’s Second Edition, published after Google’s 2004 IPO. For a contemporary rebuttal of Dreyfus, see also Ian Stoner, *In Defense of Hyperlinks: A Response to Dreyfus*, 7 TECHNÉ RES. PHILOS. TECHNOL. (2004), <https://scholar.lib.vt.edu/ejournals/SPT/v7n3/stoner.html>.

discontinuous progress across various technological benchmarks.¹⁷⁰ In some cases, such technological advances took not only the general public but even close domain experts by surprise.¹⁷¹ Indeed, occasionally even the eventual inventors themselves harboured considerable uncertainty. For instance, in 1901, a mere two years before personally building the first heavier-than-air flyer, Wilbur Wright told his brother that he believed powered flight was fifty years away.¹⁷² Moreover, we do not always recognize signals of technological breakthrough, even after they have already occurred or are ongoing, as illustrated by contemporary scepticism, even during late stages of their development or testing, regarding the possibility of launching rockets outside the Earth's atmosphere,¹⁷³ nuclear fission,¹⁷⁴ nuclear bombs,¹⁷⁵ or penicillin.¹⁷⁶

That is not to suggest that such sudden breakthroughs or advances are at all the scientific norm. However, it can be useful to consider the epistemic situation of society's relation to various failures of technological prediction. *Prima facie*, in cases where predictions of a certain new technology repeatedly fail or are postponed, we would expect to see more high-profile and protracted scientific and public debates held over a longer period of time, than in cases where a

¹⁷⁰ Katja Grace, *Discontinuous progress in history: an update*, AI IMPACTS (2020), <https://aiimpacts.org/discontinuous-progress-in-history-an-update/> (last visited May 11, 2020). This detailed report reviews 37 technological trends, and identifies 10 'large robust discontinuities'—events that "abruptly and clearly contributed more to progress on some technological metric than another century would have seen on the previous trend". These include amongst others the Pyramid of Djoser (2650 BC—structure height trends), the SS Great Eastern (1858—ship size trends), the first and second telegraph (1858, 1866—speed of sending a message across the Atlantic Ocean), the first non-stop transatlantic flight (1919—speed of passenger or military payload travel), first nuclear weapons (1945—relative effectiveness of explosives), first ICBM (1958—average speed of military payload), and the discovery of $\text{YBa}_2\text{Cu}_3\text{O}_7$ as a superconductor (1987—warmest temperature of superconducting). In addition, they also identify 5 'moderate robust discontinuities' (events that suddenly contribute around 10-100 years of progress of previous trends).

¹⁷¹ As late as four years after the first flight of the Wright Flyer, some esteemed US intellectuals still publicly contended that heavier-than-air flight was impossible. Baron Schwartz, *Heavier-Than-Air Flight Is Impossible* (2017), <https://www.xaprb.com/blog/flight-is-impossible/> (last visited Jun 6, 2020).

¹⁷² DONALD B. HOLMES, WILBUR'S STORY 91 (2008). As cited in Yudkowsky, *supra* note 168.

¹⁷³ For instance, in 1929, a Massachusetts newspaper published a sarcastic headline "Moon rocket misses target by 238,799 1/2 miles." They later published an apology. Lihi Gibor, *The Brilliant Rocket Scientist to Come Before His Time*, DAVIDSON INSTITUTE (2016), <http://davidson.weizmann.ac.il/en/online/sciencehistory/brilliant-rocket-scientist-come-his-time> (last visited Jun 8, 2020).

¹⁷⁴ In 1939, three years before he personally oversaw the first critical uranium chain reaction, Enrico Fermi estimated with 90% confidence that it was impossible to use uranium in a fission chain reaction. RICHARD RHODES, THE MAKING OF THE ATOMIC BOMB 280 (1986). Likewise, in 1962, Edward Teller related how "[t]he possibilities of developing an atomic weapon and the desirability of doing it secretly were discussed at a Princeton University conference in which I participated in March 1939...Bohr said this rare variety could not be separated from common uranium except by turning the country into a gigantic factory. Bohr was worried that this could be done and that an atomic bomb could be developed—but he hoped that neither could be accomplished. Years later, when Bohr came to Los Alamos, I was prepared to say, "You see . . ." But before I could open my mouth, he said: "You see, I told you it couldn't be done without turning the whole country into a factory. You have done just that." EDWARD TELLER, THE LEGACY OF HIROSHIMA. 210–211 (1962).

¹⁷⁵ Jean Harrington, *Don't Worry—it Can't Happen*, 162 SCI. AM. 268–268 (1940).

¹⁷⁶ As an illustration of our difficulty in perceiving transformative capability, Allan Dafoe has noted how "it took about 10 years from Fleming's discovery of penicillin to the production of a compelling proof-of-concept and recognition by a major funder (Rockefeller) that this was worth seriously investing in." DAFOE, *supra* note 130 at 11. Referring to ROBERT BUD, PENICILLIN: TRIUMPH AND TRAGEDY 23–34 (2009). Though for a critical analysis, exploring whether, or in what ways penicillin should be considered a 'discontinuous' scientific breakthrough, see AI Impacts, *Penicillin and syphilis*, AI IMPACTS (2015), <https://aiimpacts.org/penicillin-and-syphilis/> (last visited Sep 21, 2020).

predicted technology arrives more or less on schedule, or where an unexpected breakthrough occurred (where there may have been little public anticipation in the preceding years). If that is so, we would expect frustrated technological predictions to generally produce a bigger cultural footprint over a longer period of time, than do successful predictions or unexpected breakthroughs. This outsized footprint in turn may shape or skew our idea of technological prediction as being categorically over-optimistic.

All this is not to suggest that advanced AI (let alone AGI) will arrive out of the blue to prove the sceptics wrong. Outside view arguments based on distinct technologies are far from enough to settle such debates. Rather, it is to emphasize our pervasive epistemic uncertainty—an uncertainty that is only occluded, but not resolved, if we reach for either categorical futurist optimism or sceptical ‘counterhype’. From an outside perspective, it can be hard to understand what other cases should be considered as part of the appropriate ‘reference class’ for debates around AI development. Instead, it is simply far from obvious which trajectory or scenario we should expect to be more accurate. Thoughtful people have at least entertained each, but we remain unsure.

2.1.6.3 Inside view: scaling laws and the ‘Bitter Lesson’?

As such, an alternative approach would be to take an inside view on recent developments and trends in AI. This asks, even if considerably more advanced AI capabilities are possible, how would they be achieved? What are the arguments or evidence available or presented?

There is no clear inside view answer here. On the one hand, the earlier discussed shortfalls in AI systems do provide significant reason for caution. On the other hand, there have been advances on addressing existing limits to AI, with progress on various fronts pushing the envelope of performance onwards. Optimists might point out that AI progress has been considerable over the past few years, and that even on modest readings, current trends in AI inputs (compute, data, labour) of the ‘AI production function’ might be expected to continue for at least some more time.¹⁷⁷

However, will mere increases in training data or computing power be enough to sustain progress? Many AI researchers hold that, incremental progress notwithstanding, many current approaches in machine learning are at some deep level limited or at least insufficient, and that something fundamentally new is needed if we are ever to approach more general or human-like AI systems.¹⁷⁸

Nonetheless, some have argued that the path of AI progress may well be more continuous, and might involve the simple application of greater computational power or data.¹⁷⁹ Richard

¹⁷⁷ This is explored in detail in the next section (2.1.7).

¹⁷⁸ See also GARY MARCUS & ERNEST DAVIS, REBOOTING AI: BUILDING ARTIFICIAL INTELLIGENCE WE CAN TRUST (2019). For one review, AI Impacts, *supra* note 147.

¹⁷⁹ For instance, Google DeepMind in 2017 pioneered ‘distributional’ reinforcement learning, which since then has proved the foundation for its own breakthrough algorithms. Marc G Bellemare, Will Dabney & Remi Munos, *A Distributional Perspective on Reinforcement Learning* 10 (2017). Recent work suggests that this technique may also play a role in biological brains, suggesting this could serve as one technique that could scale well towards more (though not necessarily fully) general capabilities. See Will Dabney et al., *A distributional code for value in dopamine-based reinforcement learning*, NATURE 1–5 (2020); And for commentary, see Karen Hao, *An algorithm that learns through rewards may show how our brain does too*, MIT TECHNOLOGY REVIEW (2020),

Sutton has influentially referred to this as ‘The Bitter Lesson’—the insight that it is ultimately not carefully crafted human domain knowledge distilled in bespoke machines, but rather “general methods that leverage computation [that] are ultimately the most effective, and by a large margin”.¹⁸⁰ Others have called this the ‘scaling hypothesis’, and while it remains a minority position amongst AI researchers, it has at least achieved some successes over recent years.¹⁸¹

This is because on a number of domains, qualitative increases in performance have been achieved without deep conceptual breakthroughs, and instead simply by ‘scaling up’ existing machine learning approaches or algorithms with more training data or computing power.¹⁸² For instance, GPT-2 and GPT-3, two OpenAI language models released to high visibility in 2019 and 2020, illustrated how apparently qualitative increases in performance at text processing and generation could be achieved through such simple ‘grind’.¹⁸³ GPT-3, a massive system revealed in May 2020, demonstrated a remarkable (though not flawless) ability to solve a wide variety of language-based problems it had never or rarely encountered before—from arithmetic to translation, and from anagram puzzles to SAT analogy tests—without specialized training or fine-tuning. Moreover, it was also able to generate news stories that human readers were unable to distinguish from real ones.¹⁸⁴ This work, along with other research, suggests that rather than hitting diminishing returns, neural language models continue to gain in performance as they are made larger and more efficient.¹⁸⁵

Of course, this is a far shot from anything like an AGI-like system: indeed, systems like GPT-3 may not even be on the correct path. Others have pointed out that while the GPT-3 system is impressive as a more general-purpose reasoner at various tasks, it performed worse at some specific tasks than AIs specifically trained for those tasks. On another benchmark of 57 standardised tests, GPT-3 did better than other systems, but still showed considerable limits.¹⁸⁶ Some have argued that such shortfalls illustrate the *intrinsic limits* of the ‘scaling’ approach.¹⁸⁷

<https://www.technologyreview.com/s/615054/deepmind-ai-reinforcement-learning-reveals-dopamine-neurons-in-brain/> (last visited Jan 15, 2020).

¹⁸⁰ Richard Sutton, *The Bitter Lesson* (2019), <http://www.incompleteideas.net/IncIdeas/BitterLesson.html> (last visited May 18, 2020); Though for a response, see also Rodney Brooks, *A Better Lesson* (2019), <https://rodneybrooks.com/a-better-lesson/> (last visited May 18, 2020).

¹⁸¹ Gwern Branwen, *On GPT-3 - Meta-Learning, Scaling, Implications, And Deep Theory* (2020), <https://www.gwern.net/newsletter/2020/05> (last visited Sep 21, 2020); Jack Clark, *Import AI 215: The Hardware Lottery; micro GPT3; and, the Peace Computer*, IMPORT AI (2020), <https://mailchi.mp/jack-clark/import-ai-215-the-hardware-lottery-micro-gpt3-and-the-peace-computer?e=524806dd16> (last visited Sep 21, 2020).

¹⁸² Matthew Hutchinson et al., *Accuracy and Performance Comparison of Video Action Recognition Approaches*, ARXIV200809037 Cs, 6 (2020), <http://arxiv.org/abs/2008.09037> (last visited Aug 24, 2020). (showing how, in the context of video action recognition, simpler approaches which scale better outperform more complex algorithms).

¹⁸³ Radford et al., *supra* note 44; Brown et al., *supra* note 44.

¹⁸⁴ Human accuracy was around 52%; Brown et al., *supra* note 44 at 26. For samples of the system’s writing, see also Gwern Branwen, *GPT-3 Creative Fiction*, 3 (2020), <https://www.gwern.net/GPT-3> (last visited Sep 4, 2020).

¹⁸⁵ Jared Kaplan et al., *Scaling Laws for Neural Language Models*, ARXIV200108361 Cs STAT (2020), <http://arxiv.org/abs/2001.08361> (last visited Feb 3, 2020). See also previous work on scaling relationships in deep learning: Joel Hestness et al., *Deep Learning Scaling is Predictable, Empirically*, ARXIV171200409 Cs STAT (2017), <http://arxiv.org/abs/1712.00409> (last visited Sep 20, 2020).

¹⁸⁶ Dan Hendrycks et al., *Measuring Massive Multitask Language Understanding*, ARXIV200903300 Cs (2020), <http://arxiv.org/abs/2009.03300> (last visited Sep 9, 2020).

¹⁸⁷ Cf. Kyle Wiggers, *OpenAI’s massive GPT-3 model is impressive, but size isn’t everything*, VENTUREBEAT (2020), <https://venturebeat.com/2020/06/01/ai-machine-learning-openai-gpt-3-size-isnt-everything/> (last visited Jun 10, 2020) (“while GPT-3 and similarly large systems are impressive with respect to their performance, they don’t

Others argue that the progress achieved by the scaling approach may be interesting, but that such an approach is simply *not sustainable* for much longer, because advances in computing hardware are hitting limits and therefore will not be able to keep pace with the sharply increasing computational demands of this paradigm.¹⁸⁸ Yet on the other hand, it is unclear whether, how soon, or for whom this would pose an obstacle: the computing costs of training GPT-3 (~\$4.6 million) may have been unprecedented from the perspective of AI research, but are relatively modest from the perspective of the resources which other principals such as states have regularly dedicated to other scientific projects,¹⁸⁹ or especially to strategically salient technologies. To give one comparison, the full training cost of GPT-3 would cover less than four minutes of running the U.S. military.¹⁹⁰ At any rate, other work has shown it is possible to produce more specific but much smaller language models which can reach performance competitive with that of GPT-3.¹⁹¹

Nonetheless, there are compelling intrinsic reasons why systems such as GPT-3 would not likely master all human language tasks. For instance, such systems are not grounded in other domains of experience (they are not embodied, and lack physical interaction), and their objective function may be too limited.¹⁹² As a result, sceptics have objected that where it comes to recreating ‘human intelligence’, no core breakthroughs have been achieved in GPT-2 or GPT-3, suggesting, for instance, that such systems only *appear* to produce coherent text, but do not have a true understanding of what they are generating, as shown by occasional idiosyncratic mistakes.¹⁹³

move the ball forward on the research side of the equation. Rather, they’re prestige projects that simply demonstrate the scalability of existing techniques.”).

¹⁸⁸ Thompson et al., *supra* note 122. Although this evaluation appears to omit the findings of Kaplan et al., *supra* note 185. Note, on avenues for increased performance, see also Charles E. Leiserson et al., *There’s plenty of room at the Top: What will drive computer performance after Moore’s law?*, 368 SCIENCE (2020), <https://science.sciencemag.org/content/368/6495/eaam9744> (last visited Jul 13, 2020). More generally, because language models are so computationally expensive, substantial effort is being put into improving efficiency. I thank Ross Gruetzmacher for this point.

¹⁸⁹ Branwen, *supra* note 181; Levin and Maas, *supra* note 162 at 2.

¹⁹⁰ Author’s calculations, using for reference the US military budget for FY2020 (\$738 billion). Joe Gould, *Pentagon finally gets its 2020 budget from Congress*, DEFENSE NEWS (2019), <https://www.defensenews.com/congress/2019/12/19/pentagon-finally-gets-its-2020-budget-from-congress/> (last visited Sep 26, 2020). In another comparison, the costs of training GPT-3 are equivalent to that of ten helmets for F-35 fighter pilots (at \$400,000 apiece). Roger Mola, *Super Helmet*, AIR & SPACE MAGAZINE (2017), <https://www.airspacemag.com/military-aviation/super-helmet-180964342/> (last visited Jun 26, 2020).

¹⁹¹ Timo Schick & Hinrich Schütze, *It’s Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners*, ARXIV200907118 Cs (2020), <http://arxiv.org/abs/2009.07118> (last visited Sep 21, 2020). (displaying ALBERT, which achieved competitive results on one language benchmark using only 0.1% of GPT-3’s parameters).

¹⁹² As acknowledged by the GPT-3 team; Brown et al., *supra* note 44 at 33–34. I also thank Miles Brundage for clarifying points.

¹⁹³ Gary Marcus, *GPT-2 and the Nature of Intelligence*, THE GRADIENT (2020), <https://thegradient.pub/gpt2-and-the-nature-of-intelligence/> (last visited Feb 3, 2020). Gary Marcus & Ernest Davis, *GPT-3, Bloviator: OpenAI’s language generator has no idea what it’s talking about*, MIT TECHNOLOGY REVIEW, 2020, <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/> (last visited Sep 21, 2020).

2.1.6.3.1 Philosophical debates: the possible anthropocentrism of ‘general intelligence’

Such objections may well be correct, although that need not make such advances less philosophically interesting.¹⁹⁴ Indeed, rather than necessarily showing progress towards human-like AI, future advances might further put pressure on our prevailing notions of whether, how, or in what sense any intelligence (even human intelligence) can be said to be ‘general’: it could reveal a degree of the anthropocentrism in assuming that ‘general intelligence’—or functional intelligence—must necessarily be identical to ‘human intelligence’.¹⁹⁵

In doing so, such progress could refract our understanding of ‘intelligence’, just as that concept—along with other overburdened suitcase terms such as ‘sentience’ or ‘consciousness’—has already been put under pressure in recent years through advances in various biosciences. Even in nature, there may not be easy or rigid links between ‘intelligence’, ‘sentience’, and ‘consciousness’. At the very least, such qualities cannot be clearly and unambiguously determined through many tests we possess. For instance, as far as we can test, ant colonies pass many of the ‘prerequisites’ for conscious awareness met by humans;¹⁹⁶ cleaner wrasse fish can pass all the stages of the ‘mirror self-recognition’ test.¹⁹⁷ Parrots, monkeys, and bees all appear to possess the mathematical concept of ‘zero’.¹⁹⁸ Moreover, animals can exceed humans at specific tasks or tests: chimpanzees may have a better working memory for visual information, enabling them to beat humans at various numerical recollection tasks or strategic games;¹⁹⁹ and gray parrots in test outperformed Harvard undergraduates in complex versions of a shell memorisation game.²⁰⁰ Finally, the acellular slime mould *Physarum polycephalum*, an organism that lacks a brain or central nervous system, can display complex behaviour at a level where “it is capable of finding the shortest path through a maze, it can construct networks as efficient as those designed by

¹⁹⁴ For a recent discussion of GPT-3 by philosophers (with a ‘response’ by GPT-3), see Justin Weinberg, *Philosophers On GPT-3 (updated with replies by GPT-3)*, DAILY NOUS (2020), <http://dailynous.com/2020/07/30/philosophers-gpt-3/> (last visited Aug 10, 2020).

¹⁹⁵ Indeed, as John McGinnis has once noted, “confusing the proposition that AI may soon gain human capabilities with the proposition that AI may soon partake of human nature is the single greatest systemic mistake made in thinking about computational intelligence.” John O. McGinnis, *Accelerating AI*, 104 NORTHWEST. UNIV. LAW REV., 369 (2010), https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1193&context=nulr_online (last visited Dec 7, 2018).

¹⁹⁶ Daniel A. Friedman & Eirik Søvik, *The ant colony as a test for scientific theories of consciousness*, SYNTHESE (2019), <https://doi.org/10.1007/s11229-019-02130-y> (last visited Feb 24, 2019).

¹⁹⁷ Masanori Kohda et al., *Cleaner wrasse pass the mark test. What are the implications for consciousness and self-awareness testing in animals?*, BIORXIV 397067 (2018); Though some have taken such results to be more revealing of the limits of the test: Elizabeth Preston, *A ‘Self-Aware’ Fish Raises Doubts About a Cognitive Test*, QUANTA MAGAZINE, 2018, <https://www.quantamagazine.org/a-self-aware-fish-raises-doubts-about-a-cognitive-test-2018-1212/> (last visited Mar 9, 2019).

¹⁹⁸ Scarlett R. Howard et al., *Numerical ordering of zero in honey bees*, 360 SCIENCE 1124–1126 (2018); Andreas Nieder, *Honey bees zero in on the empty set*, 360 SCIENCE 1069–1070 (2018).

¹⁹⁹ Sana Inoue & Tetsuro Matsuzawa, *Working memory of numerals in chimpanzees*, 17 CURR. BIOL. R1004–R1005 (2007); Christopher Flynn Martin et al., *Chimpanzee choice rates in competitive games match equilibrium game theory predictions*, 4 SCI. REP. 5182 (2014); For an overview, see Justin Gregg, *Is your toddler really smarter than a chimpanzee?*, BBC EARTH, October 12, 2014, <http://www.bbc.com/earth/story/20141012-are-toddlers-smarter-than-chimps>.

²⁰⁰ Juan Siliezar, *African grey parrot outperforms children and college students*, HARVARD GAZETTE (2020), <https://news.harvard.edu/gazette/story/2020/07/african-grey-parrot-outperforms-children-and-college-students/> (last visited Jul 13, 2020).

humans, it can solve computationally difficult puzzles, it makes multi-objective foraging decisions, it balances its nutrient intake and it even behaves irrationally.”²⁰¹

Of course, this need not suggest that these animals have human-‘level’ intelligence or awareness, nor does it suggest that they categorically lack these properties. The point is rather that such cases show how the range and space of ‘intelligence’—of ‘possible minds’, or even of ‘cognitive processes’—might be more complex and diverse than a focus on ‘human intelligence’ as the type specimen or core set example suggests.²⁰² Even today many behaviours of AI systems—including the areas in which they perform well and where they are brittle—frustrate easy comparison with other natural (human or other) kinds of intelligence, exacerbating conceptual confusion.²⁰³ That has often been taken as evidence that these techniques, even if they are showing increasing generality, must be dead-ends that are not (by themselves) on the path to ‘human-like’ intelligence. That may be true, but it might also induce caution about assuming that all ‘intelligent processes’ must ‘scale up’ in the same way, or converge towards a type of general intelligence that is completely equivalent to human intelligence.²⁰⁴

2.1.6.3.2 Practical implications of progress

Putting these deep and (currently) intractable philosophical debates aside, the practical point from the perspective of governance is more about whether, or to what degree, we might expect increasing *disruption*. Ultimately, it may be possible or even likely that even high-profile systems such as GPT-3 reflect only limited domain progress—or even progress towards a ‘dead end’, if that end is only narrowly defined as achieving ‘AGI’. Nonetheless, from a practical perspective, some of these objections may miss the mark: which is that if, over the coming years, certain AI systems can demonstrate increasingly adequately broad and adequately good performance across still-constrained but functional task domains, then this may well suffice to make these technologies highly disruptive—even if they never ‘truly’ reach human intelligence and ‘merely’ asymptotically approach its functional task performance.

The point of this discussion is not about the merits or limits of specific AI systems (such as GPT-3), nor to adjudicate these long-standing scientific debates. It is simply to suggest three things: (1) it is plausible to expect that AI capability progress in many domains may continue to improve for at least the coming years, in some cases suddenly or with jumps that come with little warning; (2) while we should be cautious of a naïve ‘extrapolation of exponentials’, exponential trends and phenomena do exist across natural and man-made systems, and even if they eventually level off, they can have quite disruptive effects; (3) even if the resulting improvements

²⁰¹ Madeleine Beekman & Tanya Latty, *Brainless but Multi-Headed: Decision Making by the Acellular Slime Mould *Physarum polycephalum**, 427 J. MOL. BIOL. 3734–3743 (2015).

²⁰² Bhatnagar et al., *supra* note 8; Murray Shanahan, *Beyond humans, what other kinds of minds might be out there?*, AEON (2016), <https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there> (last visited May 20, 2018).

²⁰³ See also Wolpe, *supra* note 64 at 75; Arleen Salles, Kathinka Evers & Michele Farisco, *Anthropomorphism in AI*, 11 AJOB NEUROSCI. 88–95, 88 (2020).

²⁰⁴ To draw another analogy: the food industry today remains unable to assemble an edible bread ‘from the ground up’ (molecule by molecule), and current synthetic organic food substances remains complex and economically unfeasible (beyond low-molecular-weight compounds such as vitamins and amino acids). Nonetheless, just because it cannot ‘synthesize bread’ does not mean that synthetic organic chemistry is categorically incapable of producing food substances that are, functionally speaking, as edible and nourishing as naturally prepared bread.

in AI are not considered qualitatively significant from a scientific or philosophical perspective, certain AI systems can still be qualitatively significant from a sociotechnical perspective, in terms of the capabilities they enable. They can be disruptive to society, and may require governance.

2.1.6.4 *On expert prediction and governing under uncertainty*

Both an outside view and an inside view can offer interesting considerations for how to approach the debates around AI capability ceilings and rates of further progress. However—unsurprisingly given the level of extant disagreement and debate—neither approach provides definitive conclusions. A third source of consideration, then, is to suspend either object-level or referent-level analysis, and simply ‘defer to the experts’.

Given this, what should we make of expert forecasts that ‘High-Level Machine Intelligence’ or ‘AGI’ might be a few decades away? One question is how much stock can be put by such predictions, given deep methodological challenges about conducting long-range forecasts. For one, the track record of past predictions around AI is chequered at best.²⁰⁵ Moreover, even expert surveys are notoriously sensitive to slight differences in phrasing.²⁰⁶ Finally, as noted, there is prevalent disagreement amongst many AI scientists themselves.

More fundamentally, this may be an area where even domain experts find it hard to anticipate (or conversely, rule out) sudden progress or developments *ex ante*.²⁰⁷ This may especially be the case because many of the contexts that have been found to be critical for developing reliable and useful forecasting expertise (such as the ability for experts to get rapid feedback) are absent in long-range forecasting of A(G)I scenarios.²⁰⁸ Indeed, if it turns out to be the case that future advanced AI requires fundamentally new insights, it is even possible that there are today ‘no relevant experts’ on future AI architectures, such that expert elicitation processes cannot, by definition, generate accurate or well-formed predictions.²⁰⁹

However, the fact that accurate predictions are difficult might be less ground for comfort than for caution. Articulating better roadmaps of progress and ‘early warning signs’ is important

²⁰⁵ Indeed, many past predictions have proven wildly unreliable; Stuart Armstrong & Kaj Sotala, *How We’re Predicting AI--Or Failing To*, in BEYOND AI: ARTIFICIAL DREAMS 52–75 (Jan Romportl et al. eds., 2012), <https://intelligence.org/files/PredictingAI.pdf>; Stuart Armstrong, Kaj Sotala & Sean S. OhEigearthaigh, *The errors, insights and lessons of famous AI predictions – and what they mean for the future*, 26 J. EXP. THEOR. ARTIF. INTELL. (2014), <http://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf> (last visited Feb 18, 2017).

²⁰⁶ For instance, surveys have found significant differences on reported timelines, based on phrasing. Grace et al., *supra* note 131 at 731.

²⁰⁷ On the specific case of the difficulty of predicting AI progress even for AI experts, cf. Yudkowsky, *supra* note 168.

²⁰⁸ Armstrong and Sotala, *supra* note 205 at 52–75.

²⁰⁹ That is, there may be a category error here, where today’s AI experts are strictly speaking experts on contemporary machine learning approaches, which, it is hoped, will play a role in future advanced AI. On the problem of expert judgment in conditions where there may not be a relevant reference class of experts, see generally M. Granger Morgan, *Use (and abuse) of expert elicitation in support of decision making for public policy*, 111 PROC. NATL. ACAD. SCI. 7176–7184 (2014). On the other hand, as noted by Sotala and Yampolskiy, “[t]he lack of expert agreement suggests that expertise in the field does not contribute to an ability to make reliable predictions. If the judgment of experts is not reliable, then, probably, neither is anyone else’s. This suggests that it is unjustified to be highly certain of AGI being near, but also of it not being near.” R.V. Yampolskiy & K. Sotala, *Responses to catastrophic AGI risk: A survey*, 90 PHYS. SCR., 8 (2015), <https://intelligence.org/files/ResponsesAGIRisk.pdf>.

not only for AGI, but also for far less sophisticated but nonetheless disruptive AI capabilities,²¹⁰ as these may need early policy action to avoid the lock-in of interests or governance gridlock. Indeed, an improved understanding of the technical landscape of AI over the coming years is of relevance to setting societal choices and policies,²¹¹ even if one is confident that the resulting map will not include ‘human-level’ AI or AGI on it. Improving forecasting methodologies in order to predict the distinct trajectories in AI’s development and use conditions (investment; data; hardware, talent), domain capabilities (e.g. cyberwarfare; text generation) and use cases and impact over the coming decades, is difficult and remains an active research area.²¹²

Given the above, it is important to understand the ‘epistemic situation’ we find ourselves in.²¹³ Assuming a pessimistic stance regarding AI which anticipates minor future progress may be appealing (and at times warranted) as a means to balance some of the extreme hype around AI technology. However, ultimately it may be an imperfect heuristic. What debate in this space needs is perhaps less categorical statements of impossibility or inevitability, and more good faith attempts to wrestle with the best available information and estimates, under admitted circumstances of deep epistemic uncertainty on many outstanding questions.

Along with attempting to increase our certainty, we may also have to engage frankly with questions of appropriate (or precautionary) epistemic strategy. It may well be impossible to be perfectly calibrated on AI futures. Given that, what would be the differential costs of different prediction errors? Regardless of one’s expectations about AI development and deployment timelines, what might be the costs of excessive optimism or excessive scepticism—of over-preparation or under-preparation?²¹⁴ These are not simple questions either, but they may be key in reaching more nuanced debate around the considerations and estimates that could or should inform societal debate around its relation to AI.

²¹⁰ For one discussion of warning signs, see also Carla Zoe Cremer & Jess Whittlestone, *Canaries in Technology Mines: Warning Signs of Transformative Progress in AI* (2020). Their reference to ‘canaries’ refers to the set of criteria in Oren Etzioni, *How to know if artificial intelligence is about to destroy civilization*, MIT TECHNOLOGY REVIEW, 2020, <https://www.technologyreview.com/2020/02/25/906083/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/> (last visited Jun 23, 2020).

²¹¹ DAFOE, *supra* note 130 at 15–33.

²¹² Miles Brundage, *Modeling Progress in AI*, ARXIV151205849 Cs (2015), <http://arxiv.org/abs/1512.05849> (last visited Apr 30, 2017); DAFOE, *supra* note 130. This is an active research area. For a recent research agenda, see Ross Gruetzmacher et al., *Forecasting AI Progress: A Research Agenda*, ARXIV200801848 Cs (2020), <http://arxiv.org/abs/2008.01848> (last visited Aug 11, 2020). For a survey of methodologies, see Shahar Avin, *Exploring Artificial Intelligence Futures*, AIHUMANITIES (2019), http://aihumanities.org/en/journals/journal-of-aih-list/?board_name=Enjournal&order_by=fn_pid&order_type=desc&list_type=list&vid=15 (last visited Feb 25, 2019). It should be noted that for specific technological subsets (such as defence technology), forecasts have reportedly achieved reasonable accuracy (which might be explained because of the very long procurement timelines for major weapons systems, and strong strategic pressures on systems’ marginal performance). Alexander Kott & Philip Perconti, *Long-Term Forecasts of Military Technologies for a 20-30 Year Horizon: An Empirical Assessment of Accuracy*, ARXIV180708339 Cs (2018), <http://arxiv.org/abs/1807.08339> (last visited Dec 18, 2018).

²¹³ See Adrian Currie, *Existential Risk, Creativity & Well-Adapted Science*, 76 in STUDIES IN THE HISTORY & PHILOSOPHY OF SCIENCE. 39–48 (2019), <https://www.sciencedirect.com/science/article/abs/pii/S0039368117303278?via%3Dihub>; Adrian Currie & Shahar Avin, *Method Pluralism, Method Mismatch, & Method Bias*, 19 PHILOS. IMPR. (2019), <http://hdl.handle.net/2027/spo.3521354.0019.013> (last visited Apr 9, 2019).

²¹⁴ I thank Miles Brundage for good points on this.

Such debate also matters practically, because there might be many scientific and policy foundations that would need to be established today, given important path-dependencies.²¹⁵ As Kunz & Ó hÉigearthaigh note, “[i]t is still highly uncertain if and when general intelligence comparable to that of humans might be achieved in artificial systems, but taking into account the magnitude of the consequences and a certain degree of inertia and path dependency of international governance, it is worth considering potential risks and mitigation strategies early on.”²¹⁶ Rather than seeing conflict, there might be important areas for common cause.²¹⁷

More generally, there is also a plausible argument to be made that governance will increasingly have to engage proactively rather than ‘wait out’ developments in the study on the emerging risks of new technologies. Indeed, in the face of diverse technological developments, from synthetic biotechnology to geo-engineering, there has been an extensive and emerging literature emphasizing the importance of ‘Responsible Innovation’ for technology governance, which includes the principle of appropriate ‘anticipation’.²¹⁸ Some have argued that such principles should be considered beyond national technology regulation, and should also increasingly inform global governance. For instance, Rosemary Rayfuse notes that whereas the international law system has generally been reactive in regulating new technologies, in the face of emerging technologies such as AI, “international law is being called upon to regulate not just the past and present development and deployment of technologies, but also the uncertain futures these technologies pose.”²¹⁹ The drive towards anticipation in AI governance—while acknowledging the conditions of uncertainty—could therefore align with a broader shift in the orientation of global governance towards uncertain but high-stakes technological developments more broadly.

2.1.7 Change from Algorithmic Proliferation: disruptive AI is already here

So far, we explored the arguments and prospects around future trajectories of AI. However, one does not need to assume further scientific or technological developments in order to argue that AI technology will be of far-reaching and growing importance in the coming years.

²¹⁵ Stephen Cave & Seán S. Ó hÉigearthaigh, *Bridging near- and long-term concerns about AI*, 1 NAT. MACH. INTELL. 5 (2019).

²¹⁶ Martina Kunz & Seán Ó hÉigearthaigh, *Artificial Intelligence and Robotization*, in OXFORD HANDBOOK ON THE INTERNATIONAL LAW OF GLOBAL SECURITY, 14 (Robin Geiss & Nils Melzer eds., 2020), <https://papers.ssrn.com/abstract=3310421> (last visited Jan 30, 2019).

²¹⁷ Seth D. Baum, *Reconciliation between factions focused on near-term and long-term artificial intelligence*, 33 AI SOC. 565–572 (2018); Cave and Ó hÉigearthaigh, *supra* note 215; Carina Prunkl & Jess Whittlestone, *Beyond Near-and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 138–143 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375803> (last visited Feb 12, 2020).

²¹⁸ See generally: Richard Owen et al., *A Framework for Responsible Innovation*, in RESPONSIBLE INNOVATION 27–50 (2013), <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118551424.ch2> (last visited Aug 21, 2019); Jack Stilgoe, Richard Owen & Phil Macnaghten, *Developing a framework for responsible innovation*, 42 RES. POLICY 1568–1580 (2013); And specific to AI, see: Miles Brundage, *Artificial Intelligence and Responsible Innovation*, in FUNDAMENTAL ISSUES OF ARTIFICIAL INTELLIGENCE 543–554 (Vincent C. Müller ed., 2016), http://link.springer.com/chapter/10.1007/978-3-319-26485-1_32 (last visited Mar 4, 2017).

²¹⁹ Rosemary Rayfuse, *Public International Law and the Regulation of Emerging Technologies*, in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY, 500–501 (2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22> (last visited Jan 3, 2019).

Indeed, even if one were to take a pessimistic (or, depending on one's views, the optimistic) perspective, and expect that fundamental AI progress were to slow down or halt soon, the state of existing technology is such that it will have considerable societal impact in the coming years.

This is for a range of reasons: (2.1.7.1.) input thresholds and barriers to deployment are continuing to fall; (2.1.7.2.) the technology has broad usability as a general-purpose technology; (2.1.7.3.) models indicate AI could see relatively rapid global (horizontal) diffusion; (2.1.7.4.) it could also see rapid domestic (vertical) diffusion, because 'infrastructure overhang' can facilitate quick deployment rates of AI capabilities into pre-existing architectures or platforms across the 'digital lifeworld.'²²⁰

2.1.7.1 Falling input thresholds and barriers to deployment

In the first place, many of the previously discussed practical barriers to deployment of existing AI technologies are on track to fall steadily. As these existing AI capabilities mature and proliferate, they are already likely to see increasing use.

For one, *training data requirements* may increasingly be achievable for many parties. On the one hand, this is because of the increasing *supply* of data, as a product of the growing ubiquity of data-gathering devices and sensors. For instance, Microsoft has previously predicted that by 2025, 4.7 billion people (just over half the world's expected population) will access the internet through an estimated 50 billion connected devices.²²¹ Of course, this trend will—and arguably should—be subject to global and regional developments and trends in data (privacy) regulation.

At the same time, further increases in data efficiency can result in falling data *demands* or needs of AI systems. This may be because unsupervised and semi-supervised learning approaches become more prominent drivers of progress, given that it is easier and cheaper to obtain large unlabelled data for these systems, and only small labelled datasets may be necessary to fine-tune pre-trained models.²²² This may be complemented by advances in the areas of one-shot learning,²²³ simulation transfer,²²⁴ approaches that allow principals to complement real training data with AI-generated 'synthetic data',²²⁵ or which are able to work with messy, continuous, or irregularly measured datasets.²²⁶ Such increases in data efficiency may carry

²²⁰ JAMIE SUSSKIND, *FUTURE POLITICS: LIVING TOGETHER IN A WORLD TRANSFORMED BY TECH* (2018).

²²¹ DAVID BURT ET AL., *Cyberspace 2025: Today's Decisions, Tomorrow's Terrain* (2014), <https://www.microsoft.com/en-us/cybersecurity/content-hub/cyberspace-2025> (last visited Jul 3, 2020).

²²² I thank Ross Gruetzmacher for this point.

²²³ Natalie Ram, *One Shot Learning In AI Innovation*, AI PULSE (2019), <https://aipulse.org/one-shot-learning-in-ai-innovation/> (last visited Feb 26, 2019).

²²⁴ Compare the OpenAI 'Dactyl' robotic hand, which was trained to solve real-world tasks in sets of randomised simulations, without physically faithful modelling of the world. OpenAI et al., *Learning Dexterous In-Hand Manipulation*, ARXIV180800177 Cs STAT (2018), <http://arxiv.org/abs/1808.00177> (last visited Jan 11, 2019).

²²⁵ This functionally allows arbitrage of compute for data. One example here is 'RarePlanes', a 'machine learning dataset that incorporates both real and synthetically generated satellite imagery' to enable the detection of aircraft. Jacob Shermer et al., *RarePlanes: Synthetic Data Takes Flight*, ARXIV200602963 Cs (2020), <http://arxiv.org/abs/2006.02963> (last visited Jun 8, 2020). Another example is Nvidia's synthetic dataset for self-driving cars. Sivabalan Manivasagam et al., *LiDARsim: Realistic LiDAR Simulation by Leveraging the Real World*, ARXIV200609348 Cs (2020), <http://arxiv.org/abs/2006.09348> (last visited Jun 22, 2020).

²²⁶ See for instance Ricky T. Q. Chen et al., *Neural Ordinary Differential Equations*, ARXIV180607366 Cs STAT (2018), <http://arxiv.org/abs/1806.07366> (last visited Jan 11, 2019). Karen Hao, *A radical new neural network*

considerable societal impacts, potentially shifting competitive market dynamics, exacerbating surveillance and privacy concerns (because the marginal value of data increases), or lowering the threshold to access and potential misuse by many actors.²²⁷

Moreover, *computing hardware* demands must one day prove a fundamental physical constraint,²²⁸ but for the foreseeable future, the costs of computing will continue to fall.²²⁹ In spite of continuous warnings of impending limits, the 18-month doubling-time of Moore's Law appears to have held to date.²³⁰ It will certainly not do so forever, but ongoing developments in materials science may help postpone that point.²³¹ Moreover, improvements in end-to-end system designs suggest AI performance could scale even if semiconductors do not.²³² Of course, such advances may not be able to keep pace with the increasing compute demands of peak AI projects, but they may at least lower the threshold for applying existing systems.

More to the point, while in the long run hardware limits may cap progress—especially given the even-more-extreme growth in compute used by top AI systems—this may not inhibit usage in many practical and disruptive applications. Once trained, the amount of computing power necessary for AI deployment is far lower.²³³ Not all useful AI capabilities come with prohibitively restrictive hardware requirements. For instance, the 2016 'ALPHA' AI developed by the University of Cincinnati proved able to defeat expert U.S. Air Force tacticians in simulated aerial combat, using no more processing power than that afforded by a small, \$60 'Raspberry Pi' computer.²³⁴ While algorithmic improvements are harder to measure, such achievements have

design could overcome big challenges in AI, MIT TECHNOLOGY REVIEW, 2018, <https://www.technologyreview.com/s/612561/a-radical-new-neural-network-design-could-overcome-big-challenges-in-ai/> (last visited Jan 11, 2019).

²²⁷ Aaron D. Tucker, Markus Anderljung & Allan Dafoe, *Social and Governance Implications of Improved Data Efficiency*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 378–384 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375863> (last visited Feb 12, 2020).

²²⁸ Hwang, *supra* note 94.

²²⁹ For trends over recent years (using floating point operations per second (FLOPS) as metric of computational power), see also Asya Bergal, *2019 recent trends in GPU price per FLOPS*, AI IMPACTS (2020), <https://aiimpacts.org/2019-recent-trends-in-gpu-price-per-flops/> (last visited Sep 27, 2020). (estimating that in recent years, GPU (Graphics Processing Units) prices have fallen at rates that would yield an order of magnitude over 17 years (single-precision FLOPS), 10 years (half-precision FLOPS), and 5 years (for half-precision fused multiply-add FLOPS). For a recent detailed examination of how, in pure computational terms (leaving out questions of software), such trends would compare to the potential thresholds of computational power necessary to match the human brain, see JOSEPH CARLSMITH, *How Much Computational Power Does It Take to Match the Human Brain?* (2020), <https://www.openphilanthropy.org/brain-computation-report> (last visited Sep 20, 2020).

²³⁰ Hassan N. Khan, David A. Hounshell & Erica R. H. Fuchs, *Science and research policy at the end of Moore's law*, 1 NAT. ELECTRON. 14 (2018). There are arguments however that the move towards specialised AI computing hardware may 'fragment' the field of computing progress; NEIL THOMPSON & SVENJA SPANUTH, *The Decline of Computers As a General Purpose Technology: Why Deep Learning and the End of Moore's Law are Fragmenting Computing* (2018), <https://papers.ssrn.com/abstract=3287769> (last visited Jul 13, 2020).

²³¹ The Economist, *The incredible shrinking machine - A new material helps transistors become vanishingly small*, THE ECONOMIST (2020), <https://www.economist.com/science-and-technology/2020/07/18/a-new-material-helps-transistors-become-vanishingly-small> (last visited Aug 19, 2020).

²³² Leiserson et al., *supra* note 188.

²³³ For instance, while the GPT-3 system may have cost around \$4.6 million in compute (at today's prices) to train, its subsequent use is not expensive; as the authors note, even with the full model, "generating 100 pages of content from a trained model can cost on the order of 0.4 kW-hr, or only a few cents in energy costs." Brown et al., *supra* note 44 at 38.

²³⁴ Nicholas Ernest et al., *Genetic Fuzzy based Artificial Intelligence for Unmanned Combat Aerial Vehicle Control in Simulated Air Combat Missions*, 06 J. DEF. MANAG. (2016), <http://www.omicsgroup.org/journals/genetic-fuzzy->

been had over the past decade, delivering a 44-fold compute efficiency gain over the period 2012-2019, relative to an 11-fold efficiency gain expected through Moore's Law over the same period.²³⁵ Finally, in some contexts, AI techniques themselves can support in the accelerated development of computing hardware,²³⁶ as well as the software configuration of machine learning itself.²³⁷

Thirdly, in terms of *human AI expertise* or talent, the labour supply of qualified AI researchers may become less constrained over time. High industry recruitment these past few years, along with state investment, may contribute to a pipeline of upcoming students. As such, demand for talent can lead to increased supply. In addition to this, there are ongoing improvements in a range of AI development environments or software tools which automate or facilitate aspects of the machine learning tailoring process, making it easier for people to learn to work in the field, while also making individual AI researchers more 'productive'.²³⁸ These two trends may gradually reduce the 'expertise barrier' to acquiring the required computer science labour and talent to run or deploy certain capabilities.²³⁹

To be sure, the relative speeds of these trends in data, compute, and talent (availability and -requirements) may be contested; and certainly no trend lasts indefinitely. Nonetheless, for the foreseeable future, their direction appears to be towards increasing accessibility and proliferation of AI tools.

2.1.7.2 Broad usability and status as emerging General-Purpose-Technology

More than this, AI will likely have far-reaching global effects, as a result of its broad versatility. While there are many ways to segment and subdivide the functions of AI technology, generally speaking, present-day systems can perform one of several classes of functions: (i) *data classification* (supporting applications such as detecting if a skin photo shows cancer or not; facial recognition); (ii) *data generation* (e.g. synthesizing 'DeepFake' images, video, audio or other media); (iii) *anomaly or pattern detection* (e.g. detecting cybercrime or fraudulent financial transactions); (iv) *prediction* (e.g. of consumer preferences; weather patterns; loan performance); (v) *optimisation* of complex systems and tasks (e.g. optimizing energy usage in data centers); or

based-artificial-intelligence-for-unmanned-combat-aerialvehicle-control-in-simulated-air-combat-missions-2167-0374-1000144.php?aid=72227 (last visited Sep 19, 2016).

²³⁵ DANNY HERNANDEZ & TOM BROWN, *Measuring the Algorithmic Efficiency of Neural Networks* 20 (2020), https://cdn.openai.com/papers/ai_and_efficiency.pdf.

²³⁶ Azalia Mirhoseini et al., *Chip Placement with Deep Reinforcement Learning*, ARXIV200410746 CS (2020), <http://arxiv.org/abs/2004.10746> (last visited Apr 24, 2020).

²³⁷ See for instance Esteban Real et al., *AutoML-Zero: Evolving Machine Learning Algorithms From Scratch*, ARXIV200303384 CS STAT (2020), <http://arxiv.org/abs/2003.03384> (last visited Apr 20, 2020); Alexey Svyatkovskiy et al., *IntelliCode Compose: Code Generation Using Transformer*, ARXIV200508025 CS (2020), <http://arxiv.org/abs/2005.08025> (last visited Jun 6, 2020). Or Luke Metz et al., *Tasks, stability, architecture, and compute: Training more effective learned optimizers, and using them to train themselves*, ARXIV200911243 CS STAT (2020), <http://arxiv.org/abs/2009.11243> (last visited Sep 25, 2020).

²³⁸ On the impact of 'autoML' tools on AI talent acquisition processes, see also HOLGER HÜRTGEN ET AL., *Rethinking AI talent strategy as AutoML comes of age* (2020), <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/rethinking-ai-talent-strategy-as-automated-machine-learning-comes-of-age> (last visited Aug 17, 2020).

²³⁹ Belfield, *supra* note 87. See also Paul Mozur & Cade Metz, *A U.S. Secret Weapon in A.I.: Chinese Talent*, THE NEW YORK TIMES, June 9, 2020, <https://www.nytimes.com/2020/06/09/technology/china-ai-research-education.html> (last visited Jun 15, 2020). I thank Jess Whittlestone for prompting clarifications on these points.

(vi) *autonomous operation* of cyber-physical platforms (e.g. robots such as Roomba vacuum cleaners, social robots, or military drones).²⁴⁰

What is important to note about the above tasks is that, while the specific applications in question are individually narrow—such that an algorithm trained to detect cancers in skin photos would not transfer to use in facial recognition—many of these underlying capabilities are ‘domain-agnostic’. For instance, being able to automatically detect subtle anomalies or patterns in huge datasets is something that comes in useful in diverse contexts, from healthcare to law enforcement, and from advertising to traffic management. As previously discussed, it is this ‘general narrowness’ which has made this suite of AI approaches so generally applicable even today.

From a high-level perspective, this versatility contributes to the potential of AI systems to have an impact that is both broad, and (on aggregate) deep across many industries. This has led some to compare AI to past innovations like the steam engine, electricity, the internal combustion engine, or computers, which have been described as ‘General-Purpose Technologies’ (GPT) which significantly altered the trajectory of societies.²⁴¹ As originally theorised, GPTs “are characterised by pervasiveness, inherent potential for technical improvements, and ‘innovational complementarities’, giving rise to increasing returns-to-scale.”²⁴² This is because they are not only useful across many contexts, but also are ‘enabling’—they can unlock, spur or enable further technical improvements in many other sectors. For all its current warts and limits, AI technology has a strong claim to such features, and as such can be productively understood as an emerging, still-immature ‘General-Purpose Technology’.

Of course, just because various AI capabilities (prediction, anomaly detection, autonomous operation, etc.) are almost ‘universally useful’ in theory, does not mean that they will be ‘universally used’ in practice. Especially in the near term, there will be many contexts where a certain AI capability would be in-principle useful, but is simply not ‘worth it’ for any principal. In practical situations, a range of factors will play into determining the marginal added benefits of AI. Specifically, different domains may have different features where a certain (combination of) AI capabilities would add the most over human (or existing computer) performance, in terms of accuracy, speed, scale, or cost.

²⁴⁰ SCHARRE AND HOROWITZ, *supra* note 13. Though note, there are other ways to segment these functions. See for instance Corea, *supra* note 14.

²⁴¹ MANUEL TRAJTENBERG, *AI as the next GPT: a Political-Economy Perspective* (2018), <http://www.nber.org/papers/w24245> (last visited Oct 22, 2018). See also Gruetzmacher and Whittlestone, *supra* note 138. Jade Leung has argued that AI technology may also meet the criteria of a ‘strategic GPT’—by her definition, “a general purpose technology which has the potential to deliver vast economic value and substantially affect national security, and is consequently of central political interest to states, firms, and researchers.” Jade Leung, *Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies*, July, 2019, <https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665>. She compares the domestic politics around AI research with those around previous strategic GPTs, such as aerospace technology, biotechnology and cryptography.

²⁴² For the original discussion of the GPT-concept, see Timothy Bresnahan & Manuel Trajtenberg, *General purpose technologies “Engines of growth”?*, 65 J. ECONOM. 83–108, 83 (1995). There may be interesting interrelations between GPTs and infrastructure technologies. Christiaan Hogendorp & Brett Frischmann, *Infrastructure and general purpose technologies: a technology flow framework*, EUR. J. LAW ECON. (2020), <https://doi.org/10.1007/s10657-020-09642-w> (last visited Aug 10, 2020).

For example, take the capability of ‘autonomous operation’: the US Defense Science Board’s 2016 ‘Summer Study on Autonomy’ distinguished a series of task criteria which are correlated with higher value of autonomy in systems, at least in the military context. For instance, situations with *high required decision speed* could see autonomous systems contribute to cyber operations or missile defence; for situations involving *high heterogeneity & volume of data*, AI could add value in Intelligence, Surveillance, and Reconnaissance Data Integration; under conditions of *intermittent quality of data links*, autonomous systems can add value in unmanned undersea operations or in contested communication environments; robotic systems have value in situations of *high danger of mission* (e.g. CBRN attack clean-up), or to missions requiring *high persistence and endurance* (e.g. in unmanned vehicle surveillance).²⁴³ While these operational benefits might be particularly steep or critical in the competitive military realm, many of them may hold in other contexts. Considering these factors however, helps clarify the marginal added benefit of certain proposed AI applications, as well as the (absolute or relative, economic or military) pressures towards adoption. Even taking this caveat into consideration, it is clear that there are many contexts in which AI can appeal.

Moreover, while (computationally intensive) high-end applications of AI may be particularly visible, in many areas, the proliferation or commercialisation of relatively low-end applications (such as in facial recognition) can be more than sufficient to drive significant societal impacts.²⁴⁴ Importantly, widespread application may lag in-the-field breakthroughs by at least a number of years, suggesting that a certain degree of disruption is already ‘baked in’ even if underlying AI progress were to hit some wall.

2.1.7.3 Patterns and rates of global diffusion

Moreover, AI capabilities are likely to proliferate relatively fast across the world. There is an interesting argument here about the diffusion speed of historical general-purpose technologies, relative to their impact. Of course, technological diffusion is never predetermined or entirely frictionless. Horizontal transfers have been subject to constraints throughout history,²⁴⁵ even (or, especially) in cases where it concerned perceived ‘strategic assets’.²⁴⁶

Nonetheless, a wide range of technologies do eventually diffuse globally, under distinct patterns. As Daniel Drezner has noted, the rapidity with which new technologies disseminate across the international stage tends to depend on whether the technology’s production and deployment involves high or low fixed costs, and whether or not they have considerable civilian

²⁴³ DEFENSE SCIENCE BOARD, *Defense Science Board Summer Study on Autonomy* 12 (2016), <https://www.hsdl.org/?abstract&did=794641>. See Figure 4.

²⁴⁴ In the domain of (consumer) drone technology, this is illustrated, for instance, by the 2019 Houthi drone attack on Saudi oil refineries. James Rogers, *The dark side of our drone future*, BULLETIN OF THE ATOMIC SCIENTISTS (2019), <https://thebulletin.org/2019/10/the-dark-side-of-our-drone-future/> (last visited Oct 6, 2019).

²⁴⁵ See for instance Elizabeth Smithrosser, *How to get good horses in medieval China*, MEDIEVALISTS.NET (2020), <https://www.medievalists.net/2020/06/horses-medieval-china/> (last visited Jun 15, 2020). And for a more modern example, illustrating the limits of imitation, reverse engineering, and cyber espionage, see Andrea Gilli & Mauro Gilli, *Why China Has Not Caught Up Yet: Military-Technological Superiority and the Limits of Imitation, Reverse Engineering, and Cyber Espionage*, 43 INT. SECUR. 141–189 (2019).

²⁴⁶ See Jeffrey Ding & Allan Dafoe, *The Logic of Strategic Assets: From Oil to Artificial Intelligence*, ARXIV200103246 Cs ECON Q-FIN (2020), <http://arxiv.org/abs/2001.03246> (last visited Jan 15, 2020).

applications.²⁴⁷ In Drezner's analysis, this produces in four types of technological innovations: strategic tech, prestige tech, public tech, and general purpose tech (see Figure 2.1).

	<i>Public sector dominance</i>	<i>Private sector dominance</i>
<i>High fixed costs</i>	Prestige tech	Strategic tech
<i>Low fixed costs</i>	Public tech	General purpose tech

Figure 2.1. Drezner's typology of Technological Innovation.²⁴⁸

In this typology, '*strategic tech*' technologies have high fixed costs but significant civilian applications (e.g. civilian aircraft sector or 5G networks). They accordingly invite great power intervention to promote national champions and avoid dependence on foreign suppliers.²⁴⁹ '*Prestige tech*' innovations have large fixed costs and only limited civilian applications (e.g. nuclear weapons, Concorde, manned space exploration), and therefore tend to be developed predominantly by governments.²⁵⁰ '*Public tech*' innovations have low fixed costs but limited private-sector opportunities or interests (e.g. public health innovations, such as pre-pandemic vaccines research). This means that they require government investments to seize their public good qualities.²⁵¹ Finally, '*general purpose tech*' involves technologies with low fixed costs and large private-sector applications (e.g. drones or AI), meaning that private-sector activity predominantly drives the technology's direction of development.²⁵²

Drezner argues that these factors have implications for the speed of (horizontal) diffusion: “[t]he more that a technology approaches the general purpose category, the more quickly it will diffuse across the globe. The lower the fixed costs – whether material, organizational, or societal in nature – the more rapidly a technology should diffuse from leader to laggards.”²⁵³

What does that suggest for AI? Of course, AI is a diverse technology, and not all of its applications will meet the same criteria. To be sure, many if not most AI applications are characterised by low fixed costs and private sector dominance, which suggests they are '*general purpose tech*' that will see relatively rapid dissemination. On the other hand, there may be exceptions where, in spite of the private sector dominance, the fixed costs of AI are relatively high—given the increasing computational demands of training up top AI systems—producing AI

²⁴⁷ Reproduced from Daniel W Drezner, *Technological change and international relations*, 33 INT. RELAT. 286–303, 291–292 (2019).

²⁴⁸ *Id.* at 292.

²⁴⁹ *Id.* at 291.

²⁵⁰ Notably, the name '*prestige*' does not mean to suggest that these technologies solely serve symbolic functions. Indeed, in some cases such as with nuclear weapons, governments derive obvious (if controversial) benefits from pursuing them. However, Drezner argues they “function as prestige goods because of their costs.” *Id.* at 292.

²⁵¹ *Id.* at 291.

²⁵² *Id.* at 291–292.

²⁵³ *Id.* at 292.

applications that might be categorised as ‘strategic tech’, which suggest attempted export controls and intervention by major states.

There may be applications that will diffuse more slowly. As Ayoub and Payne have noted that “[the] ubiquitous access to advanced algorithms gives a misleading impression of the ease with which military relevant AI may proliferate between states”.²⁵⁴ Indeed, certain military applications of AI may in fact involve large fixed costs in terms of adjacent or substrate technologies (e.g. pre-existing fleets of long-range drone weapons delivery platforms for LAWS; satellite command networks) which are not conjured out of nothing. Alternatively, effective deployment may require large technical changes in to-be-adopted civilian technologies, as well as large organisational and cultural changes in the adopting (military) organisation.²⁵⁵ Finally, notwithstanding a growing trend in the militarisation of domestic policing practices,²⁵⁶ many of these weapons also have little civilian application, suggesting that LAWS might be most akin to ‘prestige tech’, which might proliferate relatively slowly.²⁵⁷

However, even in such cases, many of the underlying AI capabilities which LAWS leverage (e.g. pattern recognition; autonomous operation) could still qualify as ‘general purpose tech’ with low fixed costs and considerable civilian applications. The question therefore may be about the application under consideration. However, this also suggests that, while there are some domains in which the diffusion of AI capabilities may be slow, the bottleneck to its spread may be less in AI capabilities themselves, and more in substrate technologies (e.g. delivery systems) or computing hardware.²⁵⁸

2.1.7.4 Roll-out rates and infrastructure overhang

This brings us to an additional point, which is that if the proliferation of AI capabilities is in most cases bottlenecked primarily by the availability of substrate technologies, then in certain cases we might expect it to be not very constrained at all. This point concerns ‘infrastructure overhang’: the broad availability of pre-existing supporting infrastructure.²⁵⁹ For instance, even

²⁵⁴ Kareem Ayoub & Kenneth Payne, *Strategy in the Age of Artificial Intelligence*, 39 J. STRATEG. STUD. 793–819, 809 (2016); As cited in Matthijs M. Maas, *How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons*, 40 CONTEMP. SECUR. POLICY 285–311, 290 (2019). Paper [II].

²⁵⁵ Michael C. Horowitz, *Artificial Intelligence, International Competition, and the Balance of Power*, TEXAS NATIONAL SECURITY REVIEW, 2018, <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/> (last visited May 17, 2018).

²⁵⁶ For instance, Arthur Holland Michel has chronicled how the ‘Gorgon Stare’ Wide-Area Motion Imagery (WAMI) surveillance system, which was developed to identify improvised explosive devices in Iraq, was deployed in Baltimore in 2016 to support investigative law enforcement operations. ARTHUR HOLLAND MICHEL, *EYES IN THE SKY: THE SECRET RISE OF GORGON STARE AND HOW IT WILL WATCH US ALL* (2019), <https://www.hmhbooks.com/shop/books/Eyes-in-the-Sky/9780544972001> (last visited Jun 24, 2019).

²⁵⁷ On the other hand, similar arguments apply for (tele-operated) military drones. And there can be little doubt that this technology has proliferated extensively, with more than 17 countries now operating armed drones, and some 100 countries running some kind of military drone program DAN GETTINGER, *The Drone Databook* (2019), <https://dronecenter.bard.edu/files/2019/10/CSD-Drone-Databook-Web.pdf>. Nonetheless, there has long been extensive debate over the exact drivers and rate of drone proliferation. Michael C. Horowitz, Sarah E. Kreps & Matthew Fuhrmann, *Separating Fact from Fiction in the Debate over Drone Proliferation*, 41 INT. SECUR. 7–42 (2016).

²⁵⁸ See also the discussion of ‘levers’ for AI governance, under section 2.2.3.2.

²⁵⁹ It should be noted that this notion of ‘infrastructure overhang’ is slightly distinct from the idea of ‘hardware overhang’—the situation “when human-level software is created, enough computing power may already be

if one accepts the comparison of AI with other ‘General-Purpose-Technologies’ such as electricity, that would only suggest that its societal impact will eventually be large, but not that this process will be particular rapid or come in the near future. After all, electrification took many decades to be fully adopted, even in the country that invented it first. A sceptic might therefore argue that even if AI is advancing fast in lab settings, like electricity it will take many decades before it is fully rolled out across society. However, there are two counter-arguments to this.

Firstly, over the last centuries, many broad-usage technologies have seen a steady decrease in their ‘time to pervasiveness’. For instance, Indermit Gill has noted how historically, “the time between invention and widespread use was cut from about 80 years for the steam engine to 40 years for electricity, and then to about 20 years for IT”.²⁶⁰ This pattern holds not just amongst countries, but also in terms of domestic market penetration: while it took over thirty years for a quarter of all US homes to get a telephone, it took only seven years for a similar fraction to receive internet access, with tablets and smartphones having experienced even faster rates of adoption and diffusion.²⁶¹ In some regions, technologies such as mobile phones entirely and comprehensively leapfrogged older legacy systems (e.g. landlines).²⁶² If this historical pattern holds, we might expect the pervasive rollout of AI technology to occur much faster than electrification—at least as fast as the computing revolution in its time.

Secondly, there is a reason for expecting even faster diffusion and use. Critically, whereas electrification (or even the establishment of electric cars) required the up-front rollout of physical infrastructure, AI does not face these same problems. Indeed, the marginal deployment costs of (already-trained) AI systems, and with them the barriers to AI diffusion might be lower than they were for these other technologies, given that in many cases supporting digital infrastructures have already been independently developed or established. For instance, for all that AI companies at times face challenges in procuring clean and adequate datasets, it is also not the case that they have to always start from scratch. Rather, states have long sought to make the world more

available to run vast numbers of copies at great speed” BOSTROM, *supra* note 146 at 73. More generally, this reflects situations where the societal impact of even seemingly small algorithmic improvements may prove sudden and extreme, because existing computing hardware is more than adequate to rapidly scale them. The notion of ‘infrastructure overhang’ is similar, but points more to the idea that algorithmic improvements can benefit (in both training- and deployment rates) from pre-existing digital infrastructures.

²⁶⁰ Indermit Gill, *Whoever leads in artificial intelligence in 2030 will rule the world until 2100*, BROOKINGS (2020), <https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/> (last visited Jan 22, 2020) (arguing that “There are reasons to believe that the implementation lag for AI-related technologies will be about 10 years.”). Referring to Diego Comin & Martí Mestieri, *Technology Adoption and Growth Dynamics* 38 (2014); Diego Comin & Martí Mestieri, *If Technology Has Arrived Everywhere, Why Has Income Diverged?*, 10 AM. ECON. J. MACROECON. 137–178 (2018).

²⁶¹ Rita Gunther McGrath, *The Pace of Technology Adoption is Speeding Up*, HARVARD BUSINESS REVIEW, 2013, <https://hbr.org/2013/11/the-pace-of-technology-adoption-is-speeding-up> (last visited Jul 3, 2020); Drew DeSilver, *Chart of the Week: The ever-accelerating rate of technology adoption*, PEW RESEARCH CENTER (2014), <https://www.pewresearch.org/fact-tank/2014/03/14/chart-of-the-week-the-ever-accelerating-rate-of-technology-adoption/> (last visited Jul 3, 2020); Michael DeGusta, *Are Smart Phones Spreading Faster than Any Technology in Human History?*, MIT TECHNOLOGY REVIEW, 2012, <https://www.technologyreview.com/2012/05/09/186160/are-smart-phones-spreading-faster-than-any-technology-in-human-history/> (last visited Jul 3, 2020). As cited in Hagemann, Huddleston, and Thierer, *supra* note 31 at 58–59.

²⁶² Jenny C. Aker & Isaac M. Mbiti, *Mobile Phones and Economic Development in Africa*, 24 J. ECON. PERSPECT. 207–232 (2010).

'legible';²⁶³ as such, both governments and private actors have—for better or worse—spent the last decades collating diverse and comprehensive datasets. These pre-existing databases create troves of structured training data. Conversely, on the rollout level, mobile phones and personal computing enable the rollout of AI services to society.²⁶⁴ Indeed, this can already be seen in the growth of 'AI-as-a-service' products.

As a result of this, even AI applications that do not focus on 'virtual' behaviour (e.g. on social media platforms) can nonetheless be rolled out surprisingly quickly. It has proven relatively easy to 'graft' intrusive AI surveillance capabilities onto extant surveillance infrastructures consisting of "dumb" CCTV systems. For instance, in one 2019 experiment, journalists took footage from three regular cameras in New York, and ran it through 'Rekognition', Amazon's commercially-available facial recognition tool. In one day, the set-up detected 2,750 faces; the total cost of the set-up was \$60.²⁶⁵ Such AI use can scale up as far along as the pre-existing infrastructure, as seen in reports that China has begun to roll out AI capabilities to some of its estimated 200 million cameras nationwide.²⁶⁶ As such, just as the computing revolution was able, in its time, to draw on the already-widely-established infrastructures of electrification to facilitate rapid and ubiquitous dissemination, AI systems may in turn leverage those pre-existing computing infrastructures to facilitate rapid dissemination.

This illustrates how many prospective AI usage domains can benefit from an existing infrastructure overhang which can enable the rapid deployment of new AI capabilities or their layering unto existing technologies. It also illustrates how we need not even need to look far into the future to see many large-scale impacts of AI using even very 'narrow' capabilities: in some cases, we need not even look at the future at all. The use of machine learning for simple facial recognition is hardly a futuristic premise. However, the ability to layer pattern-recognition capabilities onto existing architectures of surveillance means that establishing a comprehensive surveillance society is no longer a qualitative leap, but simply a matter of scaling up investment. Barring strong political and regulatory action to the contrary, similar infrastructural patterns may enable the 'downhill' flow of many other disruptive or challenging AI applications.

2.1.8 Reflections: how does AI matter?

Over these preceding sections, we have reviewed a wide range of arguments in favour or against the global impact and importance of AI technology. These are diverse, and highlight the heterogeneity and complexity of technology assessment in this space (see Table 2.1).

²⁶³ See generally SCOTT, J.C., *SEEING LIKE A STATE - HOW CERTAIN SCHEMES TO IMPROVE THE HUMAN CONDITION HAVE FAILED* (1998).

²⁶⁴ That is not to say that coverage is ubiquitous. For instance, while 93% of the world's population live within physical reach of mobile broadband or internet, only 53.6% actually uses the internet, showing a significant digital divide. ITU, *Measuring Digital Development: Facts and Figures 2019* (2019), <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2019.pdf> (last visited Jul 10, 2020).

²⁶⁵ Sahil Chinoy, *We Built an 'Unbelievable' (but Legal) Facial Recognition Machine*, THE NEW YORK TIMES, April 16, 2019, <https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html> (last visited Apr 23, 2019).

²⁶⁶ Paul Mozur, *Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras*, THE NEW YORK TIMES, October 15, 2018, <https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html> (last visited Feb 18, 2019).

Does AI create change?	Recent Progress and promise	Recent AI successes	<ul style="list-style-type: none"> Considerable breakthroughs in past years in image recognition, games, data generation, natural language processing... AI algorithms reaching performance benchmarks that are either symbolically salient, or ‘boring but functional’
		Growing applications in diverse sectors	<ul style="list-style-type: none"> Diverse applications from finance to healthcare, policing to art, science to government...
		Intense attention and investment	<ul style="list-style-type: none"> Significant private sector investment Dozens of national AI strategies, projects, funding
	Prevailing limits and challenges	Limits in industry and applications	<ul style="list-style-type: none"> Mis-branding & ‘fauxtomation’ Hidden labour Garbage-in-garbage-out (use of insufficient, biased or ‘dirty’ data) AI snake oil Wrong tool problem Underappreciated organisational & cultural barriers to adoption
		Preconditions to deployment	<ul style="list-style-type: none"> Data needs (e.g. labelled) Computing hardware needs Human (AI) expertise needs
		Challenges to AI techniques	<ul style="list-style-type: none"> Brittleness Bias (Lack of) transparency and explainability (& others...)
		Challenges in AI Science	<ul style="list-style-type: none"> ML ‘replication crisis’ because of differences in empirical rigor Lack of standardised metrics; ‘phantom progress’ Growing computational demands of AI projects may reach limits
		C1	AI today is neither magic nor farce, but a ‘high-variance’ portfolio of technologies
What change will AI create going forward?	Change from further AI progress?	Outside view (reference class forecasting)	<ul style="list-style-type: none"> International lawyers prematurely jumped gun on other technologies (weather control; deep seabed mining) in the past Possible general scientific ‘slowdown’ across many fields Other scientific fields historically held up by... <ul style="list-style-type: none"> Hard scientific barriers (impervious to brute-force funding) Unexpected political barriers and shocks
		Inside view	<ul style="list-style-type: none"> Other scientific fields and technological benchmarks historically accelerated by sudden, unanticipated, or discontinuous breakthroughs Underestimation of compounding effects of improving tools Overestimation of phantom barriers & ‘irreducible complexity’
		Expert prediction	<ul style="list-style-type: none"> Recent successes & ML scaling laws may be suggestive of ‘bitter lesson’ (i.e. ‘just more computation’), signalling significant AI performance increases Even scientifically or philosophically ‘uninteresting’ capability improvements (<i>‘no true intelligence’</i>) can prove highly societally disruptive Recent successes (e.g. GPT-3) may prove intrinsically limited; computationally limited, or unable to scale up to ‘true intelligence’ Surveys show AI researchers expect ‘high-level machine intelligence’ by 2060s, with many disruptive breakthroughs along the way High variance in expert predictions of AI timelines Uncertainty whether today’s AI experts can provide meaningful estimates
	Change from algorithmic proliferation?	C2	Uncertainty illustrates importance of sober further study and anticipatory governance
		Falling input thresholds	<ul style="list-style-type: none"> Training data requirements increasingly achievable (growing supply; increasing data efficiency; ...) Compute barriers may fall (increasing algorithmic efficiency; ...). New ‘autoML’ tools help reduce human expertise barriers
		Broad usability	<ul style="list-style-type: none"> AI capabilities serve domain-agnostic functions in diverse contexts Potential status as ‘General-Purpose Technology’
		Patterns of global diffusion	<ul style="list-style-type: none"> Many AI applications have low fixed cost & private sector dominance, suggestive of rapid global diffusion
		Roll-out rates and infrastructure overhang	<ul style="list-style-type: none"> GPT ‘time to pervasiveness’ has been falling historically (80 yrs for steam; 40 yrs electricity; 20 yrs IT); some technologies leapfrogged legacy systems. AI can be layered onto pre-existing infrastructures (CCTV; mobile phones)
	C3	Independent of future progress, ‘disruptive’ AI capabilities already exist & will drive global change	

Table 2.1. Overview: how does AI matter? Anticipatory and sceptical arguments²⁶⁷

²⁶⁷ Note, blue cells represent arguments or pieces of evidence that support (relative) scepticism of AI’s impact or importance; red cells represent arguments or pieces of evidence that support (relative) anticipation of AI having considerable impact.

In sum, three broad conclusions can be drawn on the basis of these arguments.

Firstly (C1), AI today is neither magic nor farce, but may instead be a *high-variance technology portfolio*. As an incipient emerging technology, it still faces many barriers and growing pains; nonetheless, it is seeing enough viable applications that this should be sufficient to warrant attention. This much may not be very controversial, although it is valuable to investigate debates and considerations.

Secondly (C2), in spite of prevailing public misconceptions, *there remains uncertainty around the potential imminence, speed, or ceiling of future AI progress*, with both outside (comparative reference class) and inside (AI-focused) arguments providing grounds to expect both further breakthroughs as well as scepticism. Pragmatically, on balance, continued progress for at least the coming years seems plausible. Critically, this extant uncertainty illustrates the importance of undertaking further study, and formulating anticipatory governance strategies.

Thirdly (C3), even if one adopted an unnecessarily pessimistic expectation around further fundamental AI progress, *AI already matters today*: ‘disruptive’ AI capabilities and algorithmic applications already exist, will likely diffuse relatively quickly over the coming years, and will be more than sufficient to drive global change and disruption in many if not all domains.²⁶⁸ For all intents and purposes, *disruptive AI* systems are already out there—and may soon be ‘here’. The technology is already heralding a new chapter in law and governance, and this makes it a fruitful and indeed critical topic for scholars to engage with—today. As such, we return to the question motivating this subsection: does AI (actually) matter? The answer may lie in understanding AI technology as a *high-variance* technology, and as a ‘*general-purpose technology*’.

Of course, AI is not the first technology to raise foundational societal questions. For instance, Daniel Deudney has suggested that AI is but the latest amongst several innovations which are reshaping the stakes and rapidity of global societal and strategic changes, noting that “the emergence and global spread of a modern civilization devoted to expanding scientific knowledge and developing new technologies has radically increased the rate, magnitude, complexity, novelty, and disruptiveness of change.”²⁶⁹ At the least, the challenge of governing AI technology well—under conditions of change—may be a major theme facing societies over the coming decades. They may prove a key step (or test) in the larger historical challenge of evolving governance systems—and, if necessary, developing new ones—to be up to the task of responsibly managing diverse and potent technologies in the coming decades.

In this sense, our ability—or inability—to collectively manage our relationship to emerging technologies such as AI may prove a key objective and test of our governance systems and norms in the face of the broader ‘technology tsunami’,²⁷⁰ and all of its changes.

²⁶⁸ Parson et al., *supra* note 132; DAFOE, *supra* note 130.

²⁶⁹ Daniel Deudney, *Turbo Change: Accelerating Technological Disruption, Planetary Geopolitics, and Architectonic Metaphors*, 20 INT. STUD. REV. 223–231, 223 (2018). Notably, he does not consider this a particularly new state historically, suggesting that we have been steadily entering this era since the emergence of nuclear weapons, or even of industrial (fossil-fuel) based civilization.

²⁷⁰ Richard Danzig, *An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order*, 5 LAWFARE RES. PAP. SER. (2017), <https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf> (last visited Sep 1, 2017).

2.2 Challenges: AI needs global governance

In the face of such challenges, can we ensure that we do not find ourselves “addressing twenty-first century challenges with twentieth-century laws”?²⁷¹ In the first place, what policy challenges does AI create? In the second place, (why) would these challenges require *global* governance? Thirdly, (why) would we expect governance systems to struggle in the face of AI?

As such, we will briefly (2.2.1) survey the array of possible challenges AI has been implicated in; (2.2.2) explore which of these challenges might benefit from-, or require global cooperation; and (2.2.3.) discuss whether global AI governance is even viable, taking into consideration various barriers and challenges as well as strategies and policy levers.

2.2.1 Scoping AI challenges

To be sure, while many have hailed the benefits that AI technology could bring, others have begun to express concern about the technology’s diverse challenges. As a widely applicable technology aimed at automating human tasks, AI can, almost by definition, touch upon nearly any domain of human activity. Accordingly, the use of AI systems is driving considerable social changes—for better and for worse—across a range of dimensions.²⁷²

In the first place, AI systems have raised diverse ethical concerns, as highlighted by a growing series of controversies and scandals.²⁷³ AI applications have been critiqued for undermining the autonomy of consumers, by fostering technological ‘addiction’,²⁷⁴ and hijacking attention.²⁷⁵ The industry itself has been criticised for relying on work that is exploitative, insecure, low-paid, or degrading.²⁷⁶ AI applications have been implicated in threatening various core values or (human) rights, such as privacy, non-discrimination, or human dignity.²⁷⁷

²⁷¹ Eric Talbot Jensen, *The Future of the Law of Armed Conflict: Ostriches, Butterflies, and Nanobots*, 35 MICH. J. INT. LAW 253–317, 253 (2014).

²⁷² These dimensions anticipate the taxonomy of ‘problem-logics’, discussed in more detail in chapter 5.4.

²⁷³ See KATE CRAWFORD ET AL., *AI Now 2019 Report* 100 (2019), https://ainowinstitute.org/AI_Now_2019_Report.pdf. And an informal compilation, see: Roman Lutz, *Learning from the past to create Responsible AI: A collection of controversial, and often unethical AI use cases*, ROMAN LUTZ, <https://romanlutz.github.io/ResponsibleAI/> (last visited Jun 22, 2020).

²⁷⁴ ADAM ALTER, *IRRESISTIBLE: THE RISE OF ADDICTIVE TECHNOLOGY AND THE BUSINESS OF KEEPING US HOOKED* (Reprint edition ed. 2018).

²⁷⁵ JAMES WILLIAMS, *STAND OUT OF OUR LIGHT: FREEDOM AND RESISTANCE IN THE ATTENTION ECONOMY* (Reprint edition ed. 2018).

²⁷⁶ Lina M Khan, *Amazon’s Antitrust Paradox*, 126 YALE LAW J. 710–805 (2017).

²⁷⁷ These literatures are all extensive. On human rights, see Q. C. VAN EST, J. GERRITSEN & L. KOOL, *Human rights in the robot age: challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality* (2017), <https://research.tue.nl/en/publications/human-rights-in-the-robot-age-challenges-arising-from-the-use-of-> (last visited May 22, 2019). On non-discrimination, see Amnesty International & Access Now, *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems* (2018), https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf (last visited Aug 27, 2018); Hin-Yan Liu, *Three Types of Structural Discrimination Introduced by Autonomous Vehicles*, 51 UC DAVIS LAW REV. 32 (2018). On privacy, see Ryan Calo, *Peeping HALS: Making Sense of Artificial Intelligence and Privacy*, 2 EJLS - EUR. J. LEG. STUD. (2010), <http://www.ejls.eu/6/83UK.htm> (last visited May 13, 2017); JAY STANLEY, *The Dawn of Robot Surveillance: AI, Video Analytics, and Privacy* (2019), https://www.aclu.org/sites/default/files/field_document/061119-robot_surveillance.pdf (last visited Jun 15, 2019). On human dignity, see Roger Brownsword, *From Erewhon to AlphaGo: for the sake of human dignity, should we destroy the machines?*, 9 LAW INNOV. TECHNOL. 117–153 (2017).

Particular concerns have focused on the biased performance of AI systems,²⁷⁸ with far-reaching contexts especially in high-stakes decisions ranging from facial recognition²⁷⁹ to hiring,²⁸⁰ and from criminal risk assessments²⁸¹ to migration management.²⁸²

At a societal and political level, such ethical concerns converge around the ability of AI to enable or support pervasive systems of surveillance, disinformation or manipulation, which could come to threaten democratic norms and practices.²⁸³ Longer-term, there are concerns over the disruption of economic and social stability,²⁸⁴ including through large-scale work displacement and changes in the nature of work itself.²⁸⁵ Indeed, along with other digital technologies, AI technology is one of the constitutive infrastructures of what Jamie Susskind has called the emerging ‘digital lifeworld’, re-opening perennial political-theoretical questions around power, freedom, democracy, and justice.²⁸⁶ At the international level, the growing military use of AI—particularly but not limited to the rise of ‘lethal autonomous weapons systems’—has led to concerns over human dignity, -control, and dynamics of military escalation.

AI also gives rise to new security threats, as various actors are able to exploit AI either as a tool or as a vulnerable attack surface, to scale up old attacks, or to carry out entirely new types of crimes.²⁸⁷ In addition, AI raises novel safety concerns in terms of unanticipated behaviour of

²⁷⁸ Baracas and Selbst, *supra* note 109.

²⁷⁹ Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 in PROCEEDINGS OF MACHINE LEARNING RESEARCH 1–15 (2018), <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.

²⁸⁰ Ifeoma Ajunwa, *Automated Employment Discrimination*, SSRN ELECTRON. J. (2019), <https://www.ssrn.com/abstract=3437631> (last visited May 12, 2020). One notable example of this was provided in 2017, when Amazon shut down an AI recruitment tool for demonstrating pervasive gender bias in hiring recommendations. Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, REUTERS, October 10, 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> (last visited Aug 11, 2020).

²⁸¹ Berk et al., *supra* note 109.

²⁸² Petra Molnar, *Technology on the margins: AI and global migration management from a human rights perspective*, 8 CAMB. INT. LAW J. 305–330 (2019); Ana Beduschi, *International migration management in the age of artificial intelligence*, MIGR. STUD. (2020), <https://academic.oup.com/migration/advance-article/doi/10.1093/migration/mnaa003/5732839> (last visited Feb 14, 2020).

²⁸³ Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCIENTIFIC AMERICAN, 2017, <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/> (last visited May 29, 2017); Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. LAW REV. 1753–1820 (2019); Steven Feldstein, *The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression*, 30 J. DEMOCR. 40–52 (2019); Gillian Bolsover & Philip Howard, *Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda*, 5 BIG DATA 273–276 (2017).

²⁸⁴ Nick Bostrom, Allan Dafoe & Carrick Flynn, *Public Policy and Superintelligent AI: A Vector Field Approach*, in ETHICS OF ARTIFICIAL INTELLIGENCE (S.M. Liao ed., 2019), <http://www.nickbostrom.com/papers/aipolicy.pdf> (last visited May 13, 2017).

²⁸⁵ GRAY AND SURI, *supra* note 70; Carl Benedikt Frey & Michael A. Osborne, *The future of employment: How susceptible are jobs to computerisation?*, 114 TECHNOL. FORECAST. SOC. CHANGE 254–280 (2017). Though for a more optimistic take, see DANAHER, *supra* note 63.

²⁸⁶ SUSSKIND, *supra* note 220.

²⁸⁷ Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV180207228 Cs (2018), <http://arxiv.org/abs/1802.07228> (last visited Feb 21, 2018); Thomas C. King et al., *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, SCI. ENG. ETHICS (2019), <https://doi.org/10.1007/s11948-018-00081-0> (last visited Feb 21, 2019); Keith J Hayward & Matthijs M

AI systems; behavioural risks from unexpected inputs, unconstrained optimisation or algorithmic ‘creativity’; and challenges arising from interactions between AIs and in human-machine assemblages.²⁸⁸ Such systems could yield cascading and even catastrophic safety failures when deployed in critical infrastructure (such as major failures in traffic safety, energy grids, or network security).²⁸⁹ Less visibly but no less important, AI systems may shift the structure or texture of human interactions, altering actors’ incentives, strategic pressures or even goals in destabilizing or harmful directions.²⁹⁰ This could alter the texture of human relations at the domestic level, or destabilize global politics at the international one.²⁹¹

At the same time, along with these diverse challenges, it should also be kept in mind that there are considerable global benefits which AI could secure, which might not be achieved by default, or by relying only on market mechanisms, but which will require cooperation or coordination. Indeed, *prima facie*, the ability to support, automate, or improve upon the speed, accuracy, or scale of human decision-making could help provide tremendous benefits. That is, while there may be societal challenges that are not amenable to ‘intelligence’ alone, there are few problems where more ‘intelligence’ could not help, in some ways, in finding better solutions, or in implementing existing solutions in better ways. If governed well, AI technologies could drive considerable economic growth,²⁹² lift global standards of living,²⁹³ and help address many critical

Maas, *Artificial intelligence and crime: A primer for criminologists*, CRIME MEDIA CULT. 1741659020917434 (2020); M. Caldwell et al., *AI-enabled future crime*, 9 CRIME SCI. 14 (2020).

²⁸⁸ Amodei et al., *supra* note 107; Ram Shankar Siva Kumar et al., *Failure Modes in Machine Learning Systems*, ARXIV191111034 CS STAT (2019), <http://arxiv.org/abs/1911.11034> (last visited Jan 7, 2020); Joel Lehman et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, ARXIV180303453 CS (2018), <http://arxiv.org/abs/1803.03453> (last visited Apr 2, 2018); On interaction between systems, see Iyad Rahwan et al., *Machine behaviour*, 568 NATURE 477 (2019); Victoria Krakovna et al., *Specification gaming: the flip side of AI ingenuity*, DEEPMIND (2020), <https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity> (last visited May 12, 2020); Matthijs M. Maas, *Regulating for “Normal AI Accidents”: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment*, in PROCEEDINGS OF THE 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 223–228 (2018), <https://doi.org/10.1145/3278721.3278766> (last visited Sep 8, 2020).

²⁸⁹ OSONDE OSOBA & WILLIAM WELSER, *THE RISKS OF ARTIFICIAL INTELLIGENCE TO SECURITY AND THE FUTURE OF WORK* (2017), <https://www.rand.org/pubs/perspectives/PE237.html> (last visited Jun 22, 2020).

²⁹⁰ Remco Zwetsloot & Allan Dafoe, *Thinking About Risks From AI: Accidents, Misuse and Structure*, LAWFARE (2019), <https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure> (last visited Feb 12, 2019); Agnes Schim van der Loeff et al., *AI Ethics for Systemic Issues: A Structural Approach* (2019), <http://arxiv.org/abs/1911.03216> (last visited Jan 13, 2020).

²⁹¹ DAFOE, *supra* note 130; Paul Scharre, *Autonomous Weapons and Stability*, March, 2020; See also Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NATL. SECUR. LAW POLICY 1–25 (2019).

²⁹² For instance, it has been projected that machine learning applications could add \$3.9 trillion in added value to the global economy by 2022. Christy Pettey & Rob van der Meulen, *Gartner Says Global Artificial Intelligence Business Value to Reach \$1.2 Trillion in 2018*, GARTNER (2018), <https://www.gartner.com/en/newsroom/press-releases/2018-04-25-gartner-says-global-artificial-intelligence-business-value-to-reach-1-point-2-trillion-in-2018> (last visited Aug 11, 2020). In the longer term, it is expected this could grow to \$16 trillion by 2030; PwC, *The macroeconomic impact of artificial intelligence* 78 (2018), <https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf>; SZCZEPANSKI MARCIN, *Economic impacts of artificial intelligence (AI)* 8 (2019), [https://europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRI_BRI\(2019\)637967_EN.pdf](https://europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRI_BRI(2019)637967_EN.pdf). Upper estimates project even greater extremes, see WILLIAM D NORDHAUS, *Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth* 49 (2015), <https://www.nber.org/papers/w21547.pdf>.

²⁹³ STONE ET AL., *supra* note 10.

global challenges, such as achieving the Sustainable Development Goals, improving global health, or tackling climate change.²⁹⁴ Indeed, in some readings, AI could even support rights,²⁹⁵ or support far-reaching societal progress and empowerment.²⁹⁶ AI technologies promise considerable opportunities, but the realisation of such goods may require some forms of global coordination or cooperation rather than a laissez-faire approach.

Underlying all these challenges is the fact that AI can drive changes in systems of law and regulation themselves. The use of algorithms in legal decision-making or prediction can expand and speed up legal process. They might also challenge existing legal doctrines and concepts, or more fundamentally, alter the practices, logics or assumptions of governance systems, including even our concept of ‘law’ itself.²⁹⁷ Such shifts may be appealing or beneficial but should be critically examined. This is especially because, given their impact on the broader systems of (international) law, these changes in law and regulation may also alter the way we could or should approach the governance of (AI) technology itself.

2.2.2 Do these challenges need global cooperation?

It is clear that policymakers have their work cut out. Yet the question is, (why) do we need *global* governance for AI? To be certain, some of the above challenges raised by AI might be adequately resolved at a domestic level, through national legislation.²⁹⁸ Certain questions around social or industrial robots, the use of algorithms in domestic judicial systems, or the regulation of self-driving cars, might appear to constitute AI issues that could be *adequately* addressed at the domestic level.²⁹⁹

At the same time, many issues from AI could at least *benefit* from international cooperation. This is because, as Jacob Turner notes, cooperation can seize the benefits of

²⁹⁴ Ricardo Vinuesa et al., *The role of artificial intelligence in achieving the Sustainable Development Goals*, 11 NAT. COMMUN. (2020), <https://www.nature.com/articles/s41467-019-14108-y> (last visited May 8, 2019); David Rolnick et al., *Tackling Climate Change with Machine Learning*, ARXIV190605433 CS STAT (2019), <http://arxiv.org/abs/1906.05433> (last visited Jul 23, 2019).

²⁹⁵ See also Andrea Scripa Els, *Artificial Intelligence as a Digital Privacy Protector*, 31 HARV. J. LAW TECHNOL. 19 (2017); Urs Gasser, *Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy*, 130 HARV. LAW REV. FORUM 10 (2016).

²⁹⁶ ITU, *Artificial Intelligence for Global Good* (2018), https://www.itu.int/en/itunews/Documents/2018/2018-01/2018_ITUNews01-en.pdf (last visited Feb 13, 2019); Luciano Floridi et al., *How to Design AI for Social Good: Seven Essential Factors*, SCI. ENG. ETHICS (2020), <https://doi.org/10.1007/s11948-020-00213-5> (last visited Apr 5, 2020). For mid-term scenarios, see also Edward Parson et al., *Could AI drive transformative social progress? What would this require?*, AI PULSE (2019), <https://aipulse.org/could-ai-drive-transformative-social-progress-what-would-this-require/> (last visited Sep 28, 2019); DANAHER, *supra* note 63.

²⁹⁷ Benjamin Alarie, *The path of the law: Towards legal singularity*, 66 UNIV. TOR. LAW J. 443–455 (2016); Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 INDIANA LAW J. 1401–1447 (2017); Brian Sheppard, *Warming up to inscrutability: How technology could challenge our concept of law*, 68 UNIV. TOR. LAW J. 36–62 (2018).

²⁹⁸ Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 UC DAVIS LAW REV. 37 (2017); Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, HARV. J. LAW TECHNOL. (2016), <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf> (last visited Mar 5, 2018).

²⁹⁹ Even this might not be the case, however, as self-driving cars throw up novel issues for existing international treaties, such as the 1949 Geneva Convention on Road Traffic and the 1968 Vienna Convention on Road Traffic. Bryant Walker Smith, *New Technologies and Old Treaties*, 114 AJIL UNBOUND 152–157 (2020). This will be discussed in more detail in Chapter 5.3 (on governance ‘development’).

standardisation while avoiding the costs of trans-jurisdiction regulatory uncertainty as well as the risk of regulatory arbitrage by AI producers picking their preferred rules.³⁰⁰ In the absence of coordinated rules, he cautions, regulation for AI could become “Balkanised, with each territory setting its own mutually incompatible rules.”³⁰¹ This is exacerbated by the unequal global division of AI talent and resources, and growing practices of ‘AI protectionism’.³⁰² Indeed, there are a growing number of scholars who argue that certain challenges created by AI cannot be met through either national-level regulation, nor through industry self-regulation,³⁰³ but will require some degree of collective action at a global level.³⁰⁴

This is of course hardly unprecedented. From a historical angle, diverse other technological developments have led to global governance efforts, meeting with varying but generally good success. This includes regimes for high-stakes strategic technologies, such as the nuclear non-proliferation regime, outer space law, or internet governance.³⁰⁵ That is not to say all general-purpose technologies have received or been subject to such regulations. Nonetheless, many scholars hold that historically, new technological developments have been key drivers, directly or indirectly, of many landmark developments in international law.³⁰⁶

As such, rather than treating the matter of global cooperation for AI as an either-or question, a more productive questions would be: on what ground could global cooperation be founded? Given the technology’s breadth and diversity, what AI problems would need global governance?

2.2.2.1 Securing AI global public goods

Over the past decades, social scientists have developed diverse typologies of global public good problems, and explored their distinctive governance requirements. In one influential taxonomy, Scott Barrett distinguished between problem domains where success is dependent on, respectively, the *single best effort* (e.g. asteroid defence, peacekeeping, geo-engineering); the *weakest link* (e.g. disease eradication); *aggregate effort* (e.g. climate change mitigation); *mutual*

³⁰⁰ TURNER, *supra* note 17 at 237–239.

³⁰¹ *Id.* at 239–240.

³⁰² Feijoo et al., *supra* note 41. See also the discussion of ‘AI nationalism’ (and new dynamics of core-periphery dependency relationships amongst states) in Ian Hogarth, *AI Nationalism*, IAN HOGARTH (2018), <https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism> (last visited Jul 23, 2018).

³⁰³ Cf. Paul Nemitz, *Constitutional democracy and technology in the age of artificial intelligence*, 376 PHIL TRANS R SOC A 20180089 (2018). See also Tom Slee, *The Incompatible Incentives of Private Sector AI*, in THE OXFORD HANDBOOK OF AI ETHICS 34 (M Dubber & F. Pasquale eds., 2019). Eyal Benvenisti, *Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?*, 29 EUR. J. INT. LAW 9–82 (2018).

³⁰⁴ Amanda Askell, Miles Brundage & Gillian Hadfield, *The Role of Cooperation in Responsible AI Development* 23 (2019). See also TURNER, *supra* note 17 at 237–239 (discussing the cross-boundary nature of AI challenges, the cost of regulatory uncertainty, and the need to avoid arbitrage).

³⁰⁵ Drezner, *supra* note 247 (on the nuclear regime); TURNER, *supra* note 17 at 244–247 (on the 1967 UN Outer Space Treaty). See also Verity Harding, *Lessons from history: what can past technological breakthroughs teach the AI community today* (2020), <https://www.bennettinstitute.cam.ac.uk/blog/lessons-history-what-can-past-technological-breakthrough/> (last visited Aug 17, 2020) (referencing the Outer Space Treaty, the UK’s Warnock Committee and Human Embryology Act, the Internet Corporation for Assigned Names and Numbers [ICANN], and the European ban on genetically modified crops).

³⁰⁶ Picker, *supra* note 33; Rayfuse, *supra* note 219. And specifically on the technological impact on the laws of war, see Braden Allenby, *Are new technologies undermining the laws of war?*, 70 BULL. AT. SCI. 21–31 (2014).

restraint (e.g. non-proliferation), or *coordination* (e.g. industry standards).³⁰⁷ Each of these problem types has distinct dynamics, in terms of why international cooperation is needed, whether cost sharing is required, whether enforcement of an agreement is challenging, or what distinct types of international institutions or arrangements are involved in the provision of each good.³⁰⁸

As a result of these differences, not all collective action problems are created equal. Rather, they pose different problems for global governance. Focusing on three distinct categories of public goods, Anne van Aaken has argued that these pose differential challenges for an international law system that is traditionally based on state consent.³⁰⁹ Governance challenges may not be particularly severe for ‘single best effort’ goods, as the problem is solved so long as one country has a strong enough incentive to invest.³¹⁰ Matters are more challenging in the context of ‘weakest link’ goods, where the good is provided if and only if every country contributes sufficiently, but is not provided if at least one state does not do so, because it is unwilling (hold-out) or incapable (e.g. failing or failed states). Van Aaken argues that several mechanisms in international law are known to overcome this problem, such as international organisations (which can provide for e.g. vaccination campaigns).³¹¹ She argues that the most difficult governance challenge for international law may be found in securing ‘aggregate efforts’ public good, as these depend on the sum effort of states, and therefore are subject to free-riding and collective action concerns.³¹² To what extent, if at all, do the issues created or enabled by AI map onto various

³⁰⁷ SCOTT BARRETT, WHY COOPERATE?: THE INCENTIVE TO SUPPLY GLOBAL PUBLIC GOODS 20 (2007), <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199211890.001.0001/acprof-9780199211890> (last visited May 24, 2018).

³⁰⁸ *Id.* at 2–21.

³⁰⁹ Anne van Aaken, *Is International Law Conducive To Preventing Looming Disasters?*, 7 GLOB. POLICY 81–96, 83 (2016).

³¹⁰ Of course, this also implies that *single best effort* goods can be very severe when no single actor *perceives* a strong enough incentive to invest, in advance of a certain disaster. This can occur under epistemic conditions where there is pervasive uncertainty around certain potential risks. For a discussion of this ‘tragedy of the uncommons’, see Jonathan B. Wiener, *The Tragedy of the Uncommons: On the Politics of Apocalypse*, 7 GLOB. POLICY 67–80 (2016). And more broadly, see Milan M. Cirkovic, Nick Bostrom & Anders Sandberg, *Anthropic Shadow: Observation Selection Effects and Human Extinction Risks*, 30 RISK ANAL. (2010), <http://www.nickbostrom.com/papers/anthropicshadow.pdf>; Eliezer Yudkowsky, *Cognitive Biases Potentially Affecting Judgment of Global Risks*, in GLOBAL CATASTROPHIC RISKS (Nick Bostrom & Milan Cirkovic eds., 2011), <https://intelligence.org/files/CognitiveBiases.pdf> (last visited Mar 5, 2018). Finally, even in cases where a certain public good is achieved through a single best effort, there may still be operational and logistical concerns around adequately (or fairly) distributing the good globally, as in the role which institutions such as the World Health Organization (WHO) might (and, one hopes, will) play in allocating the public good of a vaccine to different countries.

³¹¹ Aaken, *supra* note 309 at 83. However, Sandler is more pessimistic, arguing that because weakest-link public goods are constrained by the smallest effort made, these goods can be particularly hard to secure in situations where countries possess highly different financial capabilities, and where there are furthermore many countries, as this results in situations where parties realize someone ought to shore up the financial means of the poorer members but, being uncertain of one another’s intentions, wait for others to act first. Todd Sandler, *Strategic Aspects of Difficult Global Challenges*, 7 GLOB. POLICY 33–44, 42 (2016).

³¹² Aaken, *supra* note 309 at 83. Interestingly, in the context of cooperation to mitigate potential extreme risks emerging from future AI systems, Eliezer Yudkowsky has echoed some of these distinctions, by differentiating between “Strategies which require unanimous cooperation [but which] can be catastrophically defeated by individual defectors or small groups. [...] Strategies which require majority action [requiring] most, but not all, people in a large pre-existing group to behave a particular way, [and] Strategies which require local action—a

global public good problems?³¹³ Certainly, as noted, some AI challenges may not at all create collective action problems (at least at an international level). However, some do.

For instance, (1) questions around constraining the open proliferation or use of AI capabilities to various malicious or criminal actors³¹⁴ could well be considered a *weakest link* problem, in the sense that if a single country provides a ‘safe haven’, these tools may proliferate rapidly.³¹⁵

Likewise, (2) questions over the non-development or non-deployment of adversarial AI systems—either military uses of Lethal Autonomous Weapons Systems, or more generally the use of AI systems in cyberwarfare, computational propaganda or information warfare—could be seen as constituting a *mutual restraint* public good, rendered more challenging because of the difficulties associating with monitoring and verifying compliance.³¹⁶

Moreover, (3) questions pertaining to setting general standards in AI system reliability, explainability, training data privacy and security, could be seen as *coordination* goods, and may well come to constitute the most common global governance debates around AI in recent years.

Curiously, the relative ‘scalability’ of pre-trained AI systems has distinct implications for the final two categories of global public goods. On the one hand, (4) such systems would lend themselves relatively naturally to *single best effort* problems, where it concerns the use of AI systems to secure certain global benefits or achieving certain goals (e.g. climate change modelling; systems to monitor or detect war crimes). Such systems would need to be trained up, but once available, could be shared relatively easily and at low marginal cost to many parties.

Conversely, (5), the scalability of AI technologies also suggests that perhaps surprisingly few AI challenges would be understood as *aggregate effort* public goods. As such, for AI, the key global public goods may centre on issue areas that involve coordination, mutual restraint, or weakest link problems.

In sum, while some AI issues may be governed adequately at the national level, there are many that require global cooperation of some form for distinct reasons. It is therefore to be expected that there will continue to be a strong pull towards achieving at least some form of global cooperation, for some issues. That does not mean, however, that this is a straightforward task.

2.2.3 Is AI global governance possible? Barriers and Levers

Even if it is needed, is AI regulation at the global level even at all possible? That is, could it be enforced? There are certainly challenges, but also potential levers and opportunities.

concentration of will, talent, and funding which overcomes the threshold of some specific task.” Yudkowsky, *supra* note 4 at 28.

³¹³ For previous discussions of commons problems around AI, see also AI Impacts, *Friendly AI as a global public good*, AI IMPACTS (2016), <https://aiimpacts.org/friendly-ai-as-a-global-public-good/> (last visited Aug 30, 2020); Askell, Brundage, and Hadfield, *supra* note 304.

³¹⁴ Brundage et al., *supra* note 287; Hayward and Maas, *supra* note 287.

³¹⁵ However, on the other hand, this case is also dissimilar to traditional weakest link goods, because top AI development capabilities are still somewhat unequally distributed around the world, and would be much more sensitively affected by lax standards in these states, than by the same standards in lagging countries. However, this situation is likely to change over time.

³¹⁶ See also the discussion in sections 5.4.2 (discussing how AI could support compliance monitoring) and 7.1.1.3 (discussing the monitoring and compliance challenges around AI arms control).

2.2.3.1 Barriers and challenges

Amongst legal scholars writing on AI, there is disagreement between those arguing that existing regulation or existing legal principles are more than adequate to deal with AI, and those who suggest that these may not be sufficient.

Within this second group, one can draw an additional distinction. On the one hand, it can be argued that AI technology has certain features that give rise to the need for new regulation or revision of legal concepts (it is a prolific *rationale* for regulation).³¹⁷ On the other hand, some hold that AI technology has a range of features that renders it an unusually challenging object for governance and regulation (it is a challenging *target* for governance).³¹⁸ Putting aside the first question, we can briefly focus on the second question: what would be the barriers to effective governance or regulation?

In the first place, as Matthew Scherer and others have noted, AI regulation is complicated by the difficulty of fixing a single definition of what AI actually is.³¹⁹ In addition to this, AI research and development processes are distinguished by certain features which inhibit effective regulation even at the national level.³²⁰ As Scherer notes, AI development is *discreet*, *discrete*, *diffuse*, and *opaque*.

That is, it is *discreet* because, Scherer claims, relatively little concentrated physical infrastructure is required, and AI projects may be developed without the mass institutional frameworks that were necessary for building industrial capacity in the last century.³²¹ It is *discrete* because separate components may be designed in a decentralised manner, without top-down coordination, with the full potential or risks of a system not becoming apparent until they are brought together in a new application. In particular, as has been noted in the context of software technology general, this can result in a ‘many hands’ problem that creates challenges around assigning responsibility (or liability) amongst many actors.³²² For similar reasons, AI development can be relatively *diffuse*, as software development can be geographically and organisationally dispersed, and involve diverse actors in diverse jurisdictions, enabled by open source software and widely available tools.³²³ Finally, AI development is *opaque*, as the technologies are not always well understood by regulators, and outsiders or inspectors cannot

³¹⁷ Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. LAW REV. 513–564 (2015); TURNER, *supra* note 17 at 2.

³¹⁸ This distinction between governance ‘rationales’ and governance ‘targets’ (and the role of material or other features of the technology in determining either) will be a key aspect of Chapter 5.

³¹⁹ Scherer, *supra* note 298 at 361–362. See also Parson et al., *supra* note 132; Schuett, *supra* note 15.

³²⁰ Scherer, *supra* note 298 at 369.

³²¹ Though as noted, this may be changing, as cutting-edge AI research is requiring increasingly large amounts of computational power or hardware. See section 2.1.3.3.

³²² See generally Merel Noorman, *Computing and Moral Responsibility*, in THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., Spring 2020 ed. 2018), <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/> (last visited Sep 5, 2020).

³²³ Scherer, *supra* note 298 at 370–371.

reliably detect harmful features in an AI system under development.³²⁴ All this, Scherer argues, renders *ex ante* regulation for AI challenging even in principle.³²⁵

As a result of such factors, AI technology may prove particularly vulnerable to the so-called ‘Collingridge Dilemma’,³²⁶ a well-known dictum in technology governance by David Collingridge which states that “[w]hen change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming.”³²⁷ This reflects a general regulatory problem that is said to arise in the context of new technologies. Early on a technology’s development cycle, we face an *information problem*, because the technology’s critical features, salient uses and unanticipated impacts cannot be predicted (or agreed upon) until it has had time to be more extensively developed and used. Yet at that later stage, we face a *power problem*: implementing controls or governance solutions is more difficult now, because the technology has already become embedded in society. As a result, investments have been made in the technology; infrastructure rolled out; established interests have become clear and entrenched; or early governance approaches have locked in specific ‘legal metaphors’—and therefore solution portfolios—over others.

This poses a challenge for global AI governance, because institutions can be path-dependent, and how we structure international cooperation today can prove critical in the long run, especially as the technology advances and strategic stakes scale up.³²⁸ Indeed, Morin and others have argued that AI is one of a set of emerging complex problems which “concern the unknown; they are often unprecedented, span multiple issue areas in their scope or in their consequences, and can be disruptive.”³²⁹ Because of this, existing international institutions may not be well configured to deal with issues such as AI.

2.2.3.2 Strategies and levers

The claim has sometimes been made that digital technologies are uniquely challenging to regulation by states, whether individually or collectively.³³⁰ Such arguments are not new, but indeed echo claims made during the early days of the internet, suggesting that the decentralised nature of cyberspace left it fundamentally beyond the scope of national jurisdictions.³³¹ Yet in the

³²⁴ *Id.* at 357. This is exacerbated by the so-called ‘Volkswagen problem’, which refers to a 2015 scandal where the car company was found to have embedded ‘defeat device’ algorithms in their cars’ engine computers, which detected when the car was being tested for emissions, and accordingly altered the engine’s operation to comply with environmental standards. Russell Hotten, *Volkswagen: The scandal explained*, BBC NEWS, December 10, 2015, <https://www.bbc.com/news/business-34324772> (last visited Sep 4, 2020).

³²⁵ Although he remains optimistic about avenues through which different legislative and judicial authorities could adapt to respond better. Scherer, *supra* note 298 at 376–392. See also Parson et al., *supra* note 132.

³²⁶ DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY (1981). For a brief discussion of the Collingridge Dilemma in the context of AI, see also Michael Guihot, Anne F. Matthew & Nicolas Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, VANDERBILT J. ENTERTAIN. TECHNOL. LAW, 421–422 (2017), <https://papers.ssrn.com/abstract=3017004> (last visited Jul 2, 2018).

³²⁷ COLLINGRIDGE, *supra* note 326 at 11. He calls this the ‘dilemma of control’.

³²⁸ See also Jean-Frédéric Morin et al., *How Informality Can Address Emerging Issues: Making the Most of the G7*, 10 GLOB. POLICY 267–273, 2 (2019). See also Cihon, Maas, and Kemp, *supra* note 150 at 232.

³²⁹ Morin et al., *supra* note 328 at 4.

³³⁰ KLAUS SCHWAB, THE FOURTH INDUSTRIAL REVOLUTION (2017).

³³¹ John Perry Barlow, *A Declaration of the Independence of Cyberspace*, ELECTRONIC FRONTIER FOUNDATION (1996), <https://www.eff.org/cyberspace-independence> (last visited Aug 20, 2020). In 1993, John Perry Barlow

decades since, such claims have not held up. The internet has proven not just a site for national governmental regulation and restrictions, but even a tool of regulating behaviour.³³² It has also been subject to extensive global multi-stakeholder governance.³³³ With notable issue exceptions (such as over cybersecurity), this model of internet governance has been broadly successful.³³⁴

That is not to suggest technology governance in these areas has been remotely easy or without gaps. Nonetheless, it shows how the prospects of enforcing global regulations on digital technologies may be more nuanced. A key point may be that, as Beaumier and others have argued, the modern global digital economy is constituted of very distinct ‘regulatory objects’, some of which pose considerable challenges, but others of which can certainly be subject to international regulation.³³⁵ They argue that the ease with which a certain digital technology can be subjected to global regulation depends on two factors: the artefact’s apparent *materiality*, along with the degree of (market) *centralisation*—often a function of either high capital requirements or network effects, or both. This leads to four types of digital technologies: (Type-1) those that are clearly material and centralised (e.g. submarine cables; satellites); (Type-2) those that are clearly material but decentralised (e.g. ‘smart speakers’ and other internet-of-things enabled products); (Type-3) those that are seemingly immaterial but centralised (e.g. search engines or social media platforms reliant on network effects), or (Type-4) those that are seemingly immaterial and decentralised (e.g. the Bitcoin protocol).³³⁶

As they point out, while Type-4 technologies may be particularly hard for global regulators to capture, given that it is often unclear even which actors or intermediaries to regulate,³³⁷ many of the technologies in the other three categories can be at least somewhat controlled through a mix of strategies and approaches. As such, their model suggests that, far from ethereal, many ‘digital’ technologies are in fact sufficiently ‘material’ (e.g. satellites), or sufficiently ‘centralised’ (e.g. large search engines), that global regulation is possible, at least in principle, even if it may be challenged in a tense geopolitical environment.³³⁸

Given this framework, what type of digital technology is AI? Many AI applications, such as in social robots, (semi)autonomous drones, or camera systems with built-in (‘edge computing’) facial recognition capabilities might be considered as a Type-2 technology (material but decentralised). There may also be a large amount of AI services or tools that fall under Type-4 (immortal and decentralised). Nonetheless, the most salient and most disruptive ones likely will constitute Type-3 (immortal but centralised). These would include high-end applications of AI systems—such as Google Translate, YouTube recommender algorithms, or social media

famously told Time magazine that ‘[t]he Net interprets censorship as damage and routes around it.’ As quoted in Drezner, *supra* note 247 at 294.

³³² LAWRENCE LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE, VERSION 2.0 (2nd Revised ed. edition ed. 2006), <http://coderv2.cc/download+remix/Lessig-Codev2.pdf>.

³³³ JOSEPH S. NYE, *The Regime Complex for Managing Global Cyber Activities* (2014), <https://dash.harvard.edu/bitstream/handle/1/12308565/Nye-GlobalCommission.pdf> (last visited Sep 3, 2019).

³³⁴ Drezner, *supra* note 247 at 299.

³³⁵ Guillaume Beaumier et al., *Global Regulations for a Digital Economy: Between New and Old Challenges*, n/a GLOB. POLICY (2020), <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12823> (last visited May 14, 2020).

³³⁶ *Id.* at 3–6.

³³⁷ Even here, there are steps that states can take, as seen by the fact that many private cryptocurrency-to-state-backed-currency exchanges now enforce identification requirements. *Id.* at 6.

³³⁸ *Id.* at 6.

content moderation systems. Such applications are at least somewhat constrained by hardware and capital requirements, or are particularly dependent on network effects.

However, it is interesting to note that key elements of the broader global AI development pipeline might well be more ‘material and centralised’ (Type-1) than is often assumed. While it is true that many forms of AI software development are ‘diffuse’, as argued by Scherer,³³⁹ this does not tell the full story. Specifically, while the software or talent layers of AI development are relatively may be globally dispersed (excepting some pockets), the underlying computing hardware supply chain is much more globally concentrated by comparison.³⁴⁰ This suggests that underlying hardware could offer a potentially actionable levers for national export controls or (international) regulation.³⁴¹ As noted by one study:

“Of the various inputs into AI development (including hardware, software, data, and human effort), it’s worth noting that hardware is uniquely governable, at least in principle. Computing chips, no matter how fast, can perform only a finite and known number of operations per second, and each one has to be produced using physical materials that are countable, trackable, and inspectable. Similarly, physical robots rely on supply chains and materials that are in principle trackable. Computing power and hardware platforms for robotics are thus potentially amenable to some governance tools used in other domains that revolve around tracking of physical goods (e.g., export controls and on-site inspections).”³⁴²

Of course, the relative efficacy of such levers may depend sensitively on ongoing trends in the relative importance of algorithms, data, and computing hardware to various AI applications.³⁴³ Still, this tentatively suggests that, *once negotiated and implemented*, global governance regimes for AI could be *enforceable and effective* in at least some ways.

However, that says little about the path towards achieving global agreement and establishing such regimes in the first place. As the last few years have shown, this does not mean that global regulation of these systems will come easy. Given such challenges, and the incipient stage of governance, there is understandably a growing body of work exploring how to best develop AI governance, where to ground it, and where it could or should go next.

³³⁹ Scherer, *supra* note 298 at 370–371.

³⁴⁰ Likewise, datacenters (providing AI-as-a-service solutions) may be positioned at an intermediate point on this spectrum, as they are more concentrated than the software or human talent layers, but more distributed than computer chip production, because proximity to customers matters more for them. I thank Miles Brundage for a number of these observations and comments.

³⁴¹ Cf. BUCHANAN, *supra* note 91; Hwang, *supra* note 94. See also Leung, *supra* note 241. (arguing that we will likely see an increase in state inclination and capacity to exert control over the development and proliferation of AI, including through tools such as export controls). See also Jade Leung, Sophie-Charlotte Fischer & Allan Dafoe, *Export Controls in the Age of AI*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/> (last visited Sep 2, 2019). See also CARRICK FLYNN, *Recommendations on Export Controls for Artificial Intelligence* (2020), <https://cset.georgetown.edu/research/recommendations-on-export-controls-for-artificial-intelligence/> (last visited Sep 26, 2020).

³⁴² Brundage et al., *supra* note 116 at 69.

³⁴³ See also Tucker, Anderljung, and Dafoe, *supra* note 227. (on the social and governance implications of improved data efficiency).

2.3 State: the global AI governance landscape is insufficient

Having sketched the contours of the problem, what solutions are available? What is the available or emerging AI governance landscape, and is it up to the task? It is useful to take a step back, and examine the broader developments in the AI governance landscape that have gone on, as these highlight problems and opportunities. This field is currently still in flux, so the aim here is not to provide a comprehensive or definitive overview, but to sketch main themes and trajectories. Accordingly, we will first discuss (2.3.1) the extent to which AI issues might be accommodated under various norms, frameworks or instruments of existing international law; before discussing (2.3.2) a range of recent developments and movements in AI-specific ethics and governance initiatives and institutions.

2.3.1 Extending or applying existing International Law

The first question to ask is whether—or how—AI technology could be governed within existing public international law. This question is key, because the global legal order has a long and significant pedigree which will not (and arguably should not) be discarded.³⁴⁴ For all its current challenges, fragmentation and ‘stagnation’,³⁴⁵ it is likely to remain a lynchpin or at least important core substrate of the global governance architecture for some time.³⁴⁶

Traditional accounts of public international law hold that it is primarily produced through interactions by *states*, which can, in exercising their sovereignty, voluntarily enter into (i.e. consent to) international agreements. However, in more modern times, its subjects also include private individuals, companies, international institutions, and Non-Governmental Organizations (NGOs).

The *sources* of international law are set out in Article 38(1) of the Statute of the International Court of Justice (ICJ).³⁴⁷ These are; international treaties,³⁴⁸ international custom, and the general principles of law recognized by civilized nations.³⁴⁹ Subsidiary to this, it recognizes judicial decisions, and the teachings of the most highly qualified scholars of nations.³⁵⁰ Naturally, each of these approaches has a rich subset of tools. Moreover, there have also been

³⁴⁴ Eyal Benvenisti & George W. Downs, *Comment on Nico Krisch, “The Decay of Consent: International Law in an Age of Global Public Goods”*, 108 AJIL UNBOUND 1–3, 2 (2014) (“Historically, reports about the death of international law are invariably premature.”).

³⁴⁵ J. Pauwelyn, R. A. Wessel & J. Wouters, *When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking*, 25 EUR. J. INT. LAW 733–763 (2014).

³⁴⁶ Karen J. Alter, *The Future of International Law*, 101 ICOURTS WORK. PAP. SER. (2017), <https://papers.ssrn.com/abstract=3015177> (last visited Jun 11, 2020).

³⁴⁷ UN, *Statute of the International Court of Justice*, 33 UNTS 993 (1946), <https://www.refworld.org/docid/3deb4b9c0.html> (last visited Sep 1, 2020). Article 38(1).

³⁴⁸ International treaties may be convened between States, States and international organizations or amongst international organizations, and can come in various forms. They are governed under the UNITED NATIONS, *Vienna Convention on the Law of Treaties*, 1155 UNTS 331 (1969); As well as by the UNITED NATIONS, *Vienna Convention on the Law of Treaties Between States and International Organizations or Between International Organizations*, 25 ILM 543 (1989).

³⁴⁹ UN, *supra* note 347. Article 38(1).

³⁵⁰ *Id.* Article 38(1). However, the International Law Commission has appended two more sources to this list, namely binding decisions of international organizations and the judgments of international courts and tribunals. International Law Commission, *Draft Articles on State Responsibility, Part 2* (1989). Article 5(1).

shifts in emphasis amongst these instruments over time: for instance, some have suggested that in recent decades, the role of customary international law has been in decline relative to the increasing role of international regulation by treaty.³⁵¹

Given this background framework, we can now map the application of various norms or regimes to AI. As such, in the following sections, we will first discuss (2.3.1.1) the general status of new technologies in international law, before reviewing the strengths and limits of (2.3.1.2) Customary International Law, (2.3.1.3) International Human Rights Law, (2.3.1.4) (other) existing treaty instruments and regimes, and (2.3.1.5) adjudication under various international courts.

2.3.1.1 New Technologies in International Law

In the first place, what is international law's relation to 'technology'? Doctrinally speaking: none. As Rayfuse notes, barring direct harms, "international law's concern with, or interest in, regulating technologies lies not in their inherent nature, form, development, or even deployment."³⁵² That is to say, emerging technologies do not have a distinct status under international law; and while there are certainly a wide range of technology-specific regimes to which states can and have acceded,³⁵³ there does not exist a single legally binding global treaty regime to specifically regulate all technological risks as a category.³⁵⁴

This functional technology-neutrality is reflected, for instance, in the International Court of Justice's 1996 *Advisory Opinion on the Legality of the Threat or Use of Nuclear Weapons*, which held that, in the absence of specific treaty obligations freely consented to by states, the mere development of nuclear weapons is not inherently unlawful under international law.³⁵⁵ In fact, under this opinion, even the *use* of nuclear weapons is not inherently unlawful, given certain highly circumscribed circumstances—specifically the "extreme circumstance of self defense, in

³⁵¹ See for instance Andrew T. Guzman, *Saving Customary International Law*, 27 MICH. J. INT. LAW 115, 1000 (2005) (arguing that treaties have become a more important tool of international relations than customary international law). See also Joel P. Trachtman, *The Growing Obsolescence of Customary International Law*, in CUSTOM'S FUTURE: INTERNATIONAL LAW IN A CHANGING WORLD 172–204 (Curtis A. Bradley ed., 2016), /core/books/customs-future/the-growing-obsolescence-of-customary-international-law/21B7994612971B0EFEE1378B916DDBFE (last visited Apr 17, 2019) (arguing that customary international law may not be sufficient for modern challenges). However, see also the more nuanced discussion in: Rebecca Crootof, *Jurisprudential Space Junk: Treaties and New Technologies*, in RESOLVING CONFLICTS IN THE LAW 106–129, 126–128 (Chiara Giorgetti & Natalie Klein eds., 2019), <https://brill.com/view/book/edcoll/9789004316539/BP000015.xml> (last visited Mar 15, 2019).

³⁵² Rayfuse, *supra* note 219 at 506.

³⁵³ Picker, *supra* note 33 at 164–181. There are, of course, a plethora of proposals for distinct new treaty regimes specific to various technologies. Mette Eilstrup-Sangiovanni, *Why the World Needs an International Cyberwar Convention*, 31 PHILOS. TECHNOL. 379–407 (2018); Karen Scott, *International Law in the Anthropocene: Responding to the Geoengineering Challenge*, 34 MICH. J. INT. LAW 309–358 (2013). For a discussion of governance for high-stakes risks from AI within the UN framework, see Reinmar Nindler, *The United Nation's Capability to Manage Existential Risks with a Focus on Artificial Intelligence*, 21 INT. COMMUNITY LAW REV. 5–34 (2019). See also Crootof, *supra* note 351 at 111–112. (reviewing proposed bans for diverse technologies).

³⁵⁴ Rayfuse, *supra* note 219 at 503. For a proposal for such a single global regime to cover extreme technological risks, see Grant Wilson, *Minimizing global catastrophic and existential risks from emerging technologies through international law*, 31 VA ENVTL LJ 307 (2013).

³⁵⁵ INTERNATIONAL COURT OF JUSTICE, LEGALITY OF THE THREAT OR USE OF NUCLEAR WEAPONS: ADVISORY OPINION OF 8 JULY 1996 (1996), <https://www.icj-cij.org/files/case-related/95/095-19960708-ADV-01-00-EN.pdf>. Note that advisory opinions are not binding.

which the very survival of a State would be at stake”³⁵⁶—and provided that the state using them otherwise complies with the laws of armed conflict under international law.

Of course, even if this does not categorically outlaw nuclear weapons, such a reading might still be held to *functionally* outlaw certain types of nuclear weapon assemblages. For instance, Colin Picker has argued that;

“[w]hile [the ICJ] opinion arguably leaves plenty of leeway for the subjective determination by a state as to when it feels its very survival is at stake, the opinion would appear, at a minimum, to argue against the legitimacy under international law of tactical nuclear weapons, and possibly of strategic nuclear weapons under many of the circumstances for which they were designed.”³⁵⁷

All this is not to say that states could not consent to treaties focused on categorically outlawing the use or development of nuclear weapons technology: indeed, the adoption of the 2017 Treaty on the Prohibition of Nuclear Weapons imposes exactly such (nuclear) technology-specific treaty obligations (on its signatories).³⁵⁸ The point here, however, is that, barring state consent to such technology-specific treaties, international law does not *prima facie* take an interest in a new technology, just for its being new. Rather, it responds to (potential) new uses of that technology that violate pre-existing norms.³⁵⁹

Of course, it may well be possible or likely that the use of AI technologies might violate some of these existing norms, and a key point is that international law is able to evolve and adapt. Admittedly, this may not occur quickly since, as Rosemary Rayfuse has noted, “the traditional approach of international law to the regulation of emerging technologies has been one of reaction rather than pro- action; only attempting to evaluate and regulate their development or use *ex post facto*.³⁶⁰ Nonetheless, once roused, international law has also proven quite creative at governing new technologies.³⁶¹ Existing international legal instruments can be adapted in various ways to new (technological) changes: through treaty amendment, adaptive interpretation, or (in some circumstances) new state behaviour that results in customary international law.³⁶² This means we should next ask; could AI technology be subsumed under general norms or provisions of international law? Alternately, might existing treaty instruments be extended or re-interpreted to cover it?

³⁵⁶ *Id.* at 266.

³⁵⁷ Picker, *supra* note 33 at 175.

³⁵⁸ United Nations, *United Nations Conference to Negotiate a Legally Binding Instrument to Prohibit Nuclear Weapons, Leading Towards Their Total Elimination* (2017), <https://www.un.org/disarmament/publications/library/ptnw/> (last visited Sep 4, 2020). The treaty was opened for signature 9 August 2017. In order to come into effect, it will require signature and ratification by at least 50 countries. As of August 2020, 44 states have ratified it.

³⁵⁹ The same goes for other technologies. For instance, Rayfuse discusses how “the development and use of environmental modification technologies is neither regulated nor prohibited under international law, but only their hostile use in the context of an international armed conflict”, under the 1976 Convention on the Prohibition of Military or Any Other Hostile Use for Environmental Modification Techniques. Rayfuse, *supra* note 219 at 507.

³⁶⁰ *Id.* at 500.

³⁶¹ Picker, *supra* note 33.

³⁶² Crootof, *supra* note 351; Rebecca Crootof, *Change Without Consent: How Customary International Law Modifies Treaties*, 41 YALE J. INT. LAW 65 (2016). These avenues of ‘development’ are discussed in greater detail in Chapter 5.3.6.

2.3.1.2 Customary International Law

As noted, barring explicit treaty commitments to the contrary, states are under international law free to develop and deploy new technologies at will. As Rayfuse observes, “in general, the principle of state sovereignty permits states to utilize their resources, conduct research, and develop and deploy, or allow their nationals to research, develop, and deploy, technologies as they see fit.”³⁶³ Even so, in developing any new technology, states are still bound by the established principles of customary international law (CIL).

As established in the ICJ Statute, customary international law is considered a source of states’ international legal obligations coequal with treaties.³⁶⁴ Customary international law is considered to exist when states generally and habitually engage in certain actions (there is a ‘state practice’ element), and when they do so out of a belief that those practices are legally obligatory or permitted (e.g. bound by law, the so-called *opinion juris sive necessitates* element).³⁶⁵ As Rayfuse notes, these norms circumscribe any usage of any new technology (and indeed any state behaviour), and they include:

“the basic norms of international peace and security law, such as the prohibitions on the use of force and intervention in the domestic affairs of other states [...]; the basic principles of international humanitarian law, such as the requirements of humanity, distinction and proportionality [...]; the basic principles of international human rights law, including the principles of human dignity and the right to life, liberty, and security of the person [...]; and the basic principles of international environmental law, including the no-harm principle, the obligation to prevent pollution, the obligation to protect vulnerable ecosystems and species, the precautionary principle, and a range of procedural obligations relating to cooperation, consultation, notification, and exchange of information, environmental impact assessment, and participation [...]. The general customary rules on state responsibility and liability for harm also apply.”³⁶⁶

In addition, state sovereignty is subject to the obligation, on all states, to ensure that activities under their jurisdiction and control do not cause harm to other states, the nationals of other states, the environment or the global commons.³⁶⁷

Furthermore, states are also bound by the fundamental norms of *ius cogens*, defined in the Vienna Convention on the Law of Treaties as “...a norm accepted and recognized by the international community of States as a whole as a norm from which no derogation is permitted and which can be modified only by a subsequent norm of general international law having the same character.”³⁶⁸ These peremptory norms outlaw, amongst others, genocide, maritime piracy,

³⁶³ Rayfuse, *supra* note 219 at 506.

³⁶⁴ Crootof, *supra* note 362 at 273.

³⁶⁵ As Crootof notes, “a rule of customary international law is authoritative because states generally abide by it in the belief that it is law.” *Id.* at 1.

³⁶⁶ Rayfuse, *supra* note 219 at 503.

³⁶⁷ International Court of Justice, *Corfu Channel (Merits) (UK v Albania)* (1949); INTERNATIONAL COURT OF JUSTICE, *supra* note 355 at 29. And see Rayfuse, *supra* note 219 at 507–508.

³⁶⁸ UNITED NATIONS, *supra* note 348. Article 53.

slavery, wars of aggression or territorial aggrandizement, torture, and refoulement.³⁶⁹ These foundational norms bracket any use or deployment of AI—or indeed, of any other new technology—and therefore could provide one key backstop.³⁷⁰ Importantly, while some scholars have contended that customary international law has in recent decades receded in importance relative to treaties, or may even not be fit for many modern legal challenges,³⁷¹ it could offer distinct benefits in regulating AI innovation. For instance, Rebecca Crootof has suggested that, in the context of rapidly changing technologies, relying on customary international law may offer benefits over treaties, because;

“customary international law is universally applicable, which allows states to solve the potential holdout problem associated with any system grounded on consent. [...] Relatedly, there is little risk of fragmentation. As opposed to treaty law’s opt-in approach, customary international law has a limited opt-out option: at least theoretically, a state may avoid being bound by a developing customary international law rule by consistently contesting the rule’s existence or applicability. In practice, there are few such examples of this occurring.”³⁷²

Nonetheless, the specific application of customary international law to many AI use cases or challenges may leave much to be desired. This is for two reasons: (1) it is often unclear if or how these norms actually apply to AI use cases; and (2) it is slow.

In the first place, the basic principles and rules of customary international law may be binding on all states, but they are broad, and only provide a basic framework.³⁷³ It can at times be surprisingly unclear or contested if, how or why AI applications would actually be bound by these customary international law principles.

To see why and how, one can turn to the diverse and extensive debates over this question in the context of norms around military uses of AI. While in principle, all use of AI in conflict is bound by the provisions of International Humanitarian Law (IHL), in practice there are many uncertainties around how and when these open-ended provisions would apply with concrete effects.³⁷⁴

³⁶⁹ M. Cherif Bassiouni, *International Crimes: Jus Cogens and Obligatio Erga Omnes*, 59 LAW CONTEMP. PROBL. 63–74, 68 (1996). Notably, *Jus Cogens* creates *erga omnes* obligations for states to comply with a rule. The UN’s International Law Commission has codified the *erga omnes* principle in its draft articles on State responsibility as it, in article 48 (1)(b), allows all States to invoke a State responsibility which another State incurred due to its unlawful actions, if “the obligation breached is owed to the international community as a whole”. INTERNATIONAL LAW COMMISSION, *Draft articles on Responsibility of States for Internationally Wrongful Acts* 114 (2001), https://legal.un.org/ilc/texts/instruments/english/commentaries/9_6_2001.pdf Article 48 (1)(b).

³⁷⁰ That does not mean that further treaty provisions may not be needed to clarify or affirm the scope or applicability of existing norms and instruments. See for instance the discussion of AI interrogation tools in Amanda McAllister, *Stranger than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture*, MINN. LAW REV. 47 (2018).

³⁷¹ Trachtman, *supra* note 351 at 172.

³⁷² Crootof, *supra* note 351 at 127 (citing sources). For another defense of customary international law (specifically in addressing global commons problems around large-scale disasters, see Aaken, *supra* note 309.

³⁷³ Rayfuse, *supra* note 219 at 507.

³⁷⁴ See also Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOB. POLICY 370–375 (2019); Esther Chavannes, Klaudia Klonowska & Tim Sweijns, *Governing autonomous weapon systems: Expanding the solution space, from scoping to applying*, HCSS SECUR. 39 (2020).

For instance, there has been much debate over whether today's autonomous weapons systems fail to meet the *Principle of Distinction*—or if they do, for how much longer that will remain the case as AI capabilities advance.³⁷⁵ Other IHL principles also may not provide effective guidance. For instance, it has been argued that the *Principle of Precaution in Attack* might even speak in favour of the use of (predictive) algorithms, if their decisions became more discriminate than those of human personnel, or if their use would allow for less lethal attacks.³⁷⁶ Alternatively, one can take the *Principle of Humanity*, one of the oldest principles of IHL, laid down in the 'Martens Clause'. As phrased in the 1977 Additional Protocol to the Geneva Conventions, this clause states that:

In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience.³⁷⁷

However, there remains extensive uncertainty over the meaning of many of these terms, including the 'dictates of public conscience' in relation to AI-based weapons.³⁷⁸ For instance, while global opinion surveys do show significant and rising opposition to the use of LAWS,³⁷⁹ such judgments can be contextual and subject to framings.³⁸⁰ More generally, even if consistent public condemnation were to be achieved for LAWS, it has been argued that states could simply innovate to improve military AI system's ostensible adherence to societal norms.³⁸¹ This could be especially the case for less visceral 'lethality-enabling' applications, such as in target prediction.³⁸²

³⁷⁵ Rosert and Sauer, *supra* note 374; Ben Koppelman, *How Would Future Autonomous Weapon Systems Challenge Current Governance Norms?*, 164 RUSI J. 98–109 (2019); Allenby, *supra* note 306.

³⁷⁶ Chavannes, Klonowska, and Sweijs, *supra* note 374 at 12–13. See also Rebecca Crotof, *Regulating New Weapons Technology*, in THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 1–25, 11–12 (Eric Talbot Jensen & Ronald T.P. Alcala eds., 2019).

³⁷⁷ United Nations, *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, 8 June 1977. 126 (1977). Article 1(2).

³⁷⁸ See discussion over this debate in Matthijs M. Maas, *Innovation-Proof Governance for Military AI? How I learned to stop worrying and love the bot*, 10 J. INT. HUMANIT. LEG. STUD. 129–157, 144–146 (2019). Paper [II].

³⁷⁹ Heather M. Roff, *What Do People Around the World Think About Killer Robots?*, SLATE, 2017, http://www.slate.com/articles/technology/future_tense/2017/02/what_do_people_around_the_world_think_about_killer_robots.html (last visited Apr 5, 2018); IPSOS, *Six in Ten (61%) Respondents Across 26 Countries Oppose the Use of Lethal Autonomous Weapons Systems* (2019), https://www.ipsos.com/sites/default/files/ct/news/documents/2019-01/human-rights-watch-autonomous-weapons-pr-01-22-2019_0.pdf (last visited Jan 27, 2019); IPSOS, *Three in Ten Americans Support Using Autonomous Weapons*, IPSOS (2017), <https://www.ipsos.com/en-us/news-polls/three-ten-americans-support-using-autonomous-weapons> (last visited Jan 13, 2019); OPEN ROBOETHICS INITIATIVE, *The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll* (2015), http://www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf (last visited Apr 5, 2018).

³⁸⁰ Michael C. Horowitz, *Public opinion and the politics of the killer robots debate*, 3 RES. POLIT. 2053168015627183 (2016); Darrell M. West, *Brookings survey finds divided views on artificial intelligence for warfare, but support rises if adversaries are developing it*, BROOKINGS (2018), <https://www.brookings.edu/blog/techtank/2018/08/29/brookings-survey-finds-divided-views-on-artificial-intelligence-for-warfare-but-support-rises-if-adversaries-are-developing-it/> (last visited Sep 21, 2018).

³⁸¹ Chavannes, Klonowska, and Sweijs, *supra* note 374 at 13. See also Maas, *supra* note 378 at 146.

³⁸² Arthur Holland Michel, *The Killer Algorithms Nobody's Talking About*, FOREIGN POLICY (2020), <https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobody-s-talking-about/> (last visited Jan 21, 2020); Ashley Deeks, *Predicting Enemies*, 104 VA. LAW REV. 1529–1593 (2018).

Another avenue could conceivably be found in the existing legal review mechanism for new weapons provided for under Article 36 of Additional Protocol I to the Geneva Conventions.³⁸³ However, this mechanism is already hindered by a lack of widespread or transparent implementation by states.³⁸⁴ Moreover, its application to military AI technologies might be impeded by the conceptual difficulties around determining whether reviews need only cover directly weaponised systems, or whether these should also extend to related non-weaponised AI technologies in the much broader operational network and infrastructure that assists in the targeting processes.³⁸⁵

Moving from IHL to *jus ad bellum*, one might also appeal to the general international prohibition on the use of force. For instance, if the deployment of AI systems erodes ‘strategic stability’ or sets the conditions for an unintended ‘flash war’,³⁸⁶ or conversely if conflict prediction algorithms lower the threshold to war,³⁸⁷ this could result in premature or unlawful use of force, and the violation of states’ obligations to sustain peace and security.³⁸⁸ Such norms could therefore be interpreted to outlaw the incorporation of (semi)autonomous systems into automated self-defence systems that could result in states unwillingly breaching these obligations.

Chavannes and others have also noted that the deployment of autonomous weapons systems could lead to problems with attributing accountability for wrongful acts.³⁸⁹ This is for two reasons. On the one hand, it can lead to problems over establishing *state accountability*. The International Law Commission’s principles for the Responsibility of States for Internationally Wrongful Acts note that the only conduct attributable to states is that of government organs or of people acting under these organs’ “direction, instigation or control” as “agents of the State”.³⁹⁰ However, the complicated development and usage life cycle of military AI systems may blur the responsibilities of the various actors involved in the development, deployment and use of force, specifically making it harder to establish which actors should be counted as acting within the official command structure.³⁹¹ The other problem with attribution lies in the inability to impose

³⁸³ ARTICLE 36 - NEW WEAPONS - PROTOCOL ADDITIONAL TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, AND RELATING TO THE PROTECTION OF VICTIMS OF INTERNATIONAL ARMED CONFLICTS (PROTOCOL I), 36 (1977).

³⁸⁴ Denise Garcia, *Future arms, technologies, and international law: Preventive security governance*, 1 EUR. J. INT. SECUR. 94–111, 105 (2016); Chavannes, Klonowska, and Sweijs, *supra* note 374 at 13–14.

³⁸⁵ Hin-Yan Liu, *Categorization and legality of autonomous and remote weapons systems*, 94 INT. REV. RED CROSS 627–652, 628 (2012); Chavannes, Klonowska, and Sweijs, *supra* note 374 at 13–14.

³⁸⁶ Jürgen Altmann & Frank Sauer, *Autonomous Weapon Systems and Strategic Stability*, 59 SURVIVAL 117–142 (2017); Paul Scharre, *Flash War - Autonomous Weapons and Strategic Stability* (2016), <http://www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf> (last visited Nov 17, 2017); Scharre, *supra* note 291.

³⁸⁷ Deeks, Lubell, and Murray, *supra* note 291.

³⁸⁸ Chavannes, Klonowska, and Sweijs, *supra* note 374 at 14.

³⁸⁹ *Id.* at 14–15.

³⁹⁰ INTERNATIONAL LAW COMMISSION, *supra* note 369 at 38. Para 2.

³⁹¹ However, Thomas Burri has argued these conceptual problems may be overstated, and suggests that international law has more than sufficient precedent to establish the meaning of ‘control’ over AI systems, or the limits of delegation. Specifically, he argues it can draw on extensive case law both in modern international criminal law (especially many of the cases before the International Criminal Tribunal for the Former Yugoslavia), as well as older cases by the Permanent Court of International Justice (such as its 1935 ruling on ‘*Consistency of Certain Danzig Legislative Decrees With the Constitution of the Free City*’). See Thomas Burri, *International Law and Artificial Intelligence*, 60 GER. YEARB. INT. LAW 91–108, 101–104 (2017).

criminal liability for wrongful acts on the autonomous weapons systems making the key decisions, since these systems lack the ‘intention’ to cause harm when they go awry.³⁹²

That is not to suggest that debates on the applicability of various international norms to military AI systems are intractable. Indeed, they are likely to result in greater clarity over time, and reach resolution one way or the other. Nonetheless, the conceptual difficulty and disagreement over whether these various norms apply to lethal autonomous weapons systems—the sharpest contact point between international law and AI technology to date—is a sign of the relative conceptual difficulty involved in fitting existing customary norms to AI technology. This would likely play up in many other domains of AI use.

In the second place, and more practically, customary international law norms also develop slowly. As Picker has noted, under good circumstances, customary international law has been created in as little as 10 to 15 years.³⁹³ On occasion, it has certainly been created more quickly, as in the rules regarding a state’s territorial seas.³⁹⁴ Yet often it has taken decades. To be sure, this speed has picked up. Yet while indeed certain AI applications, along with the broader data-collection abilities of the digital society could help speed up the identification of state practice,³⁹⁵ they could also muddle it, by making it easier for different (state) actors to identify and highlight counterexamples of state practice which favour their interests.³⁹⁶ More problematically, customary international law requires clear evidence of state practice, yet this may not always be available where it concerns cases of ‘hidden’ or ‘hard to attribute’ AI capabilities such as in cyberwarfare,³⁹⁷ and more generally may be hindered by definitional problems around AI systems and applications.³⁹⁸

All this suggest that there are at least some gaps and ambiguities in the general application of customary international norms. However, would it be possible to extend or adapt specific existing legal instruments or regimes to new AI uses?

2.3.1.3 International Human Rights Law

Another avenue for governing AI challenges might be found under the aegis of international human rights law (IHRL).³⁹⁹ As noted, recent years have seen a range of scholars

³⁹² For one suggested solution, see also Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 UNIV. PA. LAW REV. 1347–1402 (2016). Elizabeth Fuzaylova, *War Torts, Autonomous Weapon Systems, and Liability: Why a Limited Strict Liability Tort Regime Should be Implemented*, 40 CARDOZO LAW REV. 1327–1366 (2019). Note, this problem may not be limited to international criminal law, but may also arise in domestic law contexts. King et al., *supra* note 287.

³⁹³ Picker, *supra* note 33 at 185.

³⁹⁴ Crootof, *supra* note 362 at 250. (note 1: discussing how customary rules around states’ rights in adjacent waters evolved rapidly).

³⁹⁵ Tamar Megiddo, *Knowledge Production, Big Data and Data-Driven Customary International Law* (2019), <https://papers.ssrn.com/abstract=3497477> (last visited Jan 21, 2020).

³⁹⁶ Ashley Deeks, *High-Tech International Law*, 88 GEORGE WASH. LAW REV. 575–653, 649 (2020). See also Section 5.5.1. (on using AI for governance ‘displacement’ including the automation of rule creation), and 7.3.3.

³⁹⁷ See generally Rebecca Crootof, *International Cybertorts: Expanding State Accountability in Cyberspace*, 103 CORNELL LAW REV. 565–644 (2018).

³⁹⁸ Matthijs M. Maas, *International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order*, 20 MELB. J. INT. LAW 29–56, 53 (2019). Paper [III].

³⁹⁹ For the sake of clarity, it should be noted that under IHRL, Human Rights are primarily provided for in treaty law, notably within nine core international human rights instruments. Some of these rights are also considered

articulating or highlighting the way in which AI technologies could threaten various human rights.⁴⁰⁰ In response, there have also been growing calls by various scholars to ground AI governance in human rights frameworks, norms and instruments. For instance, Lorna McGregor and colleagues have suggested that IHRL could provide a superior framework to existing algorithmic accountability approaches, because it enables the shared understanding and means to assess harms, is able to deal with various actors and forms of responsibility, and can apply across the algorithmic life cycle.⁴⁰¹ Similarly, Karen Yeung and colleagues have argued in favour of a ‘human rights-centered design, deliberation, and oversight approach’ to AI governance,⁴⁰² and Silja Vöneky has called for a universal UN or UNESCO soft law declaration on ‘AI Ethics and Human Rights’.⁴⁰³ There are various other such proposals, which have in common that they are presented as being an improvement to the current voluntary (self)governance approaches to AI.⁴⁰⁴ As Eileen Donahoe and Megan Metzger argue, the human rights framework would have distinct advantages by comparison to these tools, because:

“1) It puts the human person at the center of any assessment of AI and makes impact on humans the focal point of governance; 2) it covers the wide range of pressing concerns that AI raises, both procedural and substantive; 3) it outlines the roles and duties of both governments and the private sector in protecting and respecting human rights; and 4) it has the geopolitical advantage of resting on a broad global consensus, shared by many countries around the world and understood to be universally applicable.”⁴⁰⁵

For instance, they note how various sections of the Universal Declaration of Human Rights (UDHR) address some of the diverse societal concerns that have been raised around AI, from challenges around avoiding data bias and ensuring fairness in machine-based decisions relied upon by governments (Article 2—the right to equal protection and non-discrimination),

to be a part of Customary International Law, and have reached *jus cogens* status (such as the prohibitions on torture, or the prohibition on discrimination). However, not all rights that are provided for in treaty law are also part of CIL. The Universal Declaration of Human Rights (UDHR) itself is a UN General Assembly declaration that does not create binding international human rights law, however, many of its provisions are considered to be part of CIL. I thank Linnéa Nordlander for discussing some of the nuances around IHRL. Any remaining errors are my own.

⁴⁰⁰ This literature is broad, but see: FILIPPO RASO ET AL., *Artificial Intelligence & Human Rights: Opportunities & Risks* (2018), https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf? (last visited Apr 8, 2019); EST, GERRITSEN, AND KOOL, *supra* note 277; Molnar, *supra* note 282. CATELLJNE MULLER, *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law* 20 (2020), <https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da>.

⁴⁰¹ Lorna McGregor, Daragh Murray & Vivian Ng, *International Human Rights Law as a Framework for Algorithmic Accountability*, 68 INT. COMP. LAW Q. 309–343 (2019).

⁴⁰² Karen Yeung, Andrew Howes & Ganna Pogrebna, *AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing*, in THE OXFORD HANDBOOK OF AI ETHICS 78–106 (M. Dubber & F. Pasquale eds., 2020), <https://papers.ssrn.com/abstract=3435011> (last visited Sep 3, 2019).

⁴⁰³ Silja Vöneky, *How Should We Regulate AI? Current Rules and Principles as Basis for “Responsible Artificial Intelligence”* (2020), <https://papers.ssrn.com/abstract=3605440> (last visited Sep 1, 2020).

⁴⁰⁴ Some of these approaches will be discussed further on, in Section 2.3.2.2.

⁴⁰⁵ Eileen Donahoe & Megan MacDuffee Metzger, *Artificial Intelligence and Human Rights*, 30 J. DEMOCR. 115–126, 119 (2019). See also Vöneky, *supra* note 403. For an exploration of what a human-rights based approach to AI would entail, and what would be some of its pitfalls, see also Nathalie A. Smuha, *Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea*, PHILOS. TECHNOL. (2020), <https://doi.org/10.1007/s13347-020-00403-w> (last visited May 25, 2020).

concerns around autonomous weapons (Article 3—the right to life and personal security), concerns around privacy in the face of AI surveillance (Article 12—the right to privacy), the challenges around algorithmic content moderation (Article 19—the right to freedom of expression), or the displacement of human workers by AI (Articles 23 and 25—the rights to work and to enjoy an adequate standard of living).⁴⁰⁶

To be certain, there are benefits to human rights-based frameworks, particularly compared to existing soft law or self-governance initiatives, on substantive, effectiveness and legitimacy grounds.⁴⁰⁷ Nonetheless, IHRL may also face a series of shortcomings or challenges in regulating AI. For one, even its advocates concede that while human rights norms are comprehensive, they still do not cover all ethical values that are implicated or threatened by AI.⁴⁰⁸ In the second place, the framework may face challenges in reorienting towards the considerable role of private technology companies in AI: whereas under IHRL, states have a direct obligation to prevent and protect human rights, the scope and content of private businesses' human rights responsibilities remains more underdeveloped, as they are currently only subject to 'expectations' to respect human rights.⁴⁰⁹

Underlying this, it has been argued that the opacity of potentially harmful AI systems might more broadly erode any rights-based responsibility and accountability mechanisms.⁴¹⁰ Indeed, Liu has generalised this argument, suggesting that several features of digital technologies—such as their status as complex systems vulnerable to 'normal' accidents; their integration in networks enabling patterned and cumulative effects; and their tendency to exert probabilistic or self-reinforcing impacts, and to shift and exacerbate power differentials—blur certain foundations or assumptions of the human rights toolbox at a core level.⁴¹¹

⁴⁰⁶ Donahoe and Metzger, *supra* note 405 at 120.

⁴⁰⁷ For another discussion of the shortcomings of both private sector self-governance approaches, but also of the existing human rights vernacular, and arguments of how the latter might be reconfigured in order to rethink human rights regulation for AI, see also Sue Anne Teo, *Artificial Intelligence and Corporate Human Rights Self-Regulation Initiatives: The Dangers of Letting Business go on as Usual*, in (WORKING PAPER) (2020). (on file with author).

⁴⁰⁸ Yeung, Howes, and Pogrebna, *supra* note 402 at 102–103. For a more fundamental philosophical critique, see John Tasioulas, *First Steps Towards an Ethics of Robots and Artificial Intelligence*, 7 J. PRACT. ETHICS 61–95, 69–70 (2019). ("Leave aside the fact that this law does not reflect all of the ethical considerations (e.g. environmental values) bearing on AI, that it does not tend to be directly binding on non-state actors, and that not all of its provisions bind all states (e.g. because they have not ratified relevant human rights treaties). The more fundamental point is that such laws—despite the powerful moral charge conferred by the words 'human rights'—are not basic ethical standards. Instead, like any other set of laws, they are themselves to be formulated and evaluated—and, sometimes, to be found seriously wanting—in terms of basic ethical standards, including the morality of human rights").

⁴⁰⁹ As per in Article 11 of the UN Guiding Principles on Business and Human Rights, adopted by the UN Human Rights Council in 2011. UNITED NATIONS, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* 13 (2011), https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf (last visited Sep 17, 2020). See also the discussion by McGregor, Murray, and Ng, *supra* note 401 at 313. But see also Donahoe and Metzger, *supra* note 405 at 120–121. (more optimistic about the ability of these UN Guiding Principles to provide guidance to IHRL in addressing AI issues).

⁴¹⁰ Hin-Yan Liu & Karolina Zawieska, *A New Human Rights Regime to Address Robotics and Artificial Intelligence*, in 2017 PROCEEDINGS OF THE 20TH INTERNATIONAL LEGAL INFORMATICS SYMPOSIUM 179–184 (2017).

⁴¹¹ See Hin-Yan Liu, *The Digital Disruption of Human Rights Foundations*, in HUMAN RIGHTS, DIGITAL SOCIETY AND THE LAW: A RESEARCH COMPANION (2019). He explores the challenges of patterned discriminatory effects in

That does not mean that such challenges could not be overcome, of course. However, advocates of this approach have noted and urged that a human rights-based governance frameworks for AI will need to take a series of steps to clarify the applicability of human rights, strengthen enforcement mechanisms, and to secure the underlying societal infrastructures or norms—such as democratic institutions and the rule of law—which enable human rights in the first place.⁴¹² That is surely a steep challenge, especially in a geopolitical landscape in which human rights have more broadly come under pressure.

2.3.1.4 Other existing Treaty Instruments

Beyond IHRL, there are also many other issue- or domain-specific regimes that could be applied to AI. Over the years, the global legal architecture has developed a dense and diverse ecology of treaty law, potentially offering a diverse array of domain-specific norms, standards or treaty instruments that could plausibly be applied or extended to these challenges.

Martina Kunz & Seán Ó hÉigearthaigh have explored in detail how various international law regimes might cover or be extended to AI, focusing specifically on the domains of safety and security.⁴¹³ They identify diverse existing governance regimes that have begun consultation on AI issues, or which could provide at least some degree of regulatory scaffolding in these areas.⁴¹⁴ For instance, they survey a range of provisions that might apply to transborder mobile robot governance, covering systems in the air, at sea, or on land.

They argue that provisions for *autonomous airborne drones* might derive from existing instruments for the safety and security of civil aircraft. These include provisions on ‘pilotless aircraft’ already embedded in the 1944 Chicago Convention on International Civil Aviation;⁴¹⁵ ongoing guidance documents on drone safety and security produced by the International Civil Aviation Organization (ICAO);⁴¹⁶ the 1970 Hague Convention for the Suppression of Unlawful Seizure of Aircraft and its supplementary 2010 Beijing Protocol;⁴¹⁷ and the 2010 Beijing Convention on the Suppression of Unlawful Acts Relating to International Civil Aviation.⁴¹⁸

In terms of international regimes for *autonomous maritime vessels*, the International Maritime Organization (IMO) has begun to investigate legal aspects of the safe development of

greater detail in Liu, *supra* note 277. As well as the idea of power differentials in greater detail in Hin-Yan Liu, *The power structure of artificial intelligence*, 10 LAW INNOV. TECHNOL. 197–229 (2018).

⁴¹² See Smuha, *supra* note 405 at 12–13.

⁴¹³ Kunz and Ó hÉigearthaigh, *supra* note 216. For articles that examine the potential of international law instruments to address more speculative longer-term challenges from AI, see also J.G. Castel & Mathew E. Castel, *The Road to Artificial Superintelligence - has international law a role to play?*, 14 CAN. J. LAW TECHNOL. (2016), <https://ojs.library.dal.ca/CJLT/article/download/7211/6256>; Nindler, *supra* note 353.

⁴¹⁴ These next few paragraphs draw at length on the excellent survey by Kunz and Ó hÉigearthaigh, *supra* note 216.

⁴¹⁵ CONVENTION ON INTERNATIONAL CIVIL AVIATION, 15 UNTS 295 (1944). Article 8. See Kunz and Ó hÉigearthaigh, *supra* note 216 at 3–4.

⁴¹⁶ Kunz and Ó hÉigearthaigh, *supra* note 216 at 4. (citing sources).

⁴¹⁷ Which includes criminalization of remote seizure or exercise of control of an aircraft ‘by any technological means’. PROTOCOL SUPPLEMENTARY TO THE CONVENTION FOR THE SUPPRESSION OF UNLAWFUL SEIZURE OF AIRCRAFT, 50 ILM 153 (2011). Article 1(1). As discussed by Kunz and Ó hÉigearthaigh, *supra* note 216 at 4.

⁴¹⁸ CONVENTION ON THE SUPPRESSION OF UNLAWFUL ACTS RELATING TO INTERNATIONAL CIVIL AVIATION, 50 ILM 144 (2011); Kunz and Ó hÉigearthaigh, *supra* note 216 at 4–5.

remote-controlled, autonomous, or semi-autonomous surface ships.⁴¹⁹ While many legal uncertainties remain, scholars exploring the security implications of maritime autonomous vehicles under international law have broadly argued that while such vehicles raise some new questions with regards to issues such as the Right of Hot Pursuit or the Right of Visit, in most cases, existing legal principles might be applied to AI-controlled vessels without major problem.⁴²⁰

In terms of *autonomous vehicles*, there have been recent efforts at modernizing existing frameworks,⁴²¹ though this has been complicated by the existing split between the 1949 Geneva Convention on Road Traffic⁴²² and the 1968 Vienna Road Traffic Convention.⁴²³ Nonetheless, as also noted by Bryant Smith, the Global Forum's 2018 Resolution on the Deployment of Highly and Fully Automated Vehicles in Road Traffic has provided one of a number of signs that conceptual uncertainties are being pragmatically resolved.⁴²⁴

Moreover, Kunz & Ó hÉigearthaigh explore the applicability of a range of cross-domain *counterterrorism* regimes that could affect AI-related security challenges, arguing that the 1997 Convention for the Suppression of Terrorist Bombings might outlaw, for instance, the terrorist use of agricultural drones to spray chemicals on crowds.⁴²⁵ Likewise, they posit that several United Nations Security Council resolutions have application in restricting potential robotic delivery systems for terrorism, and specifically for Weapons of Mass Destruction.⁴²⁶

Other mobile robotic systems, they suggest, might be subsumed under various *conventional arms control* regimes, including the 1987 Missile Technology Control Regime (MTCR), the 1996 Wassenaar Arrangement, or the 2013 Arms Trade Treaty (ATT).⁴²⁷ However, while these and other regimes provide a precedent for dealing with- or restricting the proliferation of potential dual-use technologies, there may be shortfalls with applying them to military or security-relevant AI systems. For one, while export control regimes have certainly been updated and revised in the past, this has not often been done consistently or effectively—often requiring large outside political shocks to create momentum.⁴²⁸ More generally, these control regimes have

⁴¹⁹ Kunz and Ó hÉigearthaigh, *supra* note 216 at 5. (noting “The issuance of a circular precluding the operation of autonomous ships in international waters pending the entry into force of an international regulatory framework was considered but did not receive sufficient support”).

⁴²⁰ Natalie Klein, *Maritime Autonomous Vehicles within the International Law Framework to Enhance Maritime Security*, 95 INT. LAW STUD. 29 (2019); Robert McLaughlin, *Unmanned Naval Vehicles at Sea: USVs, UUVs, and the Adequacy of the Law*, J. LAW INF. SCI. (2011), <http://www5.austlii.edu.au/journals/JILawInfoSci/2012/6.html> (last visited Jun 22, 2020).

⁴²¹ Kunz and Ó hÉigearthaigh, *supra* note 216 at 6.

⁴²² CONVENTION ON ROAD TRAFFIC, 125 UNTS 22 (1949).

⁴²³ CONVENTION ON ROAD TRAFFIC, 1042 UNTS 15705 (1968).

⁴²⁴ See also Smith, *supra* note 299.

⁴²⁵ INTERNATIONAL CONVENTION FOR THE SUPPRESSION OF TERRORIST BOMBINGS, 2149 UNTS 256 (1997); Kunz and Ó hÉigearthaigh, *supra* note 216 at 7.

⁴²⁶ Specifically UNITED NATIONS SECURITY COUNCIL, *Resolution 1373* (2001), https://www.unodc.org/pdf/crime/terrorism/res_1373_english.pdf (last visited Sep 5, 2020) (on preventing and suppressing terrorist acts). And UNITED NATIONS SECURITY COUNCIL, *Resolution 1540* (2004) (on the non-proliferation of WMDs and their means of delivery). See Kunz and Ó hÉigearthaigh, *supra* note 216 at 7.

⁴²⁷ Kunz and Ó hÉigearthaigh, *supra* note 216 at 7–8.

⁴²⁸ Amy J Nelson, *The Impact of Emerging Technologies on Arms Control Regimes* (2018), <http://www.isodarco.it/courses/andal018/paper/iso18-AmyNelson.pdf>; AMY J. NELSON, *Innovation Acceleration, Digitization, and the Arms Control Imperative* (2019), <https://papers.ssrn.com/abstract=3382956> (last visited May 29, 2020). Although

at times struggled to resolve inter-regime conflicts.⁴²⁹ There is also a complexity around regulating or constraining the diverse components and sub-systems that go into the production of a hazardous AI system. More fundamentally, cloud computing and digital technology are not easily captured and constrained through non-proliferation efforts focused on material constraints.⁴³⁰ Indeed, export controls have arguably rarely been fully effective at restricting algorithmic innovations, in spite of a number of attempts.⁴³¹ Indeed, some suggest the increase in digitalisation has already begun to eat away at a range of export regimes or national export policies.⁴³² The pressures on repurposing export control regimes for AI are also reflected in the fact that, in summer 2020, the U.S. announced an intention to unilaterally reinterpret how it would participate in the MTCR in order to allow U.S. companies to export more weaponised drones.⁴³³ It is therefore unclear if or how fraying export control regimes might be revived or scaled up to apply to hazardous AI systems, though they offer at least one tool that should be built upon before it is rapidly discarded.

There is another shortfall in many of these above-mentioned instruments and regimes, which is that most focus predominantly on (direct harms from) cyber-physical systems such as drones or embodied robots. That suggests they might experience difficulties being extended to AI challenges in- or deriving from more virtual settings, or where they automate or support organisational decision-making. That is not to say there are no alternatives. For instance, Kunz & Ó hÉigearthaigh point to a series of existing conventions that can be applied to issues such as AI-based data processing (such as the 1981 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data; and the EU General Data Protection Regulation),⁴³⁴ or to AI-based crimes (such as the 2001 Budapest Convention on Cybercrime).⁴³⁵

However, there are limits to these regimes also. Beyond fragmented membership, many of these focus in on specific use cases or functions of AI within their jurisdiction. This can lead to the production of conflicting norms or standards, given that similar AI capabilities or techniques can be shared across such domains or use cases.

see also Caroline Fehl's more positive account of regime evolution: Caroline Fehl, *Unequal power and the institutional design of global governance: the case of arms control*, 40 REV. INT. STUD. 505–531 (2014).

⁴²⁹ KOLJA BROCKMANN, *Challenges to multilateral export controls: The Case for Inter-regime Dialogue and Coordination* 40 (2019), <https://www.sipri.org/publications/2019/other-publications/challenges-multilateral-export-controls-case-inter-regime-dialogue-and-coordination>.

⁴³⁰ MARK BROMLEY & GIOVANNA MALETTA, *The challenge of software and technology transfers to non-proliferation efforts: Implementing and complying with export controls* 50 (2018), <https://www.sipri.org/publications/2018/other-publications/challenge-software-and-technology-transfers-non-proliferation-efforts-implementing-and-complying>.

⁴³¹ One notable example here might be cryptography. Whitfield Diffie & Susan Landau, *The Export of Cryptography in the 20th Century and the 21st* (2005), https://privacyink.org/pdf/export_control.pdf (last visited Sep 2, 2019). However, at the same time, an international framework for cryptography is still lacking, see for an exploration Ashley Deeks, *The International Legal Dynamics of Encryption*, 1609 HOOVER INST. AEGIS PAP. SER. 28 (2017).

⁴³² NELSON, *supra* note 428.

⁴³³ Daryl G. Kimball, *U.S. Aims to Expand Drone Sales*, ARMS CONTROL ASSOCIATION (2020), <https://www.armscontrol.org/act/2020-07/news/us-aims-expand-drone-sales> (last visited Aug 12, 2020); PAUL K KERR, *U.S.-Proposed Missile Technology Control Regime Changes* 3 (2020), <https://fas.org/sgp/crs/nuke/IF11069.pdf>.

⁴³⁴ Kunz and Ó hÉigearthaigh, *supra* note 216 at 9–12.

⁴³⁵ *Id.* at 12–13.

2.3.1.5 International Courts and Adjudication

Finally, rather than rely on gradual accumulation of customary international law, or the explicit updating or extension of treaties, a different approach might be to rely on ‘nonconsensual’ mechanisms of international law development, including gradual development of international legal norms through adjudication in international courts.⁴³⁶

Indeed, over the last decades, many scholars have noted a trend whereby international relations and international law have become increasingly ‘judicialized’.⁴³⁷ As noted by Van Aaken, judgments delivered by international courts can develop international law by including new scientific evidence or developments, such that these decisions can “clarify norms and even ‘find’ norms, such as custom”, or at least “nudge states in certain directions since they create focal points by clarifying the law”.⁴³⁸ Moreover, while not all international treaties open themselves up to interpretation to adjudication, some do so mandatorily, or by opt-in. Moreover, legally speaking, case law can have tremendous impact and is often largely followed and cited.⁴³⁹

However, while this could support the long-term readjustment of international law to the effects of AI and digital societies more broadly, adjudication has downsides. This is because courts can only deal with the precise case that is brought before them, and have at any rate limited or specific jurisdiction. This means that any resulting legal development can be quite scattershot. Moreover, given the high speed and opacity of AI innovation, awaiting specific case law decisions on AI technology might involve unnecessarily long waits.⁴⁴⁰

2.3.2 AI governance: ongoing developments, prospects, and gaps

In response to the ambiguities or gaps in existing international law, it is unsurprising that there have, in recent years, been a growing number of initiatives seeking to clarify the application of international law norms to AI, or even pursuing new global governance arrangements specifically tailored to AI.

As noted, the majority of the early international debates in this space focused on ‘lethal autonomous weapons systems’. Yet while LAWS caught the eye of the public and of international lawyers early on, these conversations have since expanded. Broadly speaking, there have been three trends: (2.3.2.1) the growth in bottom-up public scrutiny of tech companies; (2.3.2.2) the prodigious rise in ‘AI ethics’, and (2.3.2.3) early moves towards institutionalisation and the development of international partnerships.

⁴³⁶ Aaken, *supra* note 309; Nico Krisch, *The Decay of Consent: International Law in an Age of Global Public Goods*, 108 AM. J. INT. LAW 1–40 (2014).

⁴³⁷ Aaken, *supra* note 309 at 85–86; Cesare P.R. Romano, *Proliferation of International Judicial Bodies: The Pieces of the Puzzle*, 31 N. Y. UNIV. J. INT. LAW POLIT. 709 (1998).

⁴³⁸ Aaken, *supra* note 309 at 85.

⁴³⁹ This is notwithstanding its non-binding nature for future cases. I thank Theodora Valkanou for discussing some of these points.

⁴⁴⁰ For a related argument, see also Scherer, *supra* note 298 at 388–393 (on the institutional competences of courts in the common law tort system).

2.3.2.1 Public scrutiny and activism

A first trend is found in the increase of domestic or regional public scrutiny of tech companies, in the wake of a series of public scandals. In response, civil society investigations, scholarship, public pressure or employee activism has managed to achieve remarkable shifts in industry policy in at least some cases.

Initially, this focused on collaborations in defence and security, with the emergence of the ‘Campaign to Stop Killer Robots’,⁴⁴¹ along with public campaigns and employee activism which notably led to Google halting its controversial Project MAVEN collaboration.⁴⁴² However, this has in recent years developed in a broader wave of public scrutiny of the tech sector. Amongst other topics, there has been considerable focus on the biased performance of algorithms used in crime predictions,⁴⁴³ as well as in facial and voice recognition software,⁴⁴⁴ the use of robots in public spaces, and the use of facial recognition systems and surveillance.

This has resulted in remarkable successes. Naming-and-shaming campaigns to call out the biased performance in such AI platforms has had some success at shifting industry policies.⁴⁴⁵ Moreover, US activism has also appeared effective at shifting industry perceptions of- or policies on the provision of facial recognition software to police services, notably leading both Axon (in 2019) and IBM (in 2020) to forego the provision of such tools,⁴⁴⁶ as well as Amazon (in 2020) to suspend provision of such services to police departments for one year, in the wake of activism by the Black Lives Matter movement.⁴⁴⁷

Such developments are remarkable—and might have seemed far away just years ago. Such movements have already played a large role in precipitating landmark national regulation in a variety of countries. Moreover, they may play an indispensable role at shifting global norms and perceptions, and potentially creating momentum for eventual global accords.⁴⁴⁸ Nonetheless, these movements do not by themselves create global AI governance instruments. For this, we can turn to a wide constellation of governance activities.

⁴⁴¹ Campaign to Stop Killer Robots, *About Us*, <https://www.stopkillerrobots.org/about/> (last visited Sep 5, 2020).

⁴⁴² Belfield, *supra* note 87.

⁴⁴³ Lauren Kirchner et al., *Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.*, PROPUBLICA (2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (last visited May 24, 2017).

⁴⁴⁴ Buolamwini and Gebru, *supra* note 279.

⁴⁴⁵ Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products* 7 (2019).

⁴⁴⁶ AXON AI & POLICING TECHNOLOGY ETHICS BOARD, *supra* note 88; Arvind Krishna, *IBM CEO’s Letter to Congress on Racial Justice Reform*, THINKPOLICY BLOG (2020), <https://www.ibm.com/blogs/policy/facial-recognition-susset-racial-justice-reforms/> (last visited Jun 22, 2020).

⁴⁴⁷ Amazon, *We are implementing a one-year moratorium on police use of Rekognition*, DAY ONE BLOG (2020), <https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition> (last visited Sep 5, 2020).

⁴⁴⁸ Maas, *supra* note 254 at 296–299 (discussing potential epistemic community dynamics around the movement to achieve restrictions on military AI). *Paper [II]*.

2.3.2.2 The rise of AI ethics and ‘supersoft’ law

Another trend has been the prodigious rise of ‘AI ethics’.⁴⁴⁹ Between 2016 and July 2019, at least 86 sets of AI ethics principles were promulgated.⁴⁵⁰ This figure has risen to over 160 organisational, national, and international sets of AI ethics and governance principles by 2020.⁴⁵¹ These have been articulated by a wide array of actors, including governments, intergovernmental organisations, companies, professional associations, civil society advocacy groups, and multi-stakeholder initiatives. High-profile steps involved the early 2017 Asilomar AI Principles,⁴⁵² as well as the work of the Institute of Electrical and Electronics Engineers (IEEE), which from 2016–2019 worked on the publication of a landmark set of guidelines for the ‘Ethically Aligned Design’ of autonomous and intelligent systems.⁴⁵³

The prominence of AI ethics and voluntary standards is perhaps in keeping with a broader systemic trend towards what some have called the ‘stagnation’ of traditional international law, as a result of a broad move towards increasingly ‘informal international lawmaking’,⁴⁵⁴ or at least towards the ‘multi-stakeholder’ model of governance which has, over the past two decades, also characterised cyberspace governance.⁴⁵⁵ Indeed, many of these AI governance initiatives constitute what Thomas Burri has characterised as ‘supersoft law’: codes that are ‘soft’ not only in the sense of being non-binding global standards or guidelines, but also ‘soft’ because they are not derived from international legal instruments, nor have been developed in the traditional fora of international law or state diplomacy, but instead have been formulated by (industry) stakeholders themselves.⁴⁵⁶

Are AI ethics principles effective or adequate? Burri argues in favour of continued reliance on such ‘supersoft’ law for AI, since, “emerging out of the international legal nowhere, [these standards] will be persuasive on their merits and imbued with a strong compliance pull, despite

⁴⁴⁹ It is important to distinguish between ‘AI ethics’ as a field, and as a practice. In the former sense, this refers to the rise of a dedicated ‘AI ethics’ community of scholars and activists, with associated initiatives, institutes, research groups, and conferences, which have advocated for a range of ethical values and principles around AI. In the latter sense, AI ethics refers to a *practice*, namely the formulation over the past years of diverse sets of (soft law) principles or frameworks meant to guide global AI development. These principles have been formulated and advocated for by a diverse range of actors, which certainly include actors from the ‘AI ethics’ field, but also features codes and principles by a broader range of actors such as states or companies. In this section, I use ‘AI ethics’ mostly in the latter sense, although that is not meant to downplay the important contributions of the AI ethics community. I thank Carina Prunkl for prompting this clarification.

⁴⁵⁰ For overviews, see Anna Jobin, Marcello Ienca & Effy Vayena, *The global landscape of AI ethics guidelines*, NAT. MACH. INTELL. 1–11 (2019); Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI* (2020), <https://dash.harvard.edu/handle/1/42160420> (last visited Jan 16, 2020); Yi Zeng, Enmeng Lu & Cunqing Huangfu, *Linking Artificial Intelligence Principles*, ARXIV181204814 CS (2018), <http://arxiv.org/abs/1812.04814> (last visited Jan 30, 2019).

⁴⁵¹ HIGH-LEVEL PANEL ON DIGITAL COOPERATION, *Report of the Secretary-General: Roadmap for Digital Cooperation* 39 18 (2020), https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/Roadmap_for_Digital_Cooperation_EN.pdf.

⁴⁵² Future of Life Institute, *Asilomar AI Principles*, FUTURE OF LIFE INSTITUTE (2017), <https://futureoflife.org/ai-principles/> (last visited Feb 27, 2017).

⁴⁵³ IEEE, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (2019), <https://engagestandards.ieee.org/rs/211-FYL-955/images/EAD1e.pdf> (last visited Mar 26, 2019).

⁴⁵⁴ Pauwelyn, Wessel, and Wouters, *supra* note 345.

⁴⁵⁵ NYE, *supra* note 333.

⁴⁵⁶ Burri, *supra* note 391 at 106, 109.

their non-binding form".⁴⁵⁷ At the same time, other scholars have critiqued the shortcomings of ethics codes for governing AI. These critiques depart from four distinct angles.

In the first place, some scholars argue that the principles selected are *undertheorised*. That is, they are not suitable because they are too vague, or because surface agreements hide underlying tensions amongst principles.⁴⁵⁸ It is unclear to what degree the content of these ethical principles reflects deep consensus. While Jobin and colleagues identify at least 11 distinct clusters of ethical principles amongst all sets of proposed AI principles, they also find that most find common ground on the five ethical principles of transparency, justice and fairness, non-maleficence, responsibility and privacy.⁴⁵⁹ At face value, these ethical principles may suggest a gradual degree of substantive or normative convergence.⁴⁶⁰ On the other hand, Jobin et al. also observe "substantive divergences among all [identified] 11 ethical principles in relation to four major factors: (1) how ethical principles are interpreted; (2) why they are deemed important; (3) what issue, domain or actors they pertain to; and (4) how they should be implemented."⁴⁶¹

In the second place, others argue that, even if there was greater convergence or they were better operationalised, these principles are *substantively inappropriate*. Some argue this is because many of these principles imply or import a context of industry practice which is misleading; for instance, Brent Mittelstadt has argued that the convergent sets of principles encapsulated in these AI ethics documents—transparency, justice and fairness, non-maleficence, responsibility and privacy—mirror the core principles in medical ethics.⁴⁶² However, he argues that this relies on a false analogy, since the respective contexts of practice are considerably different, because "[c]ompared to medicine, AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms."⁴⁶³ According to these critiques, then, while it may not be intrinsically incorrect to pursue AI governance (at least in part) through the articulation of ethical principles, the specific principles promulgated to date may fall short substantively.⁴⁶⁴

⁴⁵⁷ *Id.* at 109.

⁴⁵⁸ Jess Whittlestone et al., *The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions*, in PROCEEDINGS OF AAAI / ACM CONFERENCE ON ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY 2019 7 (2019), http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_188.pdf.

⁴⁵⁹ Jobin, Ienca, and Vayena, *supra* note 450 at 11.

⁴⁶⁰ Likewise, Fjeld and colleagues, in a more focused review of 36 principles, find an increasing convergence towards: "Privacy (8 principles); Accountability (10 principles); Safety and security (4 principles); Transparency and explainability (8 principles); Fairness and non-discrimination (6 principles); Human control of technology (3 principles); Professional responsibility (5 principles); Promotion of human values (3 principles)." JESSICA FJELD ET AL., *Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches* 1 15 (2019), <https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf>.

⁴⁶¹ Jobin, Ienca, and Vayena, *supra* note 450 at 7.

⁴⁶² Some make the link to bioethics explicit. Mariarosaria Taddeo & Luciano Floridi, *How AI can be a force for good*, 361 SCIENCE 751–752 (2018).

⁴⁶³ Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, 1 NAT. MACH. INTELL. 501–507 (2019).

⁴⁶⁴ Although for a defence and rebuttal, from a philosophical perspective, see Elettra Bietti, *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*, in PROCEEDINGS OF ACM FAT* CONFERENCE (2019), <https://papers.ssrn.com/abstract=3513182> (last visited Jan 6, 2020). More generally, the fact that AI development lacks certain contexts of practice (such as fiduciary duties) does not mean these could not be developed. See for instance Anthony Aguirre et al., *AI loyalty: A New Paradigm for Aligning Stakeholder Interests*, 20.18 UNIV. COLO. LAW LEG. STUD. RES. PAP. (2020), <https://papers.ssrn.com/abstract=3560653> (last visited Jun

In the third place, some have critiqued the principles on the grounds of *process and inclusion*. These scholars argue that even if the principles espoused were appropriate and effective on their merits, they have not derived from an inclusive process and therefore may not be perceived as legitimate, or at least do not (and likely will not) receive a sufficiently broad global buy-in to be viable. Some express concern that (largely) industry-based codes of ethics will “upstage democratically enacted law”, given the political clout wielded by technology companies.⁴⁶⁵ Others are concerned that many of these principles reflect the perspective of the West rather than a global consensus.⁴⁶⁶ As such, some are concerned that these ethics codes are (or will be perceived) as too parochial. As Donahoe & Metzger have argued, “[w]hile each effort has hold of part of the problem, none grasps the full spectrum of risks to humans or the whole range of governance challenges [and] none of these initiatives can claim to have achieved a broad global consensus or buy-in across stakeholder groups.”⁴⁶⁷

Fourthly, some concerns are expressed about the *categorical appropriateness of relying on ethics principles* in the first place. Some have critiqued declarations of ethical principles for attempting to exhaustively list impacts in a way that must inevitably fall short or omit values.⁴⁶⁸ More concretely, others have argued that ethics guidelines are by nature not easily operationalisable, or that they lack enforceability, meaning that deviations often carry little consequence, or might even simply serve as a marketing ploy by actors.⁴⁶⁹ A frequently-expressed concern is that the ‘ethical principles’ approach of AI ethics is prone to manipulation and risks becoming a substitute for actual law and regulation.⁴⁷⁰ Others highlights the limits to private self-governance, given the incompatible incentives of firms in many contexts of AI deployment.⁴⁷¹ Finally, on a practical level, it is unclear if, or under what conditions, codes of ethics actually work to change the choices of developers or actors. For instance, one survey that primed computer

13, 2020); Danny Tobey, *Software Malpractice in the Age of AI: A Guide for the Wary Tech Company*, in AAAI / ACM CONFERENCE ON ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY 2018 6 (2018), http://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_43.pdf. I thank Charlotte Siegmann for highlighting some of these debates.

⁴⁶⁵ Tasioulas, *supra* note 408 at 66; See also generally Nemitz, *supra* note 303. Benvenisti, *supra* note 303.

⁴⁶⁶ Daniel Schiff et al., *What’s Next for AI Ethics, Policy, and Governance? A Global Overview* (2020), <https://econpapers.repec.org/paper/osfsocarx/8jaz4.htm> (last visited Jan 12, 2020). See also ANGELA DALY ET AL., *Artificial Intelligence Governance and Ethics: Global Perspectives* (2019), <https://arxiv.org/ftp/arxiv/papers/1907/1907.03848.pdf> (last visited Jun 28, 2019); Smuha, *supra* note 405. However, for a more optimistic account, arguing that cross-cultural collaboration may be possible on many aspects of AI governance, see Seán S. Ó hÉigearthaigh et al., *Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance*, PHILOS. TECHNOL. (2020), <https://doi.org/10.1007/s13347-020-00402-x> (last visited May 17, 2020).

⁴⁶⁷ Donahoe and Metzger, *supra* note 405 at 118.

⁴⁶⁸ John Tasioulas has critiqued declarations of ethical principles because they try to enumerate impacts in a manner that attempts to be exhaustive, reflecting an “implicit assumption that there is an enumerable catalogue of evaluative considerations that are especially engaged by RAIs.” Tasioulas, *supra* note 408 at 69–70. Of course, it might be argued that this is a fully general critique of any codes or lists of ethics.

⁴⁶⁹ Thilo Hagerdorff, *The Ethics of AI Ethics -- An Evaluation of Guidelines*, ARXIV190303425 CS STAT, 10 (2019), <http://arxiv.org/abs/1903.03425> (last visited Oct 7, 2019). However, it should be noted that in the biomedical context, the field has been quite successful in operationalizing sets of high-level principles, such as those in; TOM L. BEAUCHAMP & JAMES F. CHILDRESS, *PRINCIPLES OF BIOMEDICAL ETHICS* (2012). I thank Carina Prunkl for this point.

⁴⁷⁰ See Anaïs Rességuier & Rowena Rodrigues, *AI ethics should not remain toothless! A call to bring back the teeth of ethics*, 7 BIG DATA SOC. 2053951720942541 (2020).

⁴⁷¹ Nemitz, *supra* note 303; Slee, *supra* note 303.

programmers with a widespread code of computing ethics suggested that this did not materially change these programmer's behaviour or choices when afterwards they were confronted with ethical dilemmas.⁴⁷² This is in line with previous studies suggesting that ethical guidelines may have little to no effect at altering the decision-making of professionals in various fields.⁴⁷³

Ultimately, the increase in these AI ethics principles clearly reflects growing global concern. Yet while there are many advantages to such soft-law approaches,⁴⁷⁴ ethical principles may not cover all domain areas where AI introduces new problems. For instance, with a few exceptions,⁴⁷⁵ there are very few areas where standards cover to AI systems used in security contexts. Nonetheless, one major benefit to the AI ethics principles is that they have begun the process of exploring and articulating a 'normative core' for AI governance. Nonetheless, much remains to be done to translate this into practice—which is one place where intergovernmental initiatives could come in.⁴⁷⁶

2.3.2.3 Early steps towards institutionalisation

A third trend is found in early, tentative steps towards global institutionalisation. As the prominence of AI technology has grown, there has been increasing action to bring certain AI issues under the umbrella of international organisations or intergovernmental fora. This includes the Geneva Process on Lethal Autonomous Weapons Systems under the Convention on Certain Conventional Weapons; programs in the broader UN system, as well as a range of broader multilateral developments under the OECD, Council of Europe, European Commission, G20, and the Global Partnership on AI.⁴⁷⁷

2.3.2.3.1 Lethal Autonomous Weapons under the CCW

While many initiatives on AI governance are relatively new, there is, as noted, one area in which international lawyers have engaged for a longer time, and that is the governance for

⁴⁷² Andrew McNamara, Justin Smith & Emerson Murphy-Hill, *Does ACM's code of ethics change ethical decision making in software development?*, in PROCEEDINGS OF THE 2018 26TH ACM JOINT MEETING ON EUROPEAN SOFTWARE ENGINEERING CONFERENCE AND SYMPOSIUM ON THE FOUNDATIONS OF SOFTWARE ENGINEERING - ESEC/FSE 2018 729–733, 4 (2018), <http://dl.acm.org/citation.cfm?doid=3236024.3264833> (last visited Apr 8, 2019). ("Despite its stated goal, we found no evidence that the ACM code of ethics influences ethical decision making.")

⁴⁷³ Arthur P. Brief et al., *What's wrong with the treadway commission report? Experimental analyses of the effects of personal values and codes of conduct on fraudulent financial reporting*, 15 J. BUS. ETHICS 183–198 (1996); Margaret Anne Cleek & Sherry Lynn Leonard, *Can corporate codes of ethics influence behavior?*, 17 J. BUS. ETHICS 619–630 (1998); M. Osborn et al., *Do ethical Guidelines make a difference to decision-making?*, 39 INTERN. MED. J. 800–805 (2009); As discussed in Nicolas Kluge Corrêa, *Blind Spots in AI Ethics and Biases in AI governance*, 17 (2020), https://www.researchgate.net/publication/342821723_Blind_Spots_in_AI_Ethics_and_Biases_in_AI_governance?channel=doi&linkId=5f47ea4fa6fdcc14c5d0da27&showFulltext=true (last visited Aug 28, 2020).

⁴⁷⁴ Gary Marchant, "Soft Law" Governance Of Artificial Intelligence, AI PULSE (2019), <https://aipulse.org/soft-law-governance-of-artificial-intelligence/> (last visited Feb 26, 2019).

⁴⁷⁵ See MARKUS CHRISTEN ET AL., *An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications* (2017), <https://papers.ssrn.com/abstract=3063617> (last visited Apr 30, 2019).

⁴⁷⁶ Eugenio V Garcia, *Multilateralism and Artificial Intelligence: What Role for the United Nations?*, in THE GLOBAL POLITICS OF ARTIFICIAL INTELLIGENCE 18 (Maurizio Tinnirello ed., 2020).

⁴⁷⁷ *Id.*; See also the detailed recent overview in PHILIPPE LORENZ, *AI Governance through Political Fora and Standards Developing Organizations* 41 (2020), <https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations>.

Lethal Autonomous Weapons Systems. It is useful to provide a brief background on the origin and history of this governance effort.⁴⁷⁸

In fact, scholarship and academic debate on potential military uses of robots and artificial intelligence has been ongoing for several decades.⁴⁷⁹ However, while the issues around robotic weapons were raised amongst civil society as early as the mid-'00s, and sparked some early calls for 'robot arms control',⁴⁸⁰ the issue did not initially receive sufficient uptake.⁴⁸¹ This changed between 2009-2013, when the international discussion of autonomous weapons found its feet. The International Committee for Robot Arms Control (ICRAC) was established in 2009, and began to organize expert debate into more concerted outreach.⁴⁸² It was particularly in 2012 that the issue began to gain political traction,⁴⁸³ after Human Rights Watch, a key gatekeeper in the humanitarian disarmament field, adopted the issue in an influential public report.⁴⁸⁴ Matters sped up further in early 2013, when Christof Heyns, the UN Special Rapporteur on extrajudicial, summary or arbitrary executions, called for a moratorium on robotic and AI weapons.⁴⁸⁵ April 2013 also saw the official launch of the Campaign Against Killer Robots, a global coalition of international, regional and national NGOs which called for a ban against such weapons.⁴⁸⁶

These efforts by civil society were not without fruit. On the public side, the period since 2014 has seen increasing pressure from civil society, as well as a series of open letters from AI researchers and tech company employees, denouncing the use of AI technology in weapons.⁴⁸⁷ Institutionally, the efforts led the States parties to the Convention on Certain Conventional Weapons (CCW) to begin a series of informal expert meetings in 2014. At the fifth Review Conference in 2016, this led to the establishment of the Group of Governmental Experts (GGE)

⁴⁷⁸ Other reviews are offered by Şerif Onur Bahçecik, *Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames*, 0 GLOB. POLICY (2019), <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12671> (last visited Jun 17, 2019); Elvira Rosert & Frank Sauer, *How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies*, 0 CONTEMP. SECUR. POLICY 1–26, 15 (2020); Kunz and Ó hÉigearthaigh, *supra* note 216 at 8.

⁴⁷⁹ See P. W. SINGER, WIRED FOR WAR: THE ROBOTICS REVOLUTION AND CONFLICT IN THE 21ST CENTURY (2009); RONALD ARKIN, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS (1 edition ed. 2009).

⁴⁸⁰ Robert Sparrow, *Predators or plowshares? arms control of robotic weapons*, 28 IEEE TECHNOL. SOC. MAG. 25–29 (2009); ARMIN KRISHNAN, KILLER ROBOTS: LEGALITY AND ETHICALITY OF AUTONOMOUS WEAPONS (1 edition ed. 2009); Jürgen Altmann, *Preventive Arms Control for Uninhabited Military Vehicles*, in ETHICS AND ROBOTICS 69–82 (R. Capurro & M. Nagenborg eds., 2009); Wendell Wallach & Colin Allen, *Framing robot arms control*, 15 ETHICS INF. DORDR. 125–135 (2013).

⁴⁸¹ CHARLI CARPENTER, "LOST" CAUSES, AGENDA VETTING IN GLOBAL ISSUE NETWORKS AND THE SHAPING OF HUMAN SECURITY (2014), <http://www.degruyter.com/viewbooktoc/product/487489> (last visited Apr 1, 2019).

⁴⁸² Rosert and Sauer, *supra* note 478 at 15.

⁴⁸³ Rosert and Sauer, *supra* note 478.

⁴⁸⁴ HUMAN RIGHTS WATCH, LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS (2012), https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf.

⁴⁸⁵ Christof Heyns, *Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions* (2013), https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf (last visited Feb 27, 2019).

⁴⁸⁶ Campaign to Stop Killer Robots, *supra* note 441. Note, as of September 2020, the coalition involves 165 NGOs in 65 countries.

⁴⁸⁷ See for instance Belfield, *supra* note 87. Bahçecik, *supra* note 478; Future of Life Institute, *Open Letter on Autonomous Weapons*, FUTURE OF LIFE INSTITUTE (2015), <https://futureoflife.org/open-letter-autonomous-weapons/> (last visited Nov 22, 2018); Letter to Google C.E.O., (2018), <https://static01.nyt.com/files/2018/technology/googleletter.pdf> (last visited Apr 9, 2018).

on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.⁴⁸⁸ The goal of this group has been to explore and reach agreement on recommended regulatory options for the Convention, in light of the compliance of LAWS with International Humanitarian Law, International Human Rights Law, as well as broader global security risks.⁴⁸⁹

Since 2016, the GGE has met annually, however there has been no definitive resolution to date, and the process's prospects for success remain unclear. Much of its approach has focused on the concept of 'Meaningful Human Control' (MHC), a key solution that however remains somewhat underdefined. At its last meeting in November 2019, the group agreed on 11 non-binding guiding principles to provide a framework for the development and use of LAWS.⁴⁹⁰

However, there has been some growing pessimism about the ability of this process to yield a binding ban. Civil society campaigns, frustrated with limited progress and continued state opposition in 2019 and even during the latest GGE in late September 2020, have as such increasingly threatened to shift negotiations to another forum, as they in the past did for anti-personnel mines.⁴⁹¹

2.3.2.3.2 AI in the broader UN system

Even as there remains uncertainty about the long-term prospects or adequacy of the CCW GGE process for LAWS, this has not stopped broader institutional examination of various other AI issues. Over the last years there have been a wide range of AI-focused activities across the broader United Nations system.⁴⁹² In September 2018, UN Secretary General António Guterres issued a *Strategy on New Technologies*, which amongst others emphasised the positive use of AI and digital technologies in the service of key global goals.⁴⁹³ Accordingly, the UN Secretary-

⁴⁸⁸ UNITED NATIONS CONVENTION ON PROHIBITION OR RESTRICTIONS ON THE USE OF CERTAIN CONVENTIONAL WEAPONS WHICH MAY BE DEEMED TO BE EXCESSIVELY INJURIOUS OR TO HAVE INDISCRIMINATE EFFECTS, 1342 UNTS 137 (1980).

⁴⁸⁹ Kunz and Ó hÉigearthaigh, *supra* note 216 at 8. See also Decision 1 of the Fifth Review Conference of the High Contracting Parties to the CCW, see 'Final Document of the Fifth Review Conference' (23 December 2016) UN Doc CCW/CONF.V/10, 9; in conjunction with the recommendations contained in 'Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)' (10 June 2016) UN Doc CCW/CONF.V/2, Annex.

⁴⁹⁰ UNOG, *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems* 13 (2019), [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/5497DF9B01E5D9CFC125845E00308E44/\\$file/CCW_GGE.1_2019_CRP.1_Rev2.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/5497DF9B01E5D9CFC125845E00308E44/$file/CCW_GGE.1_2019_CRP.1_Rev2.pdf) (last visited Oct 7, 2019). Annex IV. These principles affirm, inter alia, that IHL applies to AWS (a), a human must always be responsible for the decision to use these systems (b, d), and states must examine the legality of these new weapons at the design stage (e, f, g). See also France Diplomatic Ministry for Europe and Foreign Affairs, *11 Principles on Lethal Autonomous Weapons Systems (LAWS)*, FRANCE DIPLOMACY - MINISTRY FOR EUROPE AND FOREIGN AFFAIRS (2019), <https://www.diplomatie.gouv.fr/en/french-foreign-policy/united-nations/multilateralism-a-principle-of-action-for-france/alliance-for-multilateralism-63158/article/11-principles-on-lethal-autonomous-weapons-systems-laws> (last visited Sep 21, 2020).

⁴⁹¹ Janosch Delcker, *How killer robots overran the UN*, POLITICO, 2019, <https://www.politico.eu/article/killer-robots-overran-united-nations-lethal-autonomous-weapons-systems/> (last visited Oct 2, 2019). Campaign to Stop Killer Robots, *Diplomatic Talks in 2020* (2020), <https://www.stopkillerrobots.org/2020/09/diplomatic2020/> (last visited Sep 26, 2020).

⁴⁹² This discussion draws on, amongst others, Garcia, *supra* note 476; James Butcher & Irakli Beridze, *What is the State of Artificial Intelligence Governance Globally?*, 164 RUSI J. 88–96 (2019). LORENZ, *supra* note 477.

⁴⁹³ ANTÓNIO GUTERRES, *UN Secretary-General's Strategy on New Technologies* (2018), <http://www.un.org/en/newtechnologies/images/pdf/SGs-Strategy-on-New-Technologies.pdf> (last visited Oct 4, 2018); See also the resulting report: UN SG HLPDC, *The Age of Digital Interdependence: Report of the UN Secretary-General's High-*

General's High-level Panel on Digital Cooperation ran a public consultation,⁴⁹⁴ following onto which it published its first report, "The Age of Digital Interdependence", in June 2019, articulating a series of recommendations to improve digital cooperation.⁴⁹⁵ In that same month, the UN Chief Executives Board for Coordination (CEB) adopted a set of guidelines articulating a strategic approach for supporting capacity development on AI throughout the UN system.⁴⁹⁶

Simultaneously, over the past years, a wide variety of UN organisations have begun at least informal activities exploring the implications of AI technology for their operations.⁴⁹⁷ These include, amongst others, initial investigations and reports by the International Labour Organization (ILO) and the United Nations Educational, Scientific and Cultural Organization (UNESCO),⁴⁹⁸ activities by the International Telecommunications Union (ITU), International Maritime Organization (IMO) and International Civil Aviation Organization (ICAO);⁴⁹⁹ and a range of initiatives by the UN Institute for Disarmament Research (UNIDIR), the Centre for Policy Research at the United Nations University (UNU); as well as a Centre for Artificial Intelligence and Robotics by the UN Interregional Crime and Justice Research Institute (UNICRI).⁵⁰⁰ Finally, UNESCO has established an Ad Hoc Expert Group of 24 members in order to publish the first draft of a global standard-setting instrument on the ethics of AI, to be approved by UNESCO's General Conference by the end of 2021.⁵⁰¹ While it is unclear whether this normative document will take the form of a declaration, recommendation, or convention, the group produced a 'zero draft' in May 2020, which laid out a number of the substantive principles as well as proposed areas of policy action.⁵⁰²

2.3.2.3.3 Other multilateral developments: ISO, OECD Principles, G20, and GPAI

Prominently, there has also been growing activity outside the UN, potentially speaking to a broader degree of convergence in the AI governance sphere.

Level Panel on Digital Cooperation (2019), <https://digitalcooperation.org/wp-content/uploads/2019/06/DigitalCooperation-report-for-web.pdf> (last visited Jun 13, 2019).

⁴⁹⁴ Which included a submission by this author; see Luke Kemp et al., *UN High-level Panel on Digital Cooperation: A Proposal for International AI Governance* (2019), https://digitalcooperation.org/wp-content/uploads/2019/02/Luke_Kemp_Submission-to-the-UN-High-Level-Panel-on-Digital-Cooperation-2019-Kemp-et-al.pdf.

⁴⁹⁵ UN SG HLPDC, *supra* note 493.

⁴⁹⁶ Chief Executives Board for Coordination, *Summary of deliberations: A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence* (2019), <http://digitallibrary.un.org/record/3811676> (last visited Sep 12, 2020); Garcia, *supra* note 476 at 10.

⁴⁹⁷ ITU, *United Nations Activities on Artificial Intelligence (AI)* 66 (2018), https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2018-1-PDF-E.pdf; ITU, *United Nations Activities on Artificial Intelligence (AI) 2019* 88 (2019), https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf.

⁴⁹⁸ BRIGITTE LASRY ET AL., *HUMAN DECISIONS: THOUGHTS ON AI* (2018), <http://unesdoc.unesco.org/images/0026/002615/261563e.pdf> (last visited Nov 6, 2018).

⁴⁹⁹ Garcia, *supra* note 476 at 16.

⁵⁰⁰ Butcher and Beridze, *supra* note 492; Garcia, *supra* note 476 at 11.

⁵⁰¹ Garcia, *supra* note 476 at 10.

⁵⁰² AD HOC EXPERT GROUP (AHEG) FOR THE PREPARATION OF A DRAFT TEXT OF A RECOMMENDATION THE ETHICS OF ARTIFICIAL INTELLIGENCE, *Outcome document: first version of a draft text of a recommendation on the Ethics of Artificial Intelligence* (2020), <https://unesdoc.unesco.org/ark:/48223/pf0000373434> (last visited Jun 25, 2020).

For its part, the International Organization for Standardization (ISO) has for several years been engaged in articulating a series of standards for different autonomous systems.⁵⁰³ While these are in principle non-binding voluntary standards, the ISO's membership consists of 162 national standards bodies who adopt ISO standards domestically. Moreover, international standards as articulated by the ISO and others can indirectly have considerable pull, because under the Agreement on Technical Barriers to Trade (TBT), voluntary international standards may be converted into binding treaty obligations for the 164 members of the WTO.⁵⁰⁴ On the other hand, as noted by Peter Cihon, international standards approaches for AI at the ISO and IEEE have focused "primarily on standards to improve market efficiency, and to address ethical concerns, respectively", leaving "a risk that these standards may fail to address further policy objectives, such as a culture of responsible deployment and use of safety specifications in fundamental research."⁵⁰⁵

More broadly, many recent institutional developments have revolved around a set of initiatives at the OECD, the G20, and the 'Global Partnership on AI', which, while they began independently, have begun to show increasing interconnection.

After examining AI-related issues since 2016-17, the OECD in May 2019 issued a set of principles on AI, which were adopted by its members in the Recommendation of the Council on Artificial Intelligence.⁵⁰⁶ These principles, the first intergovernmental standard of its kind, emphasised; 'inclusive growth, sustainable development and well-being', 'human-centred values and fairness', 'transparency and explainability', 'robustness, security and safety', and 'accountability'.⁵⁰⁷ In addition, the document highlighted 'national policies and international cooperation for trustworthy AI', which included: "i) investing in AI research and development; ii) fostering a digital ecosystem for AI; iii) shaping an enabling policy environment for AI; iv) building human capacity and preparing for labour market transformation; and v) international cooperation for trustworthy AI."⁵⁰⁸ Moreover, in February 2020, the OECD launched the 'OECD AI Policy Observatory' in order to help shape and share AI policies.⁵⁰⁹

A second line of institutional activity has been found at the *G20*. In June 2019, one month after the publication of the OECD AI Principles, the G20 adopted a set of 'human-centred AI Principles' at the Osaka Summit Track on the Digital Economy, which drew closely from- and

⁵⁰³ See also the overview in PETER CIHON, *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development* (2019), https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_FHI-Technical-Report.pdf (last visited Apr 18, 2019).

⁵⁰⁴ WTO, *Agreement on Technical Barriers to Trade*, 1868 U.N.T.S. 120 (1994), https://www.wto.org/english/docs_e/legal_e/17-tbt_e.htm. Article 2.4. See also Han-Wei Liu & Ching-Fu Lin, *Artificial Intelligence and Global Trade Governance: A Pluralist Agenda*, 61 HARV. INT. LAW J., 24 (2020), <https://papers.ssrn.com/abstract=3675505> (last visited Sep 26, 2020).

⁵⁰⁵ CIHON, *supra* note 503 at 2.

⁵⁰⁶ OECD, *Recommendation of the Council on Artificial Intelligence* (2019), <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> (last visited May 28, 2019).

⁵⁰⁷ *Id.* at 7–8.

⁵⁰⁸ *Id.* at 8–10.

⁵⁰⁹ OECD, *OECD AI Policy Observatory: A platform for AI information, evidence, and policy options* (2019), <https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policy-observatory.pdf> (last visited Oct 8, 2019); OECD, *The OECD Artificial Intelligence Policy Observatory*, <https://www.oecd.ai/> (last visited Sep 17, 2020); See also Garcia, *supra* note 476 at 8.

endorsed the OECD principles.⁵¹⁰ They also ‘took note of’—though did not explicitly extend support for—the OECD’s recommendation on ‘national policies and international cooperation’.⁵¹¹ In 2020, the G20 Digital Economy Task Force (DETF) has adopted advancing ‘trustworthy AI’ as a point on the agenda of the Riyadh Summit.⁵¹² There have also been recent informal proposals to the G20 for the establishment of a Coordinating Committee for the Governance of Artificial Intelligence (CCGAI).⁵¹³

Moreover, other AI governance activities have come from informal coalitions of like-minded parties. This can be most visibly seen in the *Global Partnership on Artificial Intelligence* (GPAI). In 2018, France and Canada first proposed the creation of an ‘International Panel on Artificial Intelligence’,⁵¹⁴ explicitly inspired by the Intergovernmental Panel on Climate Change (IPCC). Initially, this initiative was opposed by the US, on the ground that it might be unnecessarily restrictive and could slow US companies.⁵¹⁵ Accordingly, it opposed the initiative at the G7 in the summer of 2019. Nonetheless, the idea of such a body lived on in the GPAI,⁵¹⁶ which the US finally joined in May 2020, explicitly describing it as a step to balance the growing role of China in shaping the global trajectory of AI.⁵¹⁷ As such, on June 15th 2020, GPAI was established by 15 founding members, as an “international and multistakeholder platform to guide the international community on the responsible development of AI.”⁵¹⁸

Beyond these initiatives, there are also more regional activities. Europe in particular has moved relatively fast. The European Commission in June 2018 created the ‘High-Level Expert Group on Artificial Intelligence’ (HLEG), to support the development of policies and eventual regulations around AI. In 2019, this resulted in the publication of a set of ‘Ethics Guidelines for Trustworthy AI’.⁵¹⁹ On the basis of the group’s work, the European Commission has been

⁵¹⁰ G20, *G20 Ministerial Statement on Trade and Digital Economy* (2019), <https://www.mofa.go.jp/files/000486596.pdf> (last visited May 25, 2020). Moreover, OECD Secretary-General Angel Gurria in turn endorsed the G20 principles. OECD, *OECD Supporting G20 Policy Priorities at Osaka Summit* (2019), <https://www.oecd.org/g20/summits/osaka/publicationsanddocuments/oecd-supporting-g20-policy-priorities-at-osaka-summit.htm> (last visited Sep 5, 2020).

⁵¹¹ Garcia, *supra* note 476 at 31.

⁵¹² G20, *Ministerial Declaration: G20 Digital Economy Ministers Meeting, 2020 Riyadh Summit* (2020), <http://www.g20.utoronto.ca/2020/2020-g20-digital-0722.html> (last visited Sep 17, 2020).

⁵¹³ See THORSTEN JELINEK, WENDELL WALLACH & DANIL KERIMI, *Coordinating Committee for the Governance of Artificial Intelligence* (2020), https://www.g20-insights.org/policy_briefs/coordinating-committee-for-the-governance-of-artificial-intelligence/ (last visited Jul 8, 2020).

⁵¹⁴ Justin Trudeau, *Mandate for the International Panel on Artificial Intelligence*, PRIME MINISTER OF CANADA (2018), <https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence> (last visited Jul 6, 2019); See also Nature Editors, *International AI ethics panel must be independent*, 572 NATURE 415–415 (2019).

⁵¹⁵ Cf. Tom Simonite, *The World Has a Plan to Rein in AI—but the US Doesn’t Like It*, WIRED, 2020, <https://www.wired.com/story/world-plan-rein-ai-us-doesnt-like/> (last visited May 31, 2020).

⁵¹⁶ Global Partnership on Artificial Intelligence, *Joint Statement from founding members of the Global Partnership on Artificial Intelligence* (2020), <https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding>.

⁵¹⁷ Michael Kratsios, *Artificial Intelligence Can Serve Democracy*, WALL STREET JOURNAL, May 27, 2020, <https://www.wsj.com/articles/artificial-intelligence-can-serve-democracy-11590618319> (last visited Jun 10, 2020).

⁵¹⁸ Global Partnership on Artificial Intelligence, *supra* note 516. The 15 founding members of GPAI are Australia, Canada, France, Germany, India, Italy, Japan, Mexico, New Zealand, Republic of Korea, Singapore, Slovenia, United Kingdom, United States, and the European Union.

⁵¹⁹ EUROPEAN COMMISSION, *Ethics Guidelines for Trustworthy AI* (2019), <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

preparing to introduce AI legislation for various applications. The aim is that this new regulatory framework would have considerable global flow-through effects, as was previously observed with the General Data Protection Regulation (GDPR) adopted in 2016.⁵²⁰

Finally, the Council of Europe (CoE) in 2019 established an Ad Hoc Committee on Artificial Intelligence (CAHAI), which has since been discussing potential legal frameworks for AI—including, potentially, a binding treaty on AI that would apply to all of its 47 member states.⁵²¹ In summer 2020, the CAHAI held a meeting with around 150 experts from all member countries along with various organisations, and its Policy Development working group is currently undertaking a feasibility study regarding such a treaty on AI applications, and is due to provide a recommendation by the end of the year.⁵²² If carried through, such a treaty could follow the precedent of other treaties enacted by the CoE, such as the 2001 Budapest Convention on Cybercrime,⁵²³ or the 2011 Istanbul Convention on preventing and combating violence against women,⁵²⁴ both of which were able to secure some—albeit not universal—participation of countries beyond the CoE member states.⁵²⁵

It should therefore be clear that there is currently a wave of attention and momentum for AI governance. However, while encouraging, it remains unclear whether these initiatives will be up to the challenges, and remain fit to govern AI in the face of change.

2.3.2.4 *Strengths and limits of AI governance approaches*

While the early steps from the past few years are promising, there may also be limits to many of these initiatives' prospects for governing AI. First, many of the UN organisations that have begun to evaluate AI issues were not originally created to address digital information technology, let alone AI.⁵²⁶ While international organisations can certainly change or expand their mission and activities over time, many have also proven 'sticky' and path-dependent,⁵²⁷ and this does not bode well for their ability to fully adapt to AI.

⁵²⁰ See Garcia, *supra* note 476 at 8.

⁵²¹ Natalie Leal, *Council of Europe starts work on legally-binding AI treaty*, GLOBAL GOVERNMENT FORUM (2020), <https://www.globalgovernmentforum.com/council-of-europe-starts-work-on-legally-binding-ai-treaty/> (last visited Sep 5, 2020). While the Council of Europe cannot itself produce binding laws, it can enforce specific international agreements reached by European states on various topics.

⁵²² Janosch Delcker, *The AI treaty you've never heard of — A tech industry reckoning — Sabre-rattling*, POLITICO AI: DECODED (2020), <https://www.politico.eu/newsletter/ai-decoded/politico-ai-decoded-the-ai-treaty-youve-never-heard-of-a-tech-industry-reckoning-sabre-rattling/> (last visited Sep 5, 2020). See also LORENZ, *supra* note 477 at 11–13.

⁵²³ CONVENTION ON CYBERCRIME, ETS No. 185 (2001), <https://www.coe.int/en/web/conventions/full-list> (last visited Sep 17, 2020).

⁵²⁴ COUNCIL OF EUROPE CONVENTION ON PREVENTING AND COMBATING VIOLENCE AGAINST WOMEN AND DOMESTIC VIOLENCE, CETS No. 210 (2001).

⁵²⁵ At present, 64 states have ratified the Budapest Cybercrime Convention; however, since it entered into force, several major states (including India and Brazil) have declined to adopt the Convention, on the argument that they did not participate in its drafting.

⁵²⁶ Indeed, as Morin and others note, the ITU was first created in 1865, and "would certainly look very different if it had been established in the fast-changing internet age". Morin et al., *supra* note 328 at 2.

⁵²⁷ Kathleen Thelen, *Historical Institutionalism in Comparative Politics*, 2 ANNU. REV. POLIT. SCI. 369–404 (1999); See for instance the case study of the ILO, in Lucio Baccaro & Valentina Mele, *Pathology of Path Dependency? The ILO and the Challenge of New Governance*, 65 ILR REV. 195–224 (2012).

Moreover, the institutional ecology on AI remains relatively unorganised. These initiatives have somewhat fragmented membership, which also does not include many smaller states.⁵²⁸ As a result, a June 2020 report by the UN Secretary General's High-Level Panel on Digital Cooperation noted that "[c]urrent artificial intelligence-related initiatives lack overall coordination in a way that is easily accessible to other countries outside the existing groupings, other United Nations entities and other stakeholders."⁵²⁹

Furthermore, the general institutional proliferation over the past few decades increases the opportunities for further legal fragmentation and potential inconsistencies between regimes. The fragmentation of AI governance may contribute to this process, because different institutions may resolve local AI challenges in different or contradictory manners.

That is not to say there are no positive signs. For instance, with the Secretariat of the GPAI housed under the OECD,⁵³⁰ this may signal that different global initiatives may approach the norm-setting for AI from different angles,⁵³¹ but that simultaneously there is a growing degree of convergence. Nonetheless, there are grounds for concern over the adversarial nature of some of the recent developments in the digital technology governance landscape. Given previous and ongoing contestation over cyberspace regulation, as well as the growing tensions over the 5G network standard,⁵³² the fact that the US joined the GPAI explicitly to counterbalance China may signal that these AI institutions and governance efforts may become sites for increasing contestation.⁵³³ Moreover, while GPAI appears meant as a forum for discussion and coordination on technical and policy research, there are at this stage no clear mission statement on the goals the initiative seeks to achieve, plans of action, or clear articulation of how this initiative will interact with other pre-existing efforts.⁵³⁴

2.3.2.5 Proposed pathways

The field may well have entered a sensitive window of opportunity to get it right. As such, given the incipient state of global AI governance, it is unsurprising that there are many proposals for new solutions. One key question that underpins many of these debates, is over to what extend policies need to be actively coordinated. For instance, a number of scholars have argued that individual (national and institutional) efforts on AI policy should be coordinated and supported by a centralised international regulatory framework. Whereas some envision this to take the shape of 'agile governance' that is coordinated through a multi-stakeholder 'Global Governance

⁵²⁸ The question of fragmentation of membership is discussed and illustrated in greater detail in Chapter 7.2.

⁵²⁹ United Nations General Assembly, *Road map for digital cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation* 13 (2020), <https://undocs.org/pdf?symbol=en/A/74/821> (last visited Sep 1, 2020).

⁵³⁰ OECD, *OECD to host Secretariat of new Global Partnership on Artificial Intelligence* (2020), <https://www.oecd.org/going-digital/ai/OECD-to-host-Secretariat-of-new-Global-Partnership-on-Artificial-Intelligence.htm> (last visited Jul 3, 2020).

⁵³¹ HIGH-LEVEL PANEL ON DIGITAL COOPERATION, *supra* note 451 at 42.

⁵³² Beaumier et al., *supra* note 335.

⁵³³ JELINEK, WALLACH, AND KERIMI, *supra* note 513.

⁵³⁴ Arindrajit Basu & Justin Sherman, *Two New Democratic Coalitions on 5G and AI Technologies*, LAWFARE (2020), <https://www.lawfareblog.com/two-new-democratic-coalitions-5g-and-ai-technologies> (last visited Aug 27, 2020).

Network for AI,⁵³⁵ others have put forward proposals for a more traditional centralised agency for AI, such as an ‘International Artificial Intelligence Organisation’ (IAIO) which can “serve as an international forum for discussion and engage in international standard setting activities.”⁵³⁶

Likewise, along with others scholars, I have previously articulated a proposal for centralised ‘international AI governance’.⁵³⁷ We argued this could take various forms, including “a UN specialised agency (such as the World Health Organisation), a Related Organisation to the UN (such as the World Trade Organisation) or a subsidiary body to the UN General Assembly (such as the UN Environment Programme).”⁵³⁸ In our analysis, there would be distinct functions such a centralised body could be aimed at, including coordinating national AI law and catalysing multilateral treaties; measuring and forecasting progress and impact of AI systems, or pooling AI research for social good.⁵³⁹ Designed well, such an organisation could fulfil multiple criteria and support valuable multilateral goals.⁵⁴⁰ Nonetheless, it remains unclear whether such a centralised institutional arrangement is plausible, necessary, or even preferable.

In sum, while recent years have seen important and encouraging developments, the current global governance landscape for AI remains uncertain. On the one hand, in many domains, and for questions of coordination and standard-setting, there are modest steps towards convergence and coordination, as illustrated by the OECD and the G20 AI principles, and the recent efforts of the Global Partnership on AI. However, even here, success remains uncertain: as observed in a recent G20 briefing, in spite of these various regional initiatives, at a global level there continues to be a move towards ‘digital sovereignty’, which, coupled with the strategic and competitive nature of cyber space and much of AI-driven digitalisation, risks “a dysfunctional international regime complex that will weaken local and regional approaches and render them ineffective.”⁵⁴¹

Moreover, in security and military domains—paradoxically the very areas where an active legal debate was first sparked and has been ongoing for many years—governance prospects may appear strained. The GGE process under the CCW appears strained or gridlocked. Furthermore, the envisioned ban on ‘killer robots’, even if it were achieved, might even prove brittle before very long. It deals primarily with a subset of the security issues posed by military AI technologies, specifically on the maintenance of ‘meaningful human control’ to guarantee that these systems

⁵³⁵ Wendell Wallach & Gary Marchant, *Toward the Agile and Comprehensive International Governance of AI and Robotics*, 107 PROC. IEEE 505–508 (2019). The creation of such a mechanism is the aim of the ‘International Congress for the Governance of Artificial Intelligence (ICGAI), which has been re-scheduled to be held in Prague in May 2021.

⁵³⁶ Olivia J Erdelyi & Judy Goldsmith, *Regulating Artificial Intelligence: Proposal for a Global Solution*, in PROCEEDINGS OF THE 2018 AAAI / ACM CONFERENCE ON ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY 95–101 (2018), http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf; TURNER, *supra* note 17 at 237–253. (discussing the need for, and viability of, an ‘international regulatory agency’).

⁵³⁷ Kemp et al., *supra* note 494.

⁵³⁸ *Id.* at 2–3.

⁵³⁹ A proposed ‘UN AI Research Organisation’. *Id.* at 3. For a similar proposal for a neutral ‘AI research hub’, see SOPHIE-CHARLOTTE FISCHER & ANDREAS WENGER, *A Politically Neutral Hub for Basic AI Research* 4 (2019), http://www.css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.

⁵⁴⁰ Kemp et al., *supra* note 494 at 1.

⁵⁴¹ JELINEK, WALLACH, AND KERIMI, *supra* note 513 at 3.

remain *legal* under International Humanitarian Law (IHL).⁵⁴² However, that may fail to correspond well to the operational intricacies of military AI systems,⁵⁴³ their networked nature,⁵⁴⁴ or their ‘lethality-enabling’ uses.⁵⁴⁵ As such, there is the risk that ensuring narrow compliance with IHL principles occludes other AI uses that are problematic on other (political, legal, or normative) grounds.⁵⁴⁶ This approach may then mean that the regime is brittle: unable to account for—or scale up with—new advances in AI capabilities that shift the technological envelope in ways that, for instance, render the assumption that AI weapons are inherently indiscriminate less viable,⁵⁴⁷ or which shift the risk landscape in other ways.⁵⁴⁸

Ultimately, the effects of the ‘dual trend’ in AI governance (between civil and military governance debates) are unclear. Of course, when considering the history of international law, it may be unsurprising to see a technology’s military applications be simultaneously the area where that technology receives the earliest governance efforts, but also where these are the most contested and intractable. As such, it might be expected that, as Eugenio Garcia has argued, “civilian and military uses [of AI] will presumably follow different multilateral tracks on the road to future international norms.”⁵⁴⁹ Yet this might be a problem, given the generality of the underlying AI technologies. If not handled well, the gridlock of military AI security regimes might well drive regime erosion or fragmentation even in regimes focused on civilian uses of AI.

In sum, while there are early encouraging steps in various aspects of AI governance, the landscape remains incipient, fragmented, and possibly inadequate.⁵⁵⁰ It remains uncertain whether it can reckon with the first-order policy problems created by AI. That makes it even more uncertain whether these initiatives will be up to the task of reckoning with other facets of AI change, including indirect or broader patterns of sociotechnical change, or the technology’s disruptive effects on international law and global governance itself. Finally, given the degree of fragmentation, it is unsure whether these initiatives will be able to adequately manage potentially conflictive norm- institutional, or political interactions across the broader AI ‘regime complex’. AI governance is important, and urgently needed—but it may need new ways to deal with change.

⁵⁴² See Burri, *supra* note 391 at 98–101.

⁵⁴³ Merel Ekelhof, *Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation*, 10 GLOB. POLICY (2019), <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12665> (last visited Mar 27, 2019).

⁵⁴⁴ Hin-Yan Liu, *From the Autonomy Framework towards Networks and Systems Approaches for ‘Autonomous’ Weapons Systems*, 10 J. INT. HUMANIT. LEG. STUD. 89–110 (2019); Léonard Van Rompaey, *Shifting from Autonomous Weapons to Military Networks*, 10 J. INT. HUMANIT. LEG. STUD. 111–128 (2019).

⁵⁴⁵ Michel, *supra* note 382; Deeks, *supra* note 382. This has recently also received more attention, see for instance GIACOMO PERSI PAOLI ET AL., *Modernizing Arms Control: Exploring responses to the use of AI in military decision-making* 52 (2020), <https://unidir.org/publication/modernizing-arms-control>.

⁵⁴⁶ For greater discussion of a broader range of AI ‘problems-logics’, see also the taxonomy in section 4.4.

⁵⁴⁷ Rosert and Sauer, *supra* note 374.

⁵⁴⁸ Maas, *supra* note 378 at 139–141. Paper [II]. See also the discussion in section 5.3.5.

⁵⁴⁹ Garcia, *supra* note 476 at 11.

⁵⁵⁰ See also Paper [IV]: Cihon, Maas, and Kemp, *supra* note 150; See also Luke Kemp, *Fragmented, Incidental and Inadequate: Mapping the Archipelago of AGI Governance* (Forthcoming). (unpublished working draft, shared with author).

2.4 Conclusion: the world needs global governance for AI

This chapter has explored, at some length, the context and background for AI governance—why or how AI matters (2.1); why and where it needs governance (2.2); and why existing governance approaches and initiatives are possibly insufficient or at least fragmented (2.3). In sum, AI technology matters—not universally, but in key ways. It will drive key patterns of change. In many cases, it can be regulated domestically, but in a series of key domains it will benefit from or even require global cooperation. The current trajectory of global AI governance remains uncertain. The question of how best to govern the development and deployment of artificial intelligence is of considerable relevance from a practical perspective. It has potentially far-reaching implications for legal scholarship on technology governance and regulation; and it is a question that would benefit from study, policy formulation and implementation sooner rather than later. Yet it remains an early and developing field. As such, developing new conceptual frameworks may be difficult, but may also contribute to a key foundation for governance debates and initiatives that are of urgent relevance over the coming years.

The above has provided a rationale of this analysis, as well as sketched the technological, political, and institutional background against which it must be situated. Before exploring the three ‘facets’ of change in AI governance, however, we must first discuss the general methodology, approach, and assumptions of this project.

Chapter 3. Methodology, Theory, and Assumptions

I (3.1.) introduce the overall methodology and approach of this project, and discuss some of its strengths and limits. I (3.2.) sketch the three core theoretical lenses of sociotechnical change, governance disruption, and regime complexity, and justify their analytical relevance to AI governance. I (3.3.) clarify my epistemological approach, in terms of the use of cases and examples, as well as the ‘anticipatory’ orientation to future technological- and governance change. I (3.4.) briefly set out the project’s ontological assumptions: I detail the actors, instruments, and processes that make up the global governance architecture. I adopt and defend a hybrid positions about the role of interests and norms-based foundations of international regimes, and defend three soft assumptions about the interaction of technology with human agency (technological politics; unintended consequences; asymmetric governance rationale/target materiality). I (3.5.) conclude by briefly reviewing the methods used.

Part I of this work focuses on establishing the foundations for this project. Chapter 2 articulated its rationale, by discussing how AI matters, why it may require global governance, and why and how current governance approaches fall short. This chapter will introduce the analytical foundations of this project, exploring the overall approach, the theoretical roots, and the assumptions.

3.1 Methodology and approach

This work consists of conceptual and theoretical investigation into the implications of ‘change’ for AI governance. It adopts an explorative and eclectic approach that draws on a range of theories, models, and cases to advance the explanatory power of its analysis.¹ This project locates itself at the intersection of scholarship on law, regulation and technology, and the discipline of ‘International Law and International Relations’ (IL & IR), in order to deploy key concepts and examples developed within those fields. However, as is to be expected of a topic at the intersection of disparate disciplines, the analysis will also draw in insights and models from a range of fields, ranging from AI capability monitoring and forecasting to historical analysis.

¹ It does so in the spirit of the “pragmatic, analytically eclectic, tool-kit approach” that has become prominent within this discipline of international law and international relations. Jeffrey L. Dunoff & Mark A. Pollack, *Reviewing Two Decades of IL/IR Scholarship*, in *INTERDISCIPLINARY PERSPECTIVES ON INTERNATIONAL LAW AND INTERNATIONAL RELATIONS* 626–662, 653 (Jeffrey L. Dunoff & Mark A. Pollack eds., 2012), https://www.cambridge.org/core/product/identifier/CBO9781139107310A039/type/book_part (last visited Oct 3, 2018); cf. ANDREA BIANCHI, *INTERNATIONAL LAW THEORIES: AN INQUIRY INTO DIFFERENT WAYS OF THINKING* 125 (2016). On the role of ‘eclectic theory’ in international relations, see generally Peter Katzenstein & Rudra Sil, *Eclectic Theorizing in the Study and Practice of International Relations*, in *THE OXFORD HANDBOOK OF INTERNATIONAL RELATIONS* (Christian Reus-Smit & Rudra Sil eds., 2008), <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199219322.001.0001/oxfordhb-9780199219322-e-6> (last visited Jan 14, 2020).

Drawing on this scholarship, this project aims to set out three separate lenses of governance analysis. These are explored and illustrated through a range of cases involving the governance of other technologies and other regime issue areas.² Finally, I aim to use these lenses in order to deductively derive implications for AI governance approaches. Of course, there are inherent risks to an eclectic or interdisciplinary approach, which should be admitted and considered.

One potential risk is that applying these theories to the topic of AI governance may export them outside of the empirical or epistemological contexts in which they were originally formulated, eroding their applicability. However, while care is taken to reflect on the limits of some models, this ought not to be a fundamental problem to this approach. This is because many of these theories have been explicitly articulated to be relatively agnostic to the technology (in the case of sociotechnical change and governance disruption) or to the issue area (in the case of regime complexity) under examination. Indeed, others scholars have applied these frameworks to the study of AI-adjacent technologies.³ As such, we should *prima facie* expect these models to be transferable to the topic of AI governance.⁴

A second risk of the eclectic approach is that ‘mixing’ these distinct theories together could erode the coherence of the dissertation’s argument. However, it should be remembered that the primary aim of this project is not to subsume or reconcile these distinct approaches into a single overarching framework or theory, but simply to suggest three promising and commensurable avenues for AI governance, each of which, even taken on its own, might enrich analyses of AI governance. As such, Chapters 4, 5 and 6 explore each of these perspectives in relative isolation from the others. That being said, it is still allowed that these lenses could complement one another, as shown by the integrated discussion in Chapter 7. Even there, however, care will be taken to articulate where and how the distinct lenses intersect.

Finally, a third challenge of this broad methodological apparatus is that it might not leave space to fully and comprehensively work out each lens. To be sure, this project does not claim to offer an exhaustive or definitive application of each individual three lens. It rather aims to set out the sort of questions these lenses enable one to ask—to set out conceptual recipes, rather than presenting the full dish. For instance, in the discussion of the regime complex approach, I do not aim to completely chart all regime interactions (in part because the AI governance architecture remains in such flux, and this would be rapidly outdated), and rather to identify the *type* of links or overlaps which regime complex theory would highlight as potential sources or sites of conflictual interactions, and which AI governance scholars should monitor in greater detail.

² See also section 3.3.1 (on the use of case studies in this project).

³ For instance, the taxonomy of technology-driven legal change developed by Lyria Bennett Moses (which especially informed my discussion of ‘development’ in section 5.3) has also guided other scholars writing on the regulation of AI or new weapons technologies. See for instance Rebecca Crootof, *Regulating New Weapons Technology, in THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 1–25* (Eric Talbot Jensen & Ronald T.P. Alcala eds., 2019); Carlos Ignacio Gutierrez Gaviria, *The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the U.S.*, 2020, https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html (last visited Jul 11, 2020).

⁴ Indeed, if or insofar as these lenses would prove not transferable to AI governance, that would itself provide interesting insights about the possible ‘exceptionality’ of AI technology.

Keeping these caveats in mind, we will now discuss the features and choice for the three conceptual lenses.

3.2 Theoretical Foundations and Rationales

As discussed, the three conceptual lenses I have selected in order to explore patterns of change in AI governance are *sociotechnical change*, *governance disruption*, and *regime complexity*.⁵ These three lenses enable us to consider three facets of change in AI governance: (1) AI technology as *changing rationale and - target* for regulation; (2) AI as *changer of governance*; (3) AI regimes in a *changing governance architecture* (see Figure 3.1).

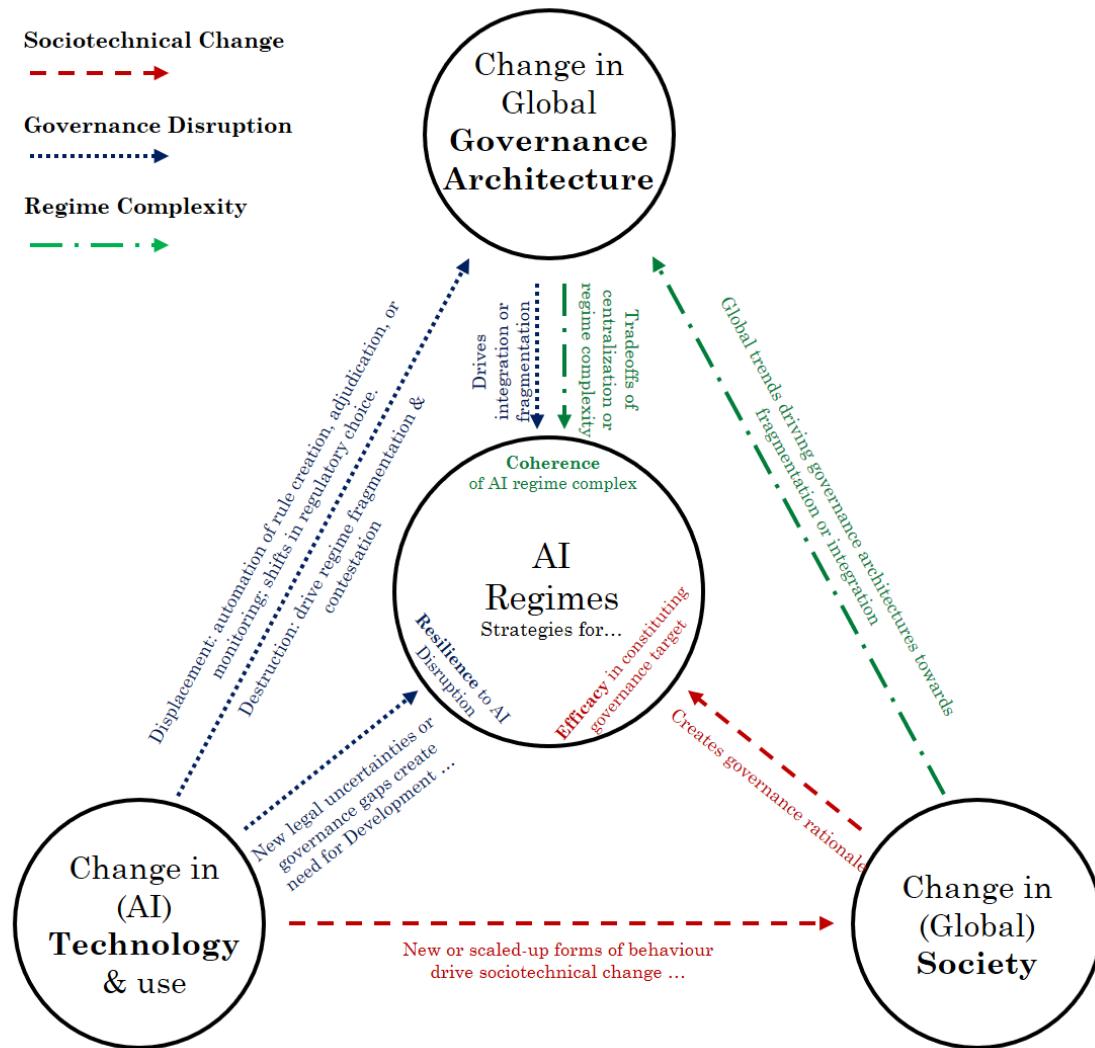


Figure 3.1. AI Governance and Three Facets of Change (revisited)

⁵ Some of these concepts are worked out in greater detail in the core chapters (4-6). However, it is valuable to briefly set out the underlying lineage and applicability of these lenses in this chapter.

Given the complexity of this model, it is valuable to disaggregate these three elements. As such, in the coming sections I introduce these distinct lenses, and for each I will discuss their underlying literatures, justify their application to the area of AI governance, and clarify why and how I draw from them in this project.

3.2.1 Sociotechnical change

The first of these three perspectives, sociotechnical change, focuses on AI-produced societal impacts as *changing rationale and - target for regulation*.

It should be no surprise that changes in technology have given rise to extensive scholarship on the relation between law and these new technologies. In some cases, such work has focused in on identifying the assumed ‘exceptional’ nature or features of a given technology.⁶ However, other scholars have influentially argued that it is less the ‘newness’ of a technology that brings about regulatory problems, but rather the ways it enables particular changes in societal practices, behaviour, or relations.⁷ That is, in what ways do changes in a given technologies translate to new ways of carrying out old conduct, or create entirely new forms of conduct, entities, or new ways of being or connecting to others? When does this create problems for societies or their legal systems? Rather than focus on ‘regulating technology’, such scholarship accordingly puts a much greater emphasis on ‘adjusting law and regulation for sociotechnical change’.⁸

This conceptual shift may be particularly salient to the context of AI governance. As noted, many proposals for legal and governance responses to AI still are reactive, focusing on the presumed newness or ‘exceptionality’ of certain applications.⁹ However, AI governance proposals could be grounded in an improved understanding of the sources of cross-sector societal change driven by AI. The lens of sociotechnical change can be a valuable conceptual tool to guide more refined ways of when and why new AI applications require new regulation—and consequently how such governance interventions could be tailored.

In developing this lens in Chapter 4, I draw particularly on Lyria Bennett Moses’s theory of ‘Law and Technological Change’,¹⁰ and her account of ‘regulating for sociotechnical change’.¹¹ This is complemented with the broader work on the ‘robolaw’ debates,¹² as well as the wide range of sector- and application-specific analyses of AI policy. The sociotechnical change perspective

⁶ Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. LAW REV. 513–564 (2015).

⁷ Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. LAW REV. CIRCUIT 17 (2015); David D Friedman, *Does Technology Require New Law?*, 71 PUBLIC POLICY 16 (2001); Lyria Bennett Moses, *Why Have a Theory of Law and Technological Change?*, 8 MINN. J. LAW SCI. TECHNOL. 589-606. (2007).

⁸ Lyria Bennett Moses, *Regulating in the Face of Sociotechnical Change*, in THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 573–596, 574 (Roger Brownsword, Eloise Scotford, & Karen Yeung eds., 2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49> (last visited May 13, 2017). Note, this approach might have similarities to Schuett’s emphasis on a ‘risk-based approach’ to AI regulation, which does not focus on the umbrella term ‘AI’, but rather on specific risk or outcomes. Jonas Schuett, *A Legal Definition of AI*, ARXIV190901095 CS, 4 (2019), <http://arxiv.org/abs/1909.01095> (last visited Jan 6, 2020).

⁹ Although this is surely not limited to AI governance alone, but has affected many fields of ‘techlaw’. See also Rebecca Crootof & B. J. Ard, *Structuring Techlaw*, 34 HARV. J. LAW TECHNOL. (2021), <https://papers.ssrn.com/abstract=3664124> (last visited Aug 28, 2020).

¹⁰ Bennett Moses, *supra* note 7.

¹¹ Bennett Moses, *supra* note 8.

¹² Calo, *supra* note 6; Balkin, *supra* note 7.

helps provide a more rigorous and theoretically grounded lens on some of the subjects examined in the papers. For instance, Paper [I] articulated four ‘governance rationales’ for arms control, and compared these to the challenges posed by military AI systems.¹³ However, this early argument simply appealed to historical analogy: it essentially argued that if in the past arms control initiatives were motivated by the fact that certain military technologies produced risks in terms of ethics, legality, safety or stability, then it might be expected or reasoned that military AI capabilities that produce such challenges should be subject to similar regimes. However, the sociotechnical change lens allows us to provide a deeper theoretical rationale for why and how particular military AI capabilities might give rise to governance rationales.¹⁴

3.2.1 Governance disruption

The second perspective focuses on AI as *changer of governance*—a ‘governance disruptor’.

This lens responds to the second analytical gap in AI governance, which is that global governance proposals often do not factor in possible technology-driven changes to the instruments and architecture of governance themselves. Of course, the idea that AI technologies might be disruptive to law and regulation is hardly new. Indeed, various scholars have explored the ways AI technology might challenge the conceptual categories of legal systems.¹⁵ Furthermore, there have been extensive exploration of the use of AI, algorithms and digital technology in diverse functions of legal systems.¹⁶ This work has predominantly focused on the effects of such ‘LawTech’ in the domestic law context, although there is a small yet growing body of work on this phenomenon at the international law level.¹⁷ Nonetheless, proposals for the global governance of AI itself still reason in relative isolation from these trends and dynamics.

To be sure, AI technology is hardly the only technology—or indeed trend—that is affecting the coherence, practices or viability of international law. Nonetheless, there may be a range of cases where the use of AI capabilities could drive change in the processes, instruments or political scaffolding of the global legal order. It is important to understand these dynamics, both to understand the trajectory and viability of distinct AI regimes in the first place, but also to ensure these governance solutions can remain adaptive to, or ‘scalable with’, potential future changes in AI technology, its capabilities, or its uses.

Accordingly, Chapter 5 aimed to address this gap, by articulating a governance disruption framework to track and interpret the potential impacts of AI on the global cooperation

¹³ Matthijs M. Maas, *How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons*, 40 CONTEMP. SECUR. POLICY 285–311, 286–287 (2019). Paper [I].

¹⁴ See also Chapter 7.1.1. (exploring emerging military AI systems through this lens).

¹⁵ Calo, *supra* note 6; Margot E. Kaminski, *Authorship, Disrupted: AI Authors in Copyright and First Amendment Law Symposium - Future-Proofing Law: From RDNA to Robots (Part 2)*, 51 UC DAVIS LAW REV. 589–616 (2017); JACOB TURNER, *ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE* 2 (2018); Léonard van Rompaey, *Discretionary Robots: Conceptual Challenges in the Legal Regulation of Machine Behaviour*, 2020.

¹⁶ Roger Brownsword, *In the year 2061: from law to technological management*, 7 LAW INNOV. TECHNOL. 1–51 (2015); Brian Sheppard, *Warming up to inscrutability: How technology could challenge our concept of law*, 68 UNIV. TOR. LAW J. 36–62 (2018); Christopher Markou & Simon Deakin, *Is Law Computable? From Rule of Law to Legal Singularity*, in *IS LAW COMPUTABLE? CRITICAL PERSPECTIVES ON LAW + ARTIFICIAL INTELLIGENCE* (Christopher Markou & Simon Deakin eds., 2020), <https://papers.ssrn.com/abstract=3589184> (last visited May 15, 2020); Rebecca Crootof, “*Cyborg Justice*” and the Risk of Technological-Legal Lock-In, *COLUMBIA LAW REV. FORUM* (2019), <https://papers.ssrn.com/abstract=3464724> (last visited Nov 18, 2019).

¹⁷ Much of which will be explored in Chapter 5 (on Governance Disruption).

architecture itself. This extends a set of themes explored within the core papers: the idea of ‘governance disruption’ was featured in a rough form in Paper [II],¹⁸ and further elaborated in Paper [III], which sketched a taxonomy of Legal ‘Development’, ‘Displacement’, and ‘Destruction’.¹⁹

However, it should be noted that the discussion of this concept, both in the papers as well as in this chapter, builds extensively on the frameworks and case studies developed by a range of scholars in the field of law, regulation and technology. For instance, the exploration of technology’s historical role in shaping and altering international law draws on detailed work by Colin Picker and Rebecca Crootof.²⁰ The taxonomy of ways in which new technology may drive a need for international legal *development* adopts and adapts an influential model, by Lyria Bennett Moses, of how, when or why technological change creates new legal situations.²¹ My discussion of the potential *displacement* of international law,²² drew extensively on work by Thomas Burri,²³ Berenice Boutin,²⁴ Steven Livingston and Matthias Risse,²⁵ and Roger Brownsword,²⁶ and in Chapter 5 was further enriched by Ashley Deeks’ in-depth account of ‘high-tech international law’.²⁷ Thirdly, my account of legal *destruction*²⁸ drew on a variety of work exploring the difficulty

¹⁸ Matthijs M. Maas, *Innovation-Proof Governance for Military AI? How I learned to stop worrying and love the bot*, 10 J. INT. HUMANIT. LEG. STUD. 129–157, 136–149 (2019). *Paper [II]*. Note that this paper drew a distinction that is somewhat different from the one presented in Paper III and in Chapter 5. Specifically, that paper distinguished between ‘direct’ governance-disrupting innovation produced by new entities or capabilities that elude inclusion in existing regimes (what, in Paper III and in section 5.3 would be grouped under ‘development’), and ‘indirect’ governance-disrupting innovation, whereby new capabilities shift the technology’s risk landscape, or change the compliance incentives for states (what, in Paper III and in section 5.5 would be grouped under ‘Destruction’).

¹⁹ Matthijs M. Maas, *International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order*, 20 MELB. J. INT. LAW 29–56 (2019). *Paper [III]*.

²⁰ Specifically, the initial review of the historical relation between international law and new technology drew on the work by Colin Picker and Rebecca Crootof. Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDOZO LAW REV. 151–219 (2001); Rebecca Crootof, *Jurisprudential Space Junk: Treaties and New Technologies*, in RESOLVING CONFLICTS IN THE LAW 106–129 (Chiara Giorgetti & Natalie Klein eds., 2019), <https://brill.com/view/book/edcoll/9789004316539/BP000015.xml> (last visited Mar 15, 2019).

²¹ Bennett Moses, *supra* note 7.

²² Maas, *supra* note 19 at 43–49. *Paper [III]*.

²³ Thomas Burri, *International Law and Artificial Intelligence*, 60 GER. YEARB. INT. LAW 91–108 (2017).

²⁴ Berenice Boutin, *Technologies for International Law & International Law for Technologies*, GRONINGEN JOURNAL OF INTERNATIONAL LAW (2018), <https://grojil.org/2018/10/22/technologies-for-international-law-international-law-for-technologies/> (last visited Oct 31, 2018).

²⁵ Steven Livingston & Matthias Risse, *The Future Impact of Artificial Intelligence on Humans and Human Rights*, 33 ETHICS INT. AFF. 141–158 (2019).

²⁶ Roger Brownsword, *Law and Technology: Two Modes of Disruption, Three Legal Mind-Sets, and the Big Picture of Regulatory Responsibilities*, 14 INDIAN J. LAW TECHNOL. 1–40 (2018).

²⁷ Ashley Deeks, *High-Tech International Law*, 88 GEORGE WASH. LAW REV. 575–653 (2020). In addition, this also draws on the brief but focused account by Bryant W. Smith in: Bryant Walker Smith, *New Technologies and Old Treaties*, 114 AJIL UNBOUND 152–157 (2020).

²⁸ Maas, *supra* note 19 at 50–55. *Paper [III]*.

of adapting international law to the regulatory texture of new AI capabilities,²⁹ as well as more general concerns over the erosive effects of emerging technology on the global legal order.³⁰

3.2.2 Regime complexity

The regime complexity lens focuses on *AI regimes in a changing governance architecture*. This lens helps focus attention beyond the specific features of one or another AI governance instrument or institution, and instead throws into relief the importance for such institutional proposals to look beyond the technology itself, and to consider underlying changes in the governance landscape in which they will emerge, as well as how (in a ‘fragmented’ regime) they relate to other institutions or regimes in a productive and non-conflictive manner. To this end, I argue we can draw valuable lessons from the lens of regime complex theory.³¹

To provide a brief primer, this theory departs from the concept of an international ‘regime’, which Krasner defined as “sets of implicit or explicit principles, norms, rules and decision-making procedures around which actors’ expectations converge in a given area of international relations”³² A regime is characterised by the norms or principles which it represents, which can end up having distinct effects on the expectations, calculations, and ultimately behaviour of actors.³³ However, it also interacts with—and for its establishment, can depend upon—the interests of these actors.³⁴

Moreover, as a result of the increasing ‘fragmentation’ of international law and the global governance architecture more broadly,³⁵ there has been increasing attention to the potential interactions of distinct regimes with one another. This led to the development of the concept of a

²⁹ Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, HARV. J. LAW TECHNOL. (2016), <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf> (last visited Mar 5, 2018).

³⁰ Richard Danzig, *An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order*, 5 LAWFARE RES. PAP. SER. (2017), <https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf> (last visited Sep 1, 2017).

³¹ Amandine Orsini, Jean-Frédéric Morin & Oran Young, *Regime Complexes: A Buzz, a Boom, or a Boost for Global Governance?*, 19 GLOB. GOV. REV. MULTILATERALISM INT. ORGAN. 27–39 (2013); Benjamin Faude & Thomas Gehring, *Regime Complexes as Governance Systems*, in RESEARCH HANDBOOK ON THE POLITICS OF INTERNATIONAL LAW 176–203 (Wayne Sandholtz & Christopher Whytock eds., 2017), <https://www.elgaronline.com/view/9781783473977.xml> (last visited Oct 15, 2019); Karen J. Alter & Kal Raustiala, *The Rise of International Regime Complexity*, 14 ANNU. REV. LAW SOC. SCI. 329–349 (2018); Laura Gómez-Mera, Jean-Frédéric Morin & Thijs Van De Graaf, *Regime Complexes*, in ARCHITECTURES OF EARTH SYSTEM GOVERNANCE: INSTITUTIONAL COMPLEXITY AND STRUCTURAL TRANSFORMATION 137–157 (Frank Biermann & Rakhyun E. Kim eds., 2020). It should be noted that authors have often used the terms ‘regime complex(es)’ or ‘regime complexity’ interchangeably, although the latter is more influenced by the work on complex systems theory; Karen J. Alter & Sophie Meunier, *The Politics of International Regime Complexity*, 7 PERSPECT. POLIT. 13–24 (2009).

³² Stephen D. Krasner, *Structural Causes and Regime Consequences: Regimes as Intervening Variables*, 36 INT. ORGAN. 185–205, 186 (1982).

³³ Arthur Stein defers from this definition, arguing that regimes are not at their core characterised by (positive) norms, since there are examples of ‘unprincipled’ regimes (such as OPEC); instead, he argues that regimes are principally constituted by the presence of ‘joint decisionmaking’ (or, the absence of independent decisionmaking). Arthur A. Stein, *Coordination and Collaboration: Regimes in an Anarchic World*, 36 INT. ORGAN. 299–324, 10 (1982). Nonetheless, for my purposes, I consider such cases as relatively narrow exceptions to the general rule.

³⁴ See also the discussion in Chapter 7.1.2 (on the role of interests and norms in establishing the viability of a given AI regime).

³⁵ Frank Biermann et al., *The Fragmentation of Global Governance Architectures: A Framework for Analysis*, 9 GLOB. ENVIRON. POLIT. 14–40 (2009); Amitav Acharya, *The Future of Global Governance: Fragmentation May Be Inevitable and Creative Global Forum*, GLOB. GOV. 453–460 (2016).

‘regime complex’. Originally, this was defined by Kal Raustiala and David Victor as “an array of partially overlapping and nonhierarchical institutions governing a particular issue-area”.³⁶ This definition has been critiqued, however, on the basis that it is unclear whether ‘institution’ refers only to formal international organisations, or all elements of a ‘regime’, such as norms.³⁷ Accordingly, this project uses the broader, Krasnerian definition of regimes (as including both institutions and norms), and instead draws on the operationalisation of regime complexity offered by Orsini, Morin and Young, who have defined it as: “[a] network of three or more international regimes on a common issue area. These should have overlapping membership and cause potentially problematic interactions.”³⁸

The regime complexity lens may provide a valuable lens for the exploration of structural changes in the governance architecture of AI governance. This is for several reasons. In the first place, the field of AI governance can benefit from cross-fertilisation with many existing and established theoretical frameworks in governance scholarship. There is an extensive and mature body of scholarship exploring regime complexes, yet to date this has not yet been applied to the subject of AI governance.³⁹ Similarly, the regime complex approach enables AI governance scholars to better consider the interaction of top-down (e.g. international legal norms) and bottom-up (e.g. behaviour by state or non-state actors) drivers of governance change. In so doing, we can build on three decades of scholarship in the paradigm of ‘International Law and International Relations’, which has tried to bridge and unite these distinct perspectives,⁴⁰ arguing that scholars in these fields can learn much from one another.⁴¹

More concretely, regime complex theory has proven both empirically and theoretically rich. While the majority of the empirical work in this field has been in the field of global

³⁶ Kal Raustiala & David G. Victor, *The Regime Complex for Plant Genetic Resources*, 58 INT. ORGAN. 277–309, 279 (2004).

³⁷ David J. Galbreath & Sascha Sauerteig, *Regime complexity and security governance*, in HANDBOOK OF GOVERNANCE AND SECURITY 82–97, 84 (James Sperling ed., 2014), <https://www.elgaronline.com/view/edcoll/9781781953167/9781781953167.00014.xml> (last visited Oct 15, 2019). As such, Alter & Meunier reasserted the broader, Krasnerian conception of ‘regimes’, and extended the term ‘international regime complexity’ as “the presence of nested, partially overlapping, and parallel international regimes that are not hierarchically ordered.” Alter and Meunier, *supra* note 31 at 13.

³⁸ Orsini, Morin, and Young, *supra* note 31 at 29. See also the discussion in Chapter 7.

³⁹ With the exception of our own Paper [IV]. Peter Cihon, Matthijs M. Maas & Luke Kemp, *Should Artificial Intelligence Governance be Centralised?: Design Lessons from History*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 228–234 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375857> (last visited Feb 12, 2020). As well as indirectly and briefly Jean-Frédéric Morin et al., *How Informality Can Address Emerging Issues: Making the Most of the G7*, 10 GLOB. POLICY 267–273 (2019).

⁴⁰ Anne-Marie Slaughter Burley, *International Law and International Relations Theory: A Dual Agenda*, 87 AM. J. INT. LAW 205–239 (1993); Anne-Marie Slaughter, *International Law and International Relations Theory: Twenty Years Later*, in INTERDISCIPLINARY PERSPECTIVES ON INTERNATIONAL LAW AND INTERNATIONAL RELATIONS 611–625 (Jeffrey L. Dunoff & Mark A. Pollack eds., 2012), https://www.cambridge.org/core/product/identifier/CBO9781139107310A038/type/book_part (last visited Oct 3, 2018); Dunoff and Pollack, *supra* note 1. However, for sharp critiques of the interaction between these two disciplines, and how it might have disproportionately favoured IR over IL, see also BIANCHI, *supra* note 1 at 6; Martti Koskeniemi, *Law, Teleology and International Relations: An Essay in Counterdisciplinarity*, 26 INT. RELAT. 3–34 (2012).

⁴¹ For instance, Gehring & Faude have suggested that, just as international legal scholars can profit from IR research on ‘cross-institutional strategizing’ by state actors, IR scholars “can learn from IL research on the endogenous dynamics of international law, in particular on legal ways to absorb conflict among overlapping sets of legal rules in the absence of a ‘final authority.’” Faude and Gehring, *supra* note 31 at 177.

environmental governance,⁴² scholars have also productively applied this lens across a wide range of other areas of global governance—human rights,⁴³ international security,⁴⁴ refugees,⁴⁵ energy,⁴⁶ maritime piracy,⁴⁷ food security,⁴⁸ and many others. This provides an extensive reference class, and allows potentially fruitful comparisons and cautious lessons to be drawn for the emerging AI governance architecture. Likewise, theoretically, the regime complexity framework has come to encapsulate diverse facets and dynamics of governance architectures—from the way norms and interests set the condition for a regime, to its evolution over time, to its productive ‘management’. This suggests it may shed light on various steps or considerations that will be critical in the emerging AI governance architecture, and indeed an exploration of some such trade-offs forms the basis of Paper [IV].⁴⁹

Indeed, the regime complex framework may be especially appropriate to studying governance systems for a rapidly changing technology. Saliently, regime complex scholars, especially those working in the environmental law field, have long been interested in identifying adaptive governance systems that are able to adapt or anticipate change, and are therefore capable of addressing the “institutional mismatch” between fixed political institutions and rapidly changing systems.⁵⁰ This may make this lens particularly relevant to investigating dynamics of change around AI governance, and especially the question of which governance systems might be more adaptable to sociotechnical change or governance disruption from AI.

Furthermore, regime complexity is concerned with how regimes may experience conflictive interactions or cross-institutional spillover or externalities.⁵¹ This again makes it promising for AI governance, because it can be hard to draw strict silos around AI’s capabilities, use cases, or resulting societal challenges.⁵² Accordingly, any institution or regime that seeks to regulate any one local AI problem is likely to inadvertently touch on regimes across many other AI issue areas. That suggests that, in a fragmented AI governance architecture, it is unlikely that individual regimes or institutions focusing on a single aspect of AI would be able to operate in strong isolation from the effects of regimes or institutions focused on other aspects, nor from existing regimes focusing on other issue areas such as human rights, security or the

⁴² Gómez-Mera, Morin, and Van De Graaf, *supra* note 31 at 137–138. (discussing reasons for the theoretical density in this particular field).

⁴³ Emilie M. Hafner-Burton, *The Power Politics of Regime Complexity: Human Rights Trade Conditionality in Europe*, 7 PERSPECT. POLIT. 33–37 (2009).

⁴⁴ Stephanie C. Hofmann, *Overlapping Institutions in the Realm of International Security: The Case of NATO and ESDP*, 7 PERSPECT. POLIT. 45–52 (2009); Galbreath and Sauerteig, *supra* note 37.

⁴⁵ Alexander Betts, *Regime Complexity and International Organizations: UNHCR as a Challenged Institution*, 19 GLOB. GOV. 69–81 (2013).

⁴⁶ Jeff D Colgan, Robert O Keohane & Thijs Van de Graaf, *Punctuated equilibrium in the energy regime complex*, 7 REV. INT. ORGAN. 117–143 (2012).

⁴⁷ Michael J. Struett, Mark T. Nance & Diane Armstrong, *Navigating the Maritime Piracy Regime Complex*, 19 GLOB. GOV. 93–104 (2013).

⁴⁸ Matias E. Margulis, *The Regime Complex for Food Security: Implications for the Global Hunger Challenge*, 19 GLOB. GOV. 53–67 (2013).

⁴⁹ Cihon, Maas, and Kemp, *supra* note 39. Paper [IV]. See also the discussion in Chapter 7.2-7.4.

⁵⁰ Gómez-Mera, Morin, and Van De Graaf, *supra* note 31 at 138.

⁵¹ *Id.* at 138.

⁵² See also the discussion in Chapter 2.1.1 (on defining AI), as well as in Chapter 4.

environment.⁵³ As such, given the extreme versatility of many AI techniques, their governance might be a ‘natural candidate’ for the regime complex lens.

That is not to say that this theory is without its methodological shortfalls or limits.⁵⁴ Moreover, it should be kept in mind that this lens has been developed in different issue domains. As such, the patterns or findings of regime complex theory should not be uncritically transferred over into AI governance. On the other hand, neither should the field of AI governance ignore these lessons or reinvent the wheel. Regime complexity offers a potentially promising and underexplored lens to many aspects of AI governance.

3.3 Epistemological Approach

Having set out the overarching methodology and justified the selection and use of these three lenses, I now briefly provide some clarification of the epistemological approach that underpins both the core papers as well as this contribution. This relates specifically to the use of cases, as well as how this argument relates to or aims to anticipate future trends.

3.3.1 Scope and use of examples

In the core chapters (4-6), each of the three frameworks will be introduced and discussed in turn. For each, I provide theoretical background and key definitions. I then explore these lenses with reference to diverse instances or areas of technological change and disruption. Finally, the implications of these models for AI governance are explored for distinct or representative issues.

This mirrors the core papers, which involved various historical comparisons or cases in order to illustrate the theory under discussion. For instance, in exploring the viability of arms control for military AI, Paper [I] explored the various historical drivers of the proliferation or non-proliferation of nuclear weapons,⁵⁵ as well as accounts of the ‘epistemic community’ involved in

⁵³ This latter case touches on the question of how different regimes focusing on distinct issue areas interact with one another—cf. BEATRIZ MARTINEZ ROMERA, REGIME INTERACTION AND CLIMATE CHANGE: THE CASE OF INTERNATIONAL AVIATION AND MARITIME TRANSPORT (2017), <https://www.taylorfrancis.com.ep.fjernadgang.kb.dk/books/9781315451817> (last visited Jan 11, 2020). This is an important research agenda, and reflects the analysis of AI regime complex ‘topology’ at the *macro* level (see section 7.2.4). However, most of the analysis in this project will remain focused on the (meso or micro-level) ‘internal’ interactions of institutions within the emerging AI regime complex, rather than those macro-level interaction of that AI regime with, say, the separate international labour regimes.

⁵⁴ For earlier critiques of the ‘regimes’ concept, see for instance Susan Strange, *Cave! Hic Dragones: A Critique of Regime Analysis*, 36 INT. ORGAN. 479–496 (1982); James F Keeley, *The Latest Wave: A Critical Review of Regime Literature*, in WORLD POLITICS: POWER, INTERDEPENDENCE AND DEPENDENCE 553–569 (David G. Haglund & Michael K. Hawes eds., 1990). The regime complex framework likewise has its challenges. For instance, Gómez-Mera and others admit that “for all the progress made in the identification of causal pathways and mechanisms through which regime complexity matters, we still lack a coherent and comprehensive theory of regime complexes and their implications” Gómez-Mera, Morin, and Van De Graaf, *supra* note 31 at 149–150. Others have argued that many of the effects of regime complexes have been theorised deductively, or that these have been considered only “with reference to a small population of specific regime complexes [such that] it is probably fair to say that a broader empirical foundation is necessary to account for a general understanding of the correlation between the characteristics of a complex and governance outcomes.” Galbreath and Sauerteig, *supra* note 37 at 87.

⁵⁵ Maas, *supra* note 13 at 291–296. Paper [I]. This drew especially on influential work in this field, such as: Scott D. Sagan, *Why Do States Build Nuclear Weapons?: Three Models in Search of a Bomb*, 21 INT. SECUR. 54–86 (1996); Etel Solingen, *The Political Economy of Nuclear Restraint*, INT. SECUR. 126–159 (1994); Sico van der Meer,

setting the foundations for the 1972 Anti-Ballistic Missile (ABM) Treaty.⁵⁶ Likewise, papers [II] and [III] drew on diverse examples of ways new technologies have historically and recently disrupted international law.⁵⁷ Finally, paper [IV], which introduced the regime complexity lens, surveyed examples and experiences drawn from the recent history of various other international regimes, such as in security, trade, or the environment.⁵⁸

It is important to clarify that the historical examples and scenarios discussed in these chapters are not meant as exhaustive and focused case studies. Rather, they are offered as clarifying cases of certain technology governance dynamics. They illustrate the operation of some of the processes discussed—for instance, the conditions under which new technological change is more or less likely to drive the obsolescence of treaty regimes—but they do not establish the relative likelihood or frequency of such situations, and do not yet suffice to ground specific predictions about AI. Instead, the taxonomies sketched are meant to identify and index distinct patterns of change (in society; in legal systems; in governance architectures) that could and should be of analytical and practical interest to AI governance. However, it is admitted that each and all will need much further investigation. They are starting points for the emerging field.

As a result, the range of examples discussed to illustrate and support these frameworks is relatively broad. This befits the analytical breadth of the three conceptual frameworks themselves. For instance, the lens of sociotechnical change explicitly aims to facilitate technology governance analysis for diverse cases of technological change. For its part, while regime complex theory has received most of its initial articulation and development in the area of international environmental governance, it too has been used to capture regime complex dynamics across diverse issue areas. This breadth of scope also befits the sheer breadth of AI governance problems, given the technology's cross-sector range of application, and its propensity to blur previous categorisations or symbolic orders. At the same time, casting a broad net of cases creates the risk that examples might be contrived, unrepresentative or cherry-picked. As such, for the most detailed and critical cases discussed, I have tried to clarify my assumptions and nuance the strengths and limits of the comparison drawn between the technology and the AI governance context.⁵⁹

Finally, I aim to apply these findings to a diverse set of AI governance challenges in different sectors. Nonetheless, a number of sections will focus in on issues in the context of international security. This includes examining the historical legal impact of military technologies, as well as applying the three lenses to the international governance of emerging

Forgoing the nuclear option: states that could build nuclear weapons but chose not to do so, 30 MED. CONFL. SURVIV. s27–s34 (2014).

⁵⁶ Maas, *supra* note 13 at 296–299. Paper [I]. Drawing especially on: Emanuel Adler, *The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control*, 46 INT. ORGAN. 101–145 (1992).

⁵⁷ Maas, *supra* note 19 at 34–37. Paper [III]. This particularly drew from Picker, *supra* note 20 at 157–163. As well as the discussion of how changing customary norms around submarine warfare rendered the earlier 1930 London Treaty and the 1936 London Protocol obsolete ‘jurisprudential space junk’. Crootof, *supra* note 20 at 113–114.

⁵⁸ Cihon, Maas, and Kemp, *supra* note 39 at 229–232. Paper [IV].

⁵⁹ One example is the discussion, in *Paper I*, on nuclear weapons as a comparator for military AI. Maas, *supra* note 13 at 288–290. See also the discussion of historical arms control in chapter 7.1.2.1.

military AI systems.⁶⁰ This in part reflects the prevailing focus of the core papers.⁶¹ Moreover, this focus area is also supported by the fact that new military technologies have historically proven a key rationale for- and disruptor of international law.⁶² As such, to some degree, security regimes for new weapons technologies might serve as a valuable (although certainly not decisive) litmus test of the broader pressures or challenges on a global governance system. Finally, as previously noted, the regulation of military AI systems has proven a particularly high-profile issue area for AI governance scholarship, making this a salient domain in which to highlight the contributions of these lenses.

That is not to say that the use of examples from the military domain is without analytical drawbacks. For instance, the fact that current AI global governance efforts appear to largely take separate and parallel tracks for military and civilian AI applications⁶³ suggests that we should be cautious about generalizing the conclusions from our analysis of one domain to the other. At the same time, rather than producing *conclusions* (e.g. about the viability of regimes, or the direction of regime evolution) that are generalizable to the entire AI governance architecture, these cases can be used to identify the *type of questions* that should be asked around AI governance regimes in these various areas.⁶⁴

In sum, the use of examples in this project will be illustrative and explorative. They are generally broad, though a number of sections will focus on (technology) regimes in the international security domain.

3.3.2 Present and future: on speculation and anticipatory governance

Finally, another key question concerns this project's relation to the future. While the three theoretical models are articulated through reference to a diverse set of past cases, their application to problems in AI governance faces the challenge that both the technology, and especially its governance landscape, remain incipient and very much in flux. Given the emphasis of this project on ongoing 'change' in AI governance, how should it balance analyses of present-day governance challenges of AI with an exploration of how these might evolve into the near-term future? Is there a way to do so without falling into the epistemic traps of either trajectory blindness or speculative prediction?

To be certain, nearly all work on technology governance must, to some extent, grapple with the question of future uncertainty. This applies doubly so at the global level, where not only

⁶⁰ Most prominently in chapter 7.1.

⁶¹ In particular Papers [I-II], though Papers [III-IV] also involved some discussion of international security challenges.

⁶² Cf. Picker, *supra* note 20; Braden Allenby, *Are new technologies undermining the laws of war?*, 70 BULL. AT. SCI. 21–31 (2014); Rosemary Rayfuse, *Public International Law and the Regulation of Emerging Technologies*, in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY (2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22> (last visited Jan 3, 2019).

⁶³ Eugenio V Garcia, *Multilateralism and Artificial Intelligence: What Role for the United Nations?*, in THE GLOBAL POLITICS OF ARTIFICIAL INTELLIGENCE 18, 11 (Maurizio Tinnirello ed., 2020) ("Civilian and military uses will presumably follow different multilateral tracks on the road to future international norms"). See also section 2.3.2.4 (on strengths and limits of current AI governance approaches).

⁶⁴ In this sense, this project could also be understood as aiming to adopt both a 'problem-finding' with a problem-solving orientation. On this distinction, see also Hin-Yan Liu & Matthijs M. Maas, 'Solving for X?' Towards a problem-finding framework that grounds long-term governance strategies for artificial intelligence, FUTURES 35 (2020).

the technology's development, but also the future trajectory of the governance landscape remains beset by uncertainty. In particular, given the relative novelty and immaturity of many AI uses, there appear to be clear methodological or epistemic difficulties in studying their second- or third-order effects of their proliferation and use.⁶⁵ Many have noted that the difficulties around predicting the trajectory and societal salience of a new technology yields practical problems not just for scholars but also for policymakers.⁶⁶ As such, the uncertainties around either the technological dimension (the future of AI), or the societal dimension (the future of global governance), help drive home the difficulty of charting future trajectories of AI governance.

However, while it is warranted to be cautious and concerned about the risks of unstructured predictions about the future, there is certainly both scientific and practical ground for pursuing 'anticipatory' analyses of potential future trajectories. On the one hand, there is established philosophical and scientific support for the role of productive 'speculative research' in contexts where key information is constrained,⁶⁷ providing one tailors the analytical tools to the epistemic context at hand.⁶⁸ Practically, many scholars of international law and technology have argued that while uncertainty around future trajectories is unavoidable when studying emerging technologies, this is exactly one reason why a 'wait-and-see' approach may not be adequate.⁶⁹ Particularly in areas involving the international regulation of new technological risks, these scholars have argued in favour of pre-emptive or anticipatory analysis.⁷⁰ Accordingly, scholars of law and technology have increasingly emphasised the importance of engaging with future uncertainty, discussing the general utility of future-oriented policymaking more broadly;⁷¹ the more specific relevance of 'legal forecasting' in the face of technological uncertainty;⁷² and the

⁶⁵ Michael Horowitz, for instance, has argued that the scholarship on military AI is even more incipient than that on drones or cyberwar. Michael C. Horowitz, *Do Emerging Military Technologies Matter for International Politics?*, 23 ANNU. REV. POLIT. SCI. 385–400 (2020).

⁶⁶ For instance, Ballard and Calo note that these challenges include 'staleness' (where outdated rules nevertheless persist through inertia or now-entrenched interests); 'waste' (as a result of regulatory over-investment in a particular instantiation of a technology that has become superseded); and 'policy paralysis', which risks "abdicate governmental responsibility to channel technology in the public interest."

Stephanie Ballard & Ryan Calo, *Taking Futures Seriously: Forecasting as Method in Robotics Law and Policy* 22, 1 (2019), https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/Calo_Taking-Futures-Seriously.pdf.

⁶⁷ As defined by Swedberg, speculation in the social sciences can be defined as "the use of guesses, conjectures and similar ways of thinking, that help the scientist to come up with explanations and redefinitions of phenomena, in situations where important facts are missing." Richard Swedberg, *Does Speculation Belong in Social Science Research?*, SOCIOLOGICAL METHODS & RES. 0049124118769092, 19 (2018).

⁶⁸ Adrian Currie & Shahar Avin, *Method Pluralism, Method Mismatch, & Method Bias*, 19 PHILOSOPHY & IMPRINTING (2019), <http://hdl.handle.net/2027/spo.3521354.0019.013> (last visited Apr 9, 2019); Adrian Currie, *Existential Risk, Creativity & Well-Adapted Science*, 76 in STUDIES IN THE HISTORY & PHILOSOPHY OF SCIENCE. 39–48 (2019), <https://www.sciencedirect.com/science/article/abs/pii/S0039368117303278?via%3Dihub>.

⁶⁹ Crootof, *supra* note 3 at 21–22.

⁷⁰ For instance, Eric Talbot Jensen is a proponent of such a anticipatory approach, arguing that "[j]ust as military practitioners work steadily to predict new threats and defend against them, [law of armed conflict] practitioners need to focus on the future of armed conflict and attempt to be proactive in evolving the law to meet future needs." See Eric Talbot Jensen, *The Future of the Law of Armed Conflict: Ostriches, Butterflies, and Nanobots*, 35 MICH. J. INT'L. LAW 253–317, 254 (2014). Likewise, Paul Scharre notes, for instance, how "one need not wait (and should not wait) until nuclear war breaks out to think about stability dynamics with nuclear weapons." Paul Scharre, *Autonomous Weapons and Stability*, March, 2020.

⁷¹ See for instance THE POLITICS AND SCIENCE OF PREVISION: GOVERNING AND PROBING THE FUTURE, (Andreas Wenger, Ursula Jasper, & Myriam Dunn Cavelty eds., 2020).

⁷² Graeme Laurie, Shawn H. E. Harmon & Fabiana Arzuaga, *Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty*, 4 LAW INNOVATION & TECHNOLOGY 1–33 (2012).

value of distinct foresight and forecasting approaches for AI policy and governance.⁷³ This suggests that anticipatory governance analysis can be warranted, provided it is bracketed by appropriate caution and reflection.

At the same time, it should be emphasised that this is not about pursuing prediction. Indeed, especially in the international context, accurate technology forecasting may not be an unalloyed good, and could even be politically *counterproductive* to establishing some regimes. Scholars have noted that negotiations to establish technology regimes can be held up by deeply unequal stakes.⁷⁴ Conversely, scholars of international institutions have argued that shared uncertainty over the state (or future) of the world can in some cases support international cooperation and the creation of robust international institutions.⁷⁵ Given this, there is a risk around prediction—which is not that it is hard to get right, but that we should not always want it to be. After all, as I previously argued in *Paper [II]*:

“the more clearly a technology’s future impacts are envisioned, the more clearly states today are able to project and articulate how their and other states’ relative proficiency in the relevant areas of science and technology will eventually translate into concrete differential strategic advantages – and the more reluctant they may be to accede to governance regimes that would see them forgo such advantages. [...] uncertainty may have its benefits; and governance regimes which are ‘blind’ but adaptive might prove more effective than governance regimes which are based on clear roadmaps of what interests and stakes, exactly, will be on the table.”⁷⁶

With all that, this work takes an inquisitive but cautious approach to the future. It recognizes that scholarship in this area cannot avoid reckoning with potential trajectories. Yet I do not aim to make specific predictions about particular future directions in either AI technology or global governance, nor do the arguments here make strong assumptions about these trends. Rather, the aim is to articulate a set of analytical frameworks that can provide insight into AI governance today, and would remain applicable and analytically relevant across a wide range of near- to mid-term scenarios for both AI usage or governance development.⁷⁷ As such, in working out the implications of these lenses for AI governance, I aim, in the first instance, to discuss existing challenges or governance initiatives around existing AI use cases, or those that would arise from the further proliferation of- or marginal improvement over existing capabilities.⁷⁸

Nonetheless, this project, and especially the lenses of sociotechnical change and governance disruption are grounded in the expectation of *some* further future change in AI technology or its usage. Notably, this does not require or assume the appearance of specific future

⁷³ Ballard and Calo, *supra* note 66.

⁷⁴ Picker, *supra* note 20 at 191–194.

⁷⁵ Barbara Koremenos, Charles Lipson & Duncan Snidal, *The Rational Design of International Institutions*, 55 INT. ORGAN. 761–799, 778 (2001).

⁷⁶ Maas, *supra* note 18 at 150. *Paper [II]*. (arguing on this basis that “the difficulty of accurate technological foresight arguably underscores rather than erodes the utility of pursuing governance strategies or instruments that are versatile even if – or precisely when – they cannot accurately predict developmental pathways”).

⁷⁷ Though not unrestricted. There are distinct scenarios that would break the boundary conditions or exceed the assumptions of the framework explored here. These could include a global backlash to AI technology; (conversely) sudden and particularly disjunctive capability breakthroughs (see Section 2.1.6); or some form of global systemic conflict.

⁷⁸ See also the discussion of the proliferation of disruptive AI in chapter 2.1.7.

AI technologies, but is rather generally rooted in a soft but defensible expectation of some forms of continued technological progress in AI—or diffusion of AI—over the coming years.⁷⁹ As such, in discussing the model of governance disruption, the argument is not that all future AI applications shall produce such disruption, nor to predict in great detail which types of AI applications *will* do so. Rather, examples and projections of cases are established simply to illustrate the conditions under (and the ways by) which certain AI capabilities could or would drive disruption, if adopted.

Likewise, this project also does not seek to provide strong predictions of the future of AI governance, but simply to explore various drivers of these trajectories, and their respective consequences and implications.⁸⁰ The hope is that this may help clarify how different assumptions about the pace and development of AI, or about broader trends in global governance, may produce different judgments about the relative success of different regime configurations, and the need for strategies to ensure the efficacy, resilience, or coherence of that governance system.

3.4 Ontological Assumptions

We will now turn to the epistemological and ontological assumptions underpinning this project. These relate to the constitution and dynamics of the global governance architecture, as well as to the interrelation between technological change and global society.

3.4.1 The global governance architecture

‘Global governance’ is a broad and vague concept, with contested analytical and normative uses.⁸¹ For the purposes of scholarship, however, I approach it as an analytical lens on the evolving structure of global cooperation. This structure is populated and characterised by an array of actors, instruments, and structures, which are often subject to distinct terms, and as such it is valuable to lay out these concepts in some greater detail.

At the highest level, *international order* “reflects the organization of the entire system of international relations”.⁸² This world order is populated by a *global governance complex* which, in a definition by Kacowicz, embraces “states, international institutions, transnational networks, and agencies (both public and private) that function, with variable effect and effectiveness, to promote, regulate, and manage the common affairs of humanity.”⁸³

⁷⁹ These assumptions were explored and defended in Chapter 2.1.6-2.1.7.

⁸⁰ In doing so, this approach takes some cue from the one used by Baum and others to model possible ‘long-term trajectories for human civilization’, where the focus is on clarifying conditions, assumptions and dependencies, more than on making specific predictions or even attempting to assign probabilities to- or between scenarios. Seth D. Baum et al., *Long-term trajectories of human civilization*, 21 FORESIGHT 53–83 (2019).

⁸¹ Klaus Dingwerth & Philipp Pattberg, *Global Governance as a Perspective on World Politics*, 12 GLOB. GOV. 185–203 (2006).

⁸² Biermann et al., *supra* note 35 at 16. Referring to HEDLEY BULL, THE ANARCHICAL SOCIETY: A STUDY OF ORDER IN WORLD POLITICS (1977).

⁸³ Arie M. Kacowicz, *Global Governance, International Order, and World Order*, OXF. HANDB. GOV., 689 (2012), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199560530.001.0001/oxfordhb-9780199560530-e-48> (last visited May 15, 2018).

As such, the global governance complex is characterised by a diverse ecology of actors and instruments, many of which focus on distinct issue areas.⁸⁴ A range of distinctions and subdivisions should be made to chart the system, its elements, and the often complex interrelations amongst them. This will necessarily be a cursory and brief overview of a complex ecology, however, one can loosely distinguish between actors, instruments, overall developments and trends, and the regime dynamics.⁸⁵

3.4.1.1 Actors

As noted, traditional accounts hold that international law is primarily produced through interactions by *states*, defined as sovereign entities with a population, a territory and a governmental organisation.⁸⁶ In exercising their sovereignty, states can voluntarily enter into (i.e. consent to) various international agreements. Nonetheless, over the past decades, there has been, if not a decline of states, at least increased roles of other actors.

A second group of actors which has taken a growing role in the development of international law is found in *international organisations*. Under international law, an international organisations is defined by an international agreement that establishes them, a capability to act independently from its member states, and subjection to international law.⁸⁷ Beyond this, however, international organisations compose a highly diverse group of entities. In membership, they can range from global institutions such as the UN to thematic organisations such as the OECD to bilateral institutions; they can be formal international financial institutions such as the World Bank or the ITU, to informal clubs such as the G7.⁸⁸ Moreover, in recent decades, there has been a sharply growing role of various *non-state actors* in global governance, including NGOs, private sector actors, public interest organisations, knowledge-based ‘epistemic communities’,⁸⁹ and various ‘trans-national issue networks’.⁹⁰ These actors have distinct roles, interests, and tasks, and create more heterogeneous dynamics in the creation of norms or laws.⁹¹

⁸⁴ From a systems theory perspective, this can be approached as a complex network of ‘nodes’ (treaties, regimes, organizations) that are linked (through interactions, shared memberships, treaty conflicts, partnerships, resource and information flows), in order to produce clusters (regime complexes, treaty clusters), and particular network effects. see Rakhyun E. Kim, *Is Global Governance Fragmented, Polycentric, or Complex? The State of the Art of the Network Approach*, INT. STUD. REV., 8 (2019), <https://academic.oup.com/isr/advance-article/doi/10.1093/isr/viz052/5571549> (last visited Feb 16, 2020).

⁸⁵ This distinction partially follows (though departs from) an influential framework on the production of international law which distinguishes between ‘actors’, ‘instruments’, and ‘processes’. See ALAN BOYLE & CHRISTINE CHINKIN, THE MAKING OF INTERNATIONAL LAW (2007). See also ROMERA, *supra* note 53 at 34.

⁸⁶ The most widely accepted formulation of these criteria is generally held to be provided in the 1933 Montevideo Convention. CONVENTION ON RIGHTS AND DUTIES OF STATES ADOPTED BY THE SEVENTH INTERNATIONAL CONFERENCE OF AMERICAN STATES, (1933), <https://treaties.un.org/doc/Publication/UNTS/LON/Volume%20165/v165.pdf>. Article 1.

⁸⁷ Under international law, requirements are an international agreement that establishes them, a capability to act independently from member States, and their subjection to international law. JOSÉ E. ALVAREZ, INTERNATIONAL ORGANIZATIONS AS LAW-MAKERS 6 (2005).

⁸⁸ Morin et al., *supra* note 39.

⁸⁹ Peter M. Haas, *Introduction: Epistemic Communities and International Policy Coordination*, 46 INT. ORGAN. 1–35 (1992).

⁹⁰ E.g. CHARLI CARPENTER, “LOST” CAUSES, AGENDA VETTING IN GLOBAL ISSUE NETWORKS AND THE SHAPING OF HUMAN SECURITY (2014), <http://www.degruyter.com/viewbooktoc/product/487489> (last visited Apr 1, 2019).

⁹¹ Indeed, their rise has created what Slaughter has called a ‘Lego model’, wherein ‘governments can be taken apart, put together with corporations, foundations, NGOs, church groups, universities, or any number of social

3.4.1.2 Instruments

In aiming to shape and regulate international behaviour, these actors avail themselves of a range of instruments, from traditional ‘hard’ international law to many varieties of ‘soft law’.

As noted earlier, the traditional sources of international law are set out in Article 38(1) of the Statute of the International Court of Justice, and include international treaties, international custom, and the general principles of law recognized by civilized nations—and, subsidiary to this, judicial decisions and the teachings of the most highly qualified scholars.⁹² However, in recent decades, these traditional tools have increasingly become complemented by diverse forms of soft law. The term ‘soft law’ can have various definitions, depending on the context. From a legal perspective, it simply refers to ‘normative provisions contained in nonbinding texts’,⁹³ a definition that covers various soft rules that are included in distinct instruments from treaties, nonbinding or voluntary resolutions, standards, or codes of conduct.⁹⁴ From a social scientific perspective, ‘soft law’ is defined less with reference to the legal validity or status of the rule, and more with consideration to the (absence of hard) mechanisms used to bind states to commitments; in such a reading, a vaguely formulated treaty without provisions for adjudication or enforcement could be considered soft law.⁹⁵ Notably, in this context, soft law standards need not only be produced by states or formal international organisations, but can also be produced by non-state or private sector actors.⁹⁶ What matters is simply the production of “nonbinding rules or instruments that interpret or inform our understanding of binding legal rules or represent promises that in turn create expectations about future conduct”.⁹⁷

3.4.1.3 Developments: the fragmentation of international law, and the rise of regime complexity

Along with the elements (actors and instruments) of the global governance complex, it is also important to sketch key developments that have characterised the recent trajectory of the global governance complex, and which have affected the interrelations between these actors and various norms.

Scholars have noted various trends. Some have noted the gradual ‘judicialization’ of international relations, with the rise in prominence of various international courts and arbitral

actors in any number of different coalitions.’ Anne-Marie Slaughter, *Remarks, The Big Picture: Beyond Hot Spots & Crises in Our Interconnected World*, 1 PENN STATE J. LAW INT. AFF. 286–302, 294–295 (2012).

⁹² UN, *Statute of the International Court of Justice*, 33 UNTS 993 (1946), <https://www.refworld.org/docid/3deb4b9c0.html> (last visited Sep 1, 2020). Article 38(1). See also the discussion of the sources of international law in Section 2.3.1.

⁹³ COMMITMENT AND COMPLIANCE: THE ROLE OF NON-BINDING NORMS IN THE INTERNATIONAL LEGAL SYSTEM, 292 (Dinah L Shelton ed., 2003); As cited in: Anne van Aaken, *Is International Law Conducive To Preventing Looming Disasters?*, 7 GLOB. POLICY 81–96, 87 (2016).

⁹⁴ Aaken, *supra* note 93 at 87.

⁹⁵ *Id.* at 87.; See generally Kenneth W. Abbott & Duncan Snidal, *Hard and Soft Law in International Governance*, 54 INT. ORGAN. 421–456 (2000); Gregory Shaffer & Mark A. Pollack, *Hard and Soft Law*, in INTERDISCIPLINARY PERSPECTIVES ON INTERNATIONAL LAW AND INTERNATIONAL RELATIONS 197–222 (Jeffrey L. Dunoff & Mark A. Pollack eds., 2012), https://www.cambridge.org/core/product/identifier/CBO9781139107310A018/type/book_part (last visited Oct 3, 2018).

⁹⁶ Indeed, Burri calls the latter ‘supersoft law’, and argues it will be particularly prominent in regulating AI. Burri, *supra* note 23 at 105–107. See also Section 2.3.2.2 on AI ethics.

⁹⁷ Andrew T Guzman & Timothy L Meyer, *International Soft Law*, 2 J. LEG. ANAL. 59, 174 (2010).

bodies.⁹⁸ Others have discussed a move away from traditional consent-based international law and towards ‘non-consensual’ lawmaking mechanisms.⁹⁹ This has been characterised by some as a process of ‘legal stagnation’, resulting in a move towards more informal and multistakeholder models of governance.¹⁰⁰

However, driving and bracketing these developments, perhaps the most important trend over the past few decades has been the steady *fragmentation* of international law. Although the term has different meanings, it has been considered as ‘the increased specialisation and diversification in international institutions, including the overlap of substantive rules and jurisdiction’.¹⁰¹ Of course, international law has never been a unified system. Yet the trend was highlighted by an influential 2006 International Law Commission report on *Fragmentation of international law: Difficulties arising from the diversification and expansion of international law*.¹⁰² In the absence of an international legislative supreme body to coordinate amongst regimes, the gradual development of specialized regimes in reaction to different challenges has had the effect that the global governance complex has become increasingly fragmented substantively (both along issue areas and geographically), normatively and institutionally.¹⁰³

As a result, recent years have seen considerable attention to the functional outcomes and effects of the normative and institutional fragmentation of governance architectures.¹⁰⁴ With it, there has also been increasing focus on the dynamics and (potentially problematic) interactions between the distinct regimes that constitute regime complexes on various given issue areas.¹⁰⁵

3.4.2 Foundations and dynamics of international regimes

Given that the literature on international regimes, and especially on regime complexity, has been informed by distinct ontologies and epistemologies, it is important to characterize this project’s approach to these debates.

Specifically, in exploring the creation, design, and effectiveness of international institutions, the field of international law and international relations has long been characterised

⁹⁸ Cesare P.R. Romano, *Proliferation of International Judicial Bodies: The Pieces of the Puzzle*, 31 N. Y. UNIV. J. INT. LAW POLIT. 709 (1998).

⁹⁹ Nico Krisch, *The Decay of Consent: International Law in an Age of Global Public Goods*, 108 AM. J. INT. LAW 1–40 (2014).

¹⁰⁰ J. Pauwelyn, R. A. Wessel & J. Wouters, *When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking*, 25 EUR. J. INT. LAW 733–763 (2014).

¹⁰¹ HARRO VAN ASSELT, THE FRAGMENTATION OF GLOBAL CLIMATE GOVERNANCE: CONSEQUENCES AND MANAGEMENT OF REGIME INTERACTIONS 22 (2014).

¹⁰² MARTTI KOSKENNIEMI & STUDY GROUP OF THE INTERNATIONAL LAW COMMISSION, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law* (2006), http://legal.un.org/documentation/english/a_cn4_1682.pdf.

¹⁰³ See also Joost Pauwelyn, *Fragmentation of International Law*, MAX PLANCK ENCYCL. PUBLIC INT. LAW 13, 2 (2006). Note however that not all agree. From a systems theory perspective, Kim argues that “the general trend over a long span of time has been, contrary to the general wisdom, governance defragmentation.” Kim, *supra* note 84 at 11.

¹⁰⁴ Biermann et al., *supra* note 35; Fariborz Zelli & Harro van Asselt, *Introduction: The Institutional Fragmentation of Global Environmental Governance: Causes, Consequences, and Responses*, 13 GLOB. ENVIRON. POLIT. 1–13 (2013); ASSELT, *supra* note 101. And in the legal field, see Gerhard Hafner, *Pros and Cons Ensuing from Fragmentation of International Law*, 25 MICH. J. INT. LAW 849–863 (2004); KOSKENNIEMI AND STUDY GROUP OF THE INTERNATIONAL LAW COMMISSION, *supra* note 102; Eyal Benvenisti & George W Downs, *The Empire’s New Clothes: Political Economy and the Fragmentation of International Law*, 60 STANFORD LAW REV. 595–632 (2007)..

¹⁰⁵ Orsini, Morin, and Young, *supra* note 31 at 29. See generally the discussion in Chapter 6.

by a distinction between *rationalist* accounts and *constructivist* accounts.¹⁰⁶ This is a rich and long-running debate, to which a brief review cannot do proper justice. Nonetheless, generally speaking, rationalist analysis emphasize the importance of *interests* to international relations. They argue that international institutions or regimes are created in order to alleviate strategic cooperation problems (e.g. the production of collective goods, or the avoidance of collective bads) in an anarchic international system.¹⁰⁷ In a rationalist perspective, these functional and interest-based considerations determine both when a regime is viable, and also what kind of institutional design choices end up getting embedded in international organisations.¹⁰⁸

In contrast, constructivist work on international institutions and regimes has explored the importance of norms, defined as “standard[s] of appropriate behaviour for actors with a given identity”.¹⁰⁹ This work explores how norms can have effects on actor’s institutional choices, and particularly how this foregrounds “logics of action other than instrumental rationality”.¹¹⁰ Conversely, it also examines the reverse relationship, by looking at how various ‘norm entrepreneurs’ such as transnational issue networks or epistemic communities can create or shift norms, and by doing so alter (state) actors’ perceptions and interests.¹¹¹ Such work can illuminate various steps in the emergence and evolution of international norms,¹¹² and show how the possibility boundary of regimes can shift over time.¹¹³

While the ‘rationalist-constructivist’ divide should be kept in mind when studying regime complexes, the difference should not be exaggerated.¹¹⁴ It has long been clear that both interests and norms matter,¹¹⁵ and that the relative role of these factors often may be a question of various contextual conditions.¹¹⁶ Indeed, in some domains, scholars have productively integrated both lenses, using constructivist analyses to ‘deepen’ rationalist arguments.¹¹⁷ For instance, Caroline

¹⁰⁶ The scholarship is extensive, but for an illuminating case study, see Caroline Fehl, *Explaining the International Criminal Court: A ‘Practice Test’ for Rationalist and Constructivist Approaches*, 10 EUR. J. INT. RELAT. 357–394 (2004).

¹⁰⁷ *Id.* at 362–367.

¹⁰⁸ See Koremenos, Lipson, and Snidal, *supra* note 75; Barbara Koremenos, Charles Lipson & Duncan Snidal, *Rational Design: Looking Back to Move Forward*, 55 INT. ORGAN. 1051–1082 (2001).

¹⁰⁹ Martha Finnemore & Kathryn Sikkink, *International Norm Dynamics and Political Change*, 52 INT. ORGAN. 887–917, 891 (1998). Note, this understanding of norms is distinct from ‘norms’ in the legal sense.

¹¹⁰ Fehl, *supra* note 106 at 359.

¹¹¹ See also the recent analytical framework in CAROLINE FEHL & ELVIRA ROSERT, *It’s Complicated: A Conceptual Framework for Studying Relations and Interactions between International Norms* 25 (2020).

¹¹² For a good example of this, see Elvira Rosert, *Norm emergence as agenda diffusion: Failure and success in the regulation of cluster munitions*, 25 EUR. J. INT. RELAT. 1103–1131 (2019).

¹¹³ See for instance in the area of arms control regimes: HARALD MÜLLER & CARMEN WUNDERLICH, *NORM DYNAMICS IN MULTILATERAL ARMS CONTROL: INTERESTS, CONFLICTS, AND JUSTICE* (2013).

¹¹⁴ Indeed, as early as 2002, Alexander Wendt and James Fearon argued that a synthesis between the theories is possible, and that they should be seen more as methodological tools than incompatible ontologies. James Fearon & Alexander Wendt, *Rationalism v. Constructivism: A Skeptical View*, in HANDBOOK OF INTERNATIONAL RELATIONS 52–72 (2002), https://sk.sagepub.com/reference/hdbk_intlrelations/n3.xml (last visited Jun 12, 2020).

¹¹⁵ Or, as Galbreath and Sauerteig put it: “We have come far enough now in the rationalist–constructivist debate to know that interests and norms matter.” Galbreath and Sauerteig, *supra* note 37 at 89.

¹¹⁶ Jeffrey T. Checkel, *International Norms and Domestic Politics: Bridging the Rationalist—Constructivist Divide*, 3 EUR. J. INT. RELAT. 473–495 (1997).

¹¹⁷ For instance, in an early study of the establishment of the International Criminal Court (ICC), Caroline Fehl has argued that a rationalist approach is able to identify the original trade-off between either a weak court backed by the US, or a strong court without US support; however, she argues a complementary constructivist approach

Fehl has surveyed the historical evolution of arms control regimes, and identified changing patterns in institutional inequality which, she argues, support both rationalist (shared interest-based) and constructivist (norms-based) accounts of international regimes, though not realist (coercion-based) models.¹¹⁸ Moreover, her account highlights the importance of changes not only in norms, but also of changes in material and technological developments.¹¹⁹ As a result, she articulates an integrated model that combines rationalist and constructivist arguments with historical institutionalist insights about the path-dependency of institutions, regimes, and legal and governance systems.

The approach to AI governance advanced in this project will broadly follow this integrated perspective. It assumes institutional arrangements may be created purposefully (by states) in order to fulfil certain functions or shared interests (per a rationalist perspective), but that a given regime's viability, legitimacy and long-term stability rest not just on its ability to fulfil such functions, but also on it remaining in accordance with changing conceptions of a just or legitimate global political order (as emphasised by constructivists).¹²⁰ This accordingly provides distinct and complementary angles through which to explore whether, or under what conditions, governance regimes on specific AI issues might be- or become viable.¹²¹

3.4.3 Technology and society

Finally, along with considering the role of ideational (interest- or norms-based) factors in global affairs and governance, this project must grapple with the interrelation of these factors with material factors.¹²² In brief, this project eschews a pure ontology, but instead adopts a mixed one, that can reckon with the intersecting roles of all three factors in global technology regime architectures. This discussion relates to the interactions between technology and society, specifically over the degree to which technology drives societal changes, and, if so, the degree to which we can anticipate or control its effects. This finds its source in long-running debates over so-called 'technological determinism'.¹²³ These debates have at times been heated and unproductive, in part because there is in fact a very broad family of claims, all with considerably distinct implications or plausibility, which are too easily subsumed under the simple label of 'technological determinism'.¹²⁴

helps explain how persuasive lobbying activities of NGOs as norm entrepreneurs tipped the scales towards the latter option. Fehl, *supra* note 106 at 377–382.

¹¹⁸ Caroline Fehl, *Unequal power and the institutional design of global governance: the case of arms control*, 40 REV. INT. STUD. 505–531, 530 (2014).

¹¹⁹ *Id.* at 530. ("changes in the distribution of material capabilities and in prevailing moral discourses create conflicting pressures for deepening and lessening institutional inequality. How these countervailing pressures are resolved depends on the constraining and conditioning impact of past institutional choices.").

¹²⁰ *Id.* at 518–519.

¹²¹ This will be especially prominent in chapter 7.1.2 (on 'regime viability and governance foundations').

¹²² This also relates to distinct definitions of 'technology', a topic that is taken up in Section 4.1.1.

¹²³ See for instance Sally Wyatt, *Technological Determinism Is Dead; Long Live Technological Determinism*, in THE HANDBOOK OF SCIENCE AND TECHNOLOGY STUDIES 165–180 (Edward J. Hackett et al. eds., 2008).

¹²⁴ Allan Dafoe, *On Technological Determinism: A Typology, Scope Conditions, and a Mechanism*, 40 SCI. TECHNOL. HUM. VALUES 1047–1076, 6 (2015). (distinguishing between theories that hold that "(1) functional entities (artifacts, techniques, institutions, and systems) exert an effect on the world independent of human choice (*technical determinism*); (2) there is a broad sequence and tempo of scientific and technological advance (*technological trends*) that seems to follow an internal logic, making technological change seem autonomous; and

This project will not seek to resolve those long-standing debates. Analytically, it rejects what is often presented as ‘hard technological determinism’,¹²⁵ recognizing that the processes by which new technologies are created can certainly be affected and shaped by social factors. For instance, Drezner has noted that while technologies can drive considerable and far-reaching shifts in international relations, they are not an exogenous shock, but are shaped by political choices which are themselves framed not only by the nature of technology but also by broader norms and factors.¹²⁶

On the other hand, this analysis also does not find itself in ‘radical social constructivism’,¹²⁷ where technology becomes largely or entirely subsumed to social processes or human choice.¹²⁸ Indeed, one does not need to envision technology as some entirely autonomous, external or irresistible force to recognize situations in which sociotechnical systems can exert patterned effects on social orders. There has been extensive work exploring such patterns, including various ‘technological trajectories’,¹²⁹ ‘technological politics’,¹³⁰ or ‘technology-induced shifts in values’,¹³¹ amongst others.

Clearly, in practice, ideational and material factors both matter in complex ways. Indeed, rationalist and constructivist scholars of international regimes alike have recognised that structural factors can shape individual actor preferences or norms, setting the conditions for the creation and maintenance of certain regimes. These structural factors traditionally include the distribution of power (in rationalist work) or norms (in constructivist work), but they may also include the existing state of scientific knowledge or of technology.¹³² Finally, scholars exploring

(3) that people are insufficiently conscious of their technological choices (*technological somnambulism*) or have been co-opted (*the magnificent bribe*), such that the social order is becoming more machine-like over time”).

¹²⁵ Wyatt, *supra* note 123 at 173–174.

¹²⁶ Daniel W Drezner, *Technological change and international relations*, 33 INT. RELAT. 286–303, 291 (2019). It should be noted, though, that apparent strategic ‘freedom’ at the ‘micro’ or even ‘meso’ level need not rule out longer-term technological trends at the ‘macro’-level, as discussed by Dafoe, who posits an account of ‘military-economic adaptationism’ (as a special case of a general theory of sociotechnical selectionism) to reconcile and integrate these accounts. Dafoe, *supra* note 124 at 12–16.

¹²⁷ Dafoe, *supra* note 124 at 4.

¹²⁸ *Id.* at 23.

¹²⁹ Giovanni Dosi, *Technological paradigms and technological trajectories*, 11 RES. POLICY 147–162 (1982). (“the pattern of ‘normal’ problem solving activity (i.e. of ‘progress’) on the ground of a technological paradigm.”).

¹³⁰ Langdon Winner, *Do Artifacts Have Politics?*, 109 DAEDALUS 121–136 (1980). Though his famous examples might be apocryphal; see also Steve Woolgar & Geoff Cooper, *Do Artefacts Have Ambivalence? Moses’ Bridges, Winner’s Bridges and Other Urban Legends in S&TS*, 29 SOC. STUD. SCI. 433–449 (1999). Bernward Joerges, *Do Politics Have Artefacts?:* SOC. STUD. SCI. (2016), <https://journals.sagepub.com/doi/10.1177/030631299029003004> (last visited Apr 24, 2020).

¹³¹ For instance, Iain Morris has suggested that the technology of energy capture available to different societies throughout history primed and affected the prevailing value systems. IAN MORRIS ET AL., *FORAGERS, FARMERS, AND FOSSIL FUELS: HOW HUMAN VALUES EVOLVE* (Stephen Macedo ed., Updated ed. edition ed. 2015); Though for a critique, see also Alberto Bisin, *The Evolution of Value Systems: A Review Essay on Ian Morris’s Foragers, Farmers, and Fossil Fuels*, 55 J. ECON. LIT. 1122–1135 (2017). For a different version of this argument, exploring the relation of different forms of agriculture (specifically hoe agriculture vs. plow agriculture) with gender inequality, see also Alberto Alesina, Paola Giuliano & Nathan Nunn, *On the Origins of Gender Roles: Women and the Plough*, 128 Q. J. ECON. 469–530 (2013). See also an accessible discussion in Sarah Constantin, *Hoe Cultures: A Type of Non-Patriarchal Society*, OTIUM (2017), <https://srconstantin.wordpress.com/2017/09/13/hoe-cultures-a-type-of-non-patriarchal-society/> (last visited Jul 4, 2020).

¹³² For instance, Stein notes how the ‘security dilemma’ is not an intrinsic feature of global politics; rather it “presumes either that offensive weapons exist and are superior to defensive ones, or that weapons systems are not easily distinguishable.” Stein, *supra* note 33 at 320.

the international regulation of new weapons technologies have argued that while ‘artefactual’ or ‘architectural’ features are not determinative of whether a ban is achieved, they can certainly affect how ‘regulation-tolerant’ or -resistant the technology is.¹³³

As noted, this project does not seek to adopt a pure ontology, whether rationalist, constructivist, or material-contextual.¹³⁴ Rather, it takes an intermediate position, one which does not assume linear, unidirectional, or hard determinist effects of technology on human society or law, but which does make three relatively modest assumptions about their interactions.

In the first place, this project assumes *technological politics*, a soft form of technological determinism which holds that choices made and inscribed in sociotechnical systems can have pervasive and lasting socio-political effects.¹³⁵ This is not a controversial position; scholars from a wide range of fields have long held that ‘artefacts have politics’.¹³⁶ Moreover, within legal scholarship, Lawrence Lessig and others have famously explored the role of ‘code’ and ‘architecture’ as distinct regulatory modalities which can be deployed to shape human behaviour in service of certain normative, political or regulatory goals.¹³⁷ This assumption simply speaks to the idea that technologies can be deployed in a ‘regulatory’ fashion. This lens especially informs the discussion of ‘governance displacement’ within the governance disruption lens.

In the second place, a core assumption in this project is the notion of technological *unintended consequences*—the argument that, as Dafoe puts it, “[d]ue to the lack of foresight or concern by the designer, or the sheer unpredictability of complex sociotechnical processes, unintended consequences can arise that fundamentally shape social relations.”¹³⁸ The argument here is that many of the longer-term, second- or third-order consequences of various technological choices or capabilities can be *unforeseeable*. This holds across different scales of analysis: at the operational level, a range of technologies can be prone to cascading ‘normal accidents’.¹³⁹ At the governance level, this underlies the difficulty of appropriately predicting or forecasting the full range of eventual applications and sociotechnical impacts of a given technology in advance, even if one could predict the underlying technological trends. The possibility of unanticipated effects also underpins much of the discussion of governance disruption (i.e. the ways AI could inadvertently affect the doctrines, norms, practices or political scaffolding of global governance).

¹³³ Sean Watts, *Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War*, 91 INT. LAW STUD. 83 (2015); Crootof, *supra* note 20 at 115–116; Crootof, *supra* note 3 at 23. See also the discussion of the role of material factors in constituting the regulatory texture of a ‘governance target’ (in Section 4.3.1.), and the application of this framework to military AI (section 7.1.1.3).

¹³⁴ As Deudney notes, when considering the prospects of large-scale societal change stemming from rapid technological progress, “it is vital to employ hybrid or mixed ontologies and to eschew the quest for a pure, or nearly pure, ontology” Daniel Deudney, *Turbo Change: Accelerating Technological Disruption, Planetary Geopolitics, and Architectonic Metaphors*, 20 INT. STUD. REV. 223–231, 228 (2018); Compare also Dafoe, *supra* note 124 at 23.

¹³⁵ Dafoe, *supra* note 124 at 7.

¹³⁶ Winner, *supra* note 130.

¹³⁷ Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. LAW REV. 501 (1999); LAWRENCE LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE, VERSION 2.0 (2nd Revised ed. edition ed. 2006), <http://codenv2.cc/download+remix/Lessig-Codev2.pdf>.

¹³⁸ Dafoe, *supra* note 124 at 8.

¹³⁹ Matthijs M. Maas, *Regulating for “Normal AI Accidents”: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment*, in PROCEEDINGS OF THE 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 223–228 (2018), <https://doi.org/10.1145/3278721.3278766> (last visited Sep 8, 2020).

Finally, this project assumes *asymmetric governance rationale/target materiality*, a clunky term for the idea that a technology's material features or alleged exceptional qualities are asymmetrically relevant to governance, because while material characteristics may be less relevant to determining whether there is a need for new law (i.e. whether the technology creates a governance rationale), they may nonetheless play a role in determining *how* to tailor effective regulation (i.e. the texture of the technology as governance target).¹⁴⁰

3.5 Methods

As noted, this research has primarily involved desk research, focused on theoretical analyses that draws on a number of frameworks and conceptual models, and deductively applies these to the study of AI governance. There is a limit to this, which is that the use of desk research, rather than qualitative interviews of policymakers or practitioners in the field, means I cannot draw on key tacit knowledge, inside institutional insights, or an understanding of the most current institutional decisions in the rapidly evolving AI governance landscape. To a degree, this might be a feature of studying many technology landscapes. However, I have tried to limit the risks of getting stuck into 'echo chambers', in part by pursuing a paper-based dissertation, allowing me to present my research regularly at a range of international conferences, to ensure active discussion and feedback on my various claims and arguments.

3.6 Conclusion: Methodology, Theory, Assumptions

This chapter has set out the methodological approach, theoretical background, and epistemological and ontological assumptions. I argued that this project takes an eclectic approach to introducing and applying three distinct lenses (sociotechnical change; governance disruption, and regime complexity), and I discussed and justified the selection of these three lenses.

I discussed the project's *epistemic approach*, in terms of its use of examples drawn from a range of regimes and technologies, and this project's orientation (anticipatory, not predictive) towards future developments. I set out the *ontological assumptions*, regarding the nature and texture of the global governance architecture, and the role of both interests and norms in constituting the foundations of international regimes. I then I briefly defended three assumptions about the interaction of technology with human choice: (1) (*technological politics*) sociotechnical systems or choices can have inscribed political or regulatory consequences; (2) (*unintended consequences*) we cannot anticipate all of the choices or (sociotechnical) effects of a technology in advance; (3) (*asymmetric governance rationale/target materiality*) a technology's material features may not be highly importance at determining whether there is a governance rationale, but will be relevant in tailoring the governance response. Finally, I concluded with reflections on the methods.

This Chapter concludes Part I ('Foundations') of this work. With these underlying concepts, we now turn, at last, to the first of three core framework chapters.

¹⁴⁰ This argument is articulated in much greater detail in the next chapter, in Sections 4.2-4.3.

Part II. FACETS OF CHANGE

Chapter 4. Sociotechnical Change: AI as Governance Rationale and Target

The sociotechnical change lens help AI governance initiatives in reconsidering societal changes produced or enabled by AI capabilities as a rationale and as target for regulation. After providing a context for these debates, I (4.1) provide a review and definition of the concept of ‘technology’, and explore how this relates to the theory of sociotechnical change. I then explore the two aspects of this concept. I (4.2) discuss sociotechnical change as governance rationale, by discussing different types of societal change that can be produced by a technology, the distinct grounds on which these can give rise to a need for new regulation under domestic or international law, and the epistemic limits of attempting to forecast or predict patterns of sociotechnical change. I then (4.3) discuss sociotechnical change as governance target, considering the role of both material factors, as well as ‘problem logics’. To chart these logics, I (4.4) articulate a taxonomy of 6 types of problems (ethical challenges; security threats; safety risks; structural shifts; common benefits; governance disruption). Finally, I (4.5) review the analytical uses and limits of this model for AI governance.

The first facet of *change* in AI governance concerns the way governance responses and initiatives can relate to technological change as a rationale and target of regulation. When and how do changes in AI translate into changes in (global) society? When do these create a need for new laws, regulation, or governance? How could or should these approach the technology or its resulting change? This provides one key lens on patterns of change which AI regimes should reckon with (see Figure 4.1).

This argument draws on existing scholarship on law, technology and regulation (‘TechLaw’), which has theorised these dynamics, predominantly at the domestic law level, but also increasingly at the international level.¹ Critically, such a perspective is relatively underdeveloped in the existing work on AI ethics and AI governance, which often takes a relatively reactive and isolated approach to problems thrown up by distinct AI applications. Of course, the space of potential issues, values, interests and goals to which AI systems could pertain is vast, and it should be no surprise that there have been many ways in which scholars have ordered or organised AI’s challenges.

For instance, in a philosophical taxonomy of the ‘ethics of AI and robotics’, Vincent Müller distinguishes between ethical issues that arise around AI systems as ‘*objects*’ (such as privacy, manipulation, opacity, bias, human-robot interaction, employment issues, and the effects of autonomy), versus issues that arise around AI systems as ‘*subjects*’ (covering machine ethics, artificial moral agency, and the implications of potential future advanced AI).²

¹ See for instance the recent exploration in Rebecca Crootof & B. J. Ard, *Structuring Techlaw*, 34 HARV. J. LAW TECHNOL. (2021), <https://papers.ssrn.com/abstract=3664124> (last visited Aug 28, 2020); Hin-Yan Liu et al., *Artificial intelligence and legal disruption: a new model for analysis*, 0 LAW INNOV. TECHNOL. 1–54 (2020).

² Vincent C. Müller, *Ethics of AI and robotics*, in STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2020), <https://plato.stanford.edu/entries/ethics-ai/#Sing> (last visited Apr 2, 2019). In another framework, John

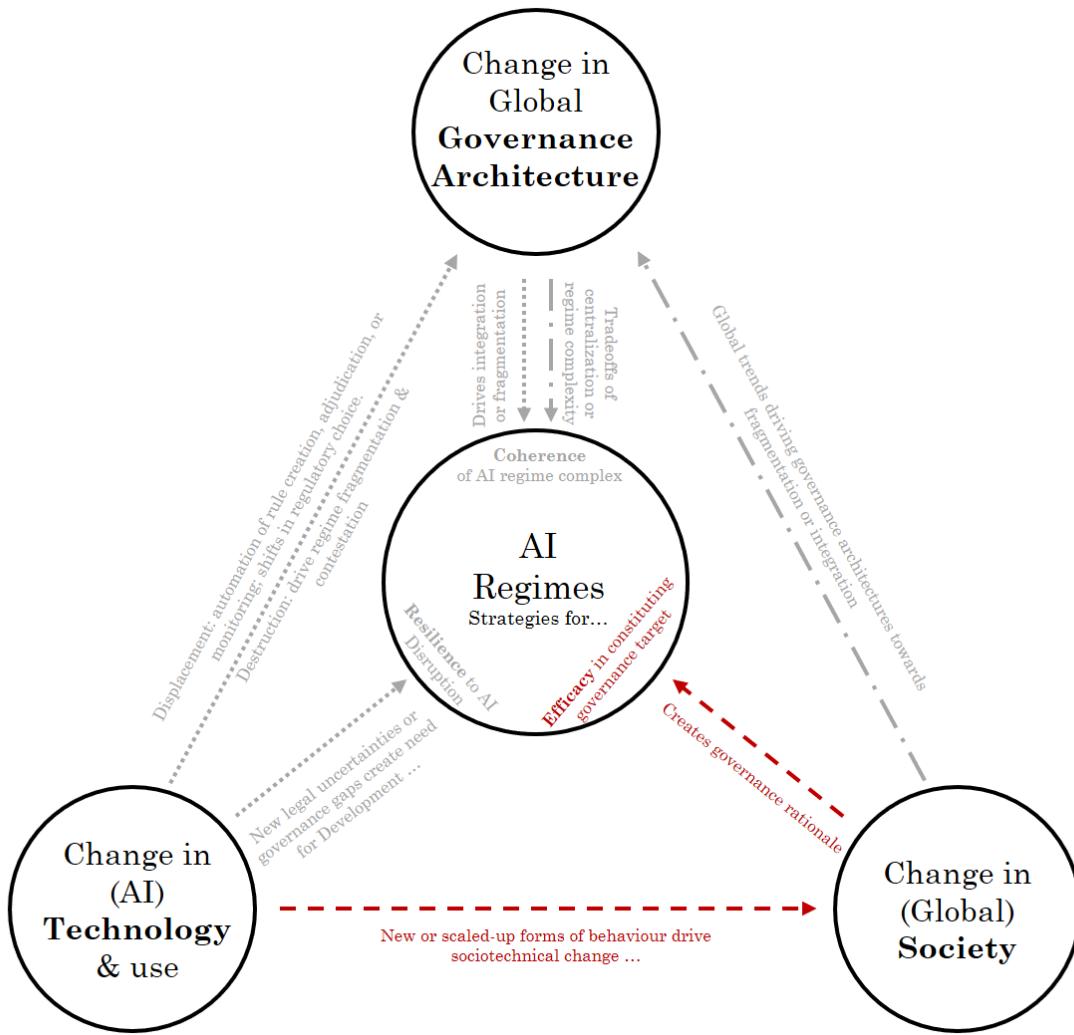


Figure 4.1. Conceptual sketch: Sociotechnical Change

Both ‘objects’- and ‘subject’-focused questions have received extensive treatment within philosophy, as well as in some legal scholarship.³ However, most recent writing on AI law and

Tasioulas characterizes ethical problems of robots and AIs (RAIs) as deriving from “functionality, inherent significance, rights and responsibilities, side-effects, and threats”. John Tasioulas, *First Steps Towards an Ethics of Robots and Artificial Intelligence*, 7 J. PRACT. ETHICS 61–95, 61 (2019). He also distinguishes between three levels of ethical analysis; (1) questions around whether AI systems need regulation (and if so, whether we ought to fashion specific laws or rely on general legal principles); (2) questions of social morality; (3) questions that arise for individuals or associations in their engagement with RAIs. *Id.* at 65.

³ See for instance, amongst many in a diverse literature; S. M. Solaiman, *Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy*, 25 ARTIF. INTELL. LAW 155–179 (2017); Joanna J. Bryson, Mihailis E. Diamantis & Thomas D. Grant, *Of, for, and by the people: the legal lacuna of synthetic persons*, 25 ARTIF. INTELL. LAW 273–291 (2017); John Danaher, *Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism*, SCI. ENG. ETHICS (2019); JACOB TURNER, *ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE* (2018); DAVID J. GUNKEL, *ROBOT RIGHTS* (2018).

policy has predominantly focused on AI systems as ‘objects’.⁴ Accordingly, Ryan Calo has distinguished AI policy problems as revolving around the themes of ‘Justice and Equity’; ‘Use of Force’; ‘Safety and Certification’; ‘Privacy and Power’; ‘Taxation and Displacement of Labor’, and a residual set of ‘Cross-Cutting Questions’.⁵ At the international level, challenges have been predominantly bucketed by pre-existing international legal domains, or thematically, distinguishing between challenges around global politics, international security, and international political economy.⁶

These are all key questions, and many of the policy responses articulated through these frameworks are valuable and insightful. Nonetheless, some may also fall short. In many of these cases, the regulatory or governance analyses and responses remain relatively siloed. They are focused either on particular AI *applications*,⁷ or segmented on the basis of AI’s *impact on specific conventional legal subjects* or approaches.⁸ Moreover, a common thread in these debates is the implicit dependence on ‘exceptionalism’—a debate over whether or not an AI application is sufficiently new and unprecedented in its ‘essential characteristics’, that it requires new laws.⁹ Yet, as noted by Crootof and Ard;

“... both the exceptionalist approach and its inverse are flawed. Not only do they place too much import on the architecture of a particular technology, they also foster narrow analyses.

⁴ One interesting exception to this, straddling both paradigms, is Nick Bostrom, Allan Dafoe & Carrick Flynn, *Public Policy and Superintelligent AI: A Vector Field Approach*, in ETHICS OF ARTIFICIAL INTELLIGENCE (S.M. Liao ed., 2019), <http://www.nickbostrom.com/papers/aipolicy.pdf> (last visited May 13, 2017). This deals both with extensive challenges to, and opportunities for human society (including desiderata of ‘magnanimity’), but also to considerations of ‘mind crime’ against future digital entities.

⁵ Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 UC DAVIS LAW REV. 37 (2017). This taxonomy is also followed by Gasser et al. in their ‘layered model for AI governance’. Urs Gasser & Virgilio A.F. Almeida, *A Layered Model for AI Governance*, 21 IEEE INTERNET COMPUT. 58–62 (2017). Others have articulated overlapping but complementary typologies; Guihot et al. list ‘bias’, ‘safety’, ‘legal decision-making’, ‘privacy’ and ‘unemployment’. Michael Guihot, Anne F. Matthew & Nicolas Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, VANDERBILT J. ENTERTAIN. TECHNOL. LAW (2017), <https://papers.ssrn.com/abstract=3017004> (last visited Jul 2, 2018). Walz & Firth-Butterfield distinguish “likely job losses, causation of damages (liability), lack of transparency, increasing loss of humanity in social relationships, loss of privacy and personal autonomy, potential information biases and the error proneness, and susceptibility to manipulation of AI and autonomous systems, [...] surveillance.” Axel Walz & Kay Firth-Butterfield, *Implementing Ethics into Artificial Intelligence: A Contribution, From a Legal Perspective, To The Development of an AI Governance Regime*, 18 DUKE LAW TECHNOL. REV. 166–231 (2019). Finally, a recent study by the European Parliamentary Research Service distinguishes between ‘Impact on society’ (labour market, inequality, privacy, human rights and dignity, bias, democracy); ‘impact on human psychology’ (relationships, personhood); ‘impact on the financial system’, ‘impact on the legal system’ (criminal law, tort law), ‘impact on the environment and the planet’ (use of natural resources, pollution and waste, energy concerns, ways AI could help), and ‘impact on trust’. ELEANOR BIRD ET AL., *The ethics of artificial intelligence: Issues and initiatives* (2020), [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU\(2020\)634452_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf) (last visited Jun 16, 2020).

⁶ MARY L. CUMMINGS ET AL., ARTIFICIAL INTELLIGENCE AND INTERNATIONAL AFFAIRS: DISRUPTION ANTICIPATED (2018), <https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>.

⁷ TURNER, *supra* note 3 at 218–219 (reflecting on UK governmental policies on driverless cars and drones in 2015–2018, which did not mention any overlap in possible underlying technology).

⁸ Petit calls these the ‘technological’ and ‘legalistic’ approaches NICOLAS PETIT, *Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications* 2 (2017), <https://papers.ssrn.com/abstract=2931339> (last visited May 11, 2020). See also Crootof and Ard, *supra* note 1 at 3–4.

⁹ Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. LAW REV. 513–564, 549 (2015).

At best, a compartmentalized assessment is a missed opportunity. At worst, it leads to ineffective, counterproductive, or even harmful rules and policy prescriptions.”¹⁰

This is unfortunate, because there are many common themes that could be articulated between AI ‘techlaw’ questions in different legal domains or different applications.¹¹ To understand what a better approach like this could look like, we will need to explore more deeply what type of change is of concern for governance.

4.1 Technology and Change

Both scholars and practitioners of international law regularly face a complicated relation to new technology. In the face of any technological change, they face four questions: are new regulations needed?¹² *Why* are they needed? *When* should we regulate? *How* should we regulate?

To understand the uses, strengths and limits of an analytical shift towards ‘sociotechnical change’ for AI governance, we should explore foundational questions in technology governance: what is ‘technology’? What is ‘technological change’? What is the relation of technological change with societal change?

4.1.1 A definition of Technology

The first step is to come to terms with the meaning of the term itself. What is ‘technology’? For all that the term itself is relatively modern,¹³ it has already seen extremely diverse usage. As noted by Lyria Bennett Moses, the word ‘technology’ has historically been used to refer to:

“(1) tools and techniques; (2) organized systems such as factories; (3) applied science; (4) those methods that achieve, or are intended to achieve, a particular goal such as efficiency, the satisfaction of human needs and wants, or control over the environment; and (5) the study of or knowledge about such things.”¹⁴

Moreover, even today, debates around technology—and related debates around the causes and effects of technological change—are beset by definitional ambiguity and varied usage of the term. For instance, as noted by Allan Dafoe, “[t]echnology can refer to vast sociotechnical systems, such as the Internet, as well as specific artifacts, standards, routines, and beliefs that make up

¹⁰ Crootof and Ard, *supra* note 1 at 4.

¹¹ *Id.* at 11. (“For example, there are obviously different concerns associated with having human beings in, on, or above the loop in content moderation, medicine, and warfare. But there is also much to be gained by considering the shared application, normative, and institutional uncertainties that arise from where a human being is situated in the decisionmaking process in all three scenarios”).

¹² Kristen Eichensehr has called this the “international law step zero” question”; Kristen E Eichensehr, *Cyberwar & International Law Step Zero*, 50 TEX. INT. LAW J. 24, 358 (2015). See also Rebecca Crootof, *Regulating New Weapons Technology*, in THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 1–25, 3–4 (Eric Talbot Jensen & Ronald T.P. Alcala eds., 2019).

¹³ Eric Schatzberg, “Technik” Comes to America: Changing Meanings of “Technology” before 1930, 47 TECHNOL. CULT. 486–512 (2006). See also Leo Marx, “Technology”: The Emergence of a Hazardous Concept, 64 SOC. RES. CAMDEN N J 965–988 (1997).

¹⁴ Lyria Bennett Moses, *Why Have a Theory of Law and Technological Change?*, 8 MINN. J. LAW SCI. TECHNOL. 589-606., 591 (2007).

these systems, such as computers, the Internet protocol, e-mail routines, and beliefs about the reliability of online information”.¹⁵ This drives home how ‘technology’ remains an extremely broad term, extending to profoundly varied and distinct artefacts, entities, practices or processes.

Given this breadth, it could be argued that ‘technology’ is not an analytically relevant or useful concept, but rather a suitcase word that attempts to wrap together some deeply dissimilar phenomena under one moniker. Leo Marx has argued as much, suggesting that the term ‘technology’, by conflating specific artifacts with extremely broad or abstract sociotechnical systems, “is almost completely vacuous”, and ends up inducing erroneous deterministic thinking.¹⁶ There is something to say for such caution. Still, Dafoe argues that in many contexts, such ambiguity can be avoided through the use of more specific terms such as ‘artifact’, ‘technique’, ‘institution’ or ‘sociotechnical system’.¹⁷ Disambiguating these terms that are normally subsumed under ‘technology’ is valuable, as it highlights the diverse factors and components that are involved in making a certain local technological artifact actually functional and viable (rather than merely ensuring it ‘works’), by situating it in a broader sociotechnical system.

Nonetheless, with Dafoe, I argue that there is still analytical value to the umbrella term of ‘technology’. This is because this term enables us to zoom out from the local particulars, specific materiality, or surface details of individual sociotechnical entities—the operational or material details of this or that institution, technique, or machine—and instead draws attention to these entities’ defining characteristic: their *functionality*.¹⁸ As such, in the context of governance, we would do well to adopt a broad and functional perspective on technology. What would such a definition look like? We can take, for instance, the theoretical working definition provided by Brownsword, Scotford and Yeung, who suggest ‘technology’ extends to;

“...those entities and processes, both material and immaterial, which are created by the application of mental and/or physical effort in order to achieve some value or evolution in the state of relevant behaviour or practice. Hence, technology is taken to include tools, machines, products, or processes that may be used to solve real-world problems or to improve the status quo”.¹⁹

¹⁵ Allan Dafoe, *On Technological Determinism: A Typology, Scope Conditions, and a Mechanism*, 40 SCI. TECHNOL. HUM. VALUES 1047–1076, 5 (2015).

¹⁶ Marx, *supra* note 13 at 982–983. As quoted in Dafoe, *supra* note 15 at 5.

¹⁷ Dafoe, *supra* note 15 at 5 (“Artifact can stand for specific objects intended for a function, such as machines, devices, and tools. Technique can refer to “softer” functional configurations, such as habits of mind, analytical methods, and behavioral routines. Institutions can refer to organizational hierarchies, legal codes, and incentive structures. Sociotechnical systems can refer to the vast functional configurations of all these components.”).

¹⁸ *Id.* at 5. (“[t]echnology thus, [1] denotes those entities—artifacts, techniques, institutions, systems—that are or were functional and [2] emphasizes the functional dimension of those entities.”). For a broader philosophical discussion of the ‘functionality’ of artefacts from an etiological perspective, see also A. W. Eaton, *Artifacts and Their Functions*, in THE OXFORD HANDBOOK OF HISTORY AND MATERIAL CULTURE (Ivan Gaskell & Sarah Anne Carter eds., 2020).

¹⁹ Roger Brownsword, Eloise Scotford & Karen Yeung, *Law, Regulation, and Technology: The Field, Frame, and Focal Questions*, 1 in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY, 6 (Roger Brownsword, Eloise Scotford, & Karen Yeung eds., 2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-1> (last visited Jan 3, 2019).

This definition is broad, but promising, because it emphasises the functional ‘output’ of a given sociotechnical system (its effects upon real-world situations; an evolution in the state of relevant behaviour) rather than merely background changes in practice or knowledge that, for one or another reason, are disregarded or not picked up widely.

However, while this definition can provide a starting point, it too is not flawless, as it emphasises deliberate attempts to ‘solve problems’ or ‘improve the status quo’—an essentially normative judgment which may often become more debatable or subjective, and nowhere more so than on the anarchic international stage.

As such, this project draws back on Donald Schön’s 1967 definition of “technology” as “any tool or technique, any product or process, any physical equipment or method of doing or making, by which human capability is extended”.²⁰ This notion emphasises the key point of changes in *capability* and functioning produced by technology, or by broad sociotechnical systems. However, it does exclude from consideration changes in systems that have sometimes been described as ‘technologies’, such as the ‘social technologies’ of legal systems, social norms, or (the relative costs of inputs on) markets.²¹

4.1.2 Change in Technology and Change from Technology

The next step is to articulate the relation between ‘technology’ and ‘change’. There are at least two ways to approach this relation.

In the first place, one can examine *changes in technology*. Here, technological change is a dependent variable, and the aim is to assess what factors affect when, why and how technological progress occurs. This question of what produces, constrains, or directs change in technology is a valuable one, and, in the case of AI, remains still relatively under-studied.²² As noted previously, this pertains to long-standing debates over the drivers of innovation, focusing on the relative roles of intrinsic ‘material’ factors, human agency, norms or political choices, or structural strategic pressures, in determining which technologies are invented, in which sequence, and whether or how fast they spread.²³ In many techlaw debates, this can also include debates over the role of

²⁰ DONALD A. SCHÖN, TECHNOLOGY AND CHANGE: THE NEW HERACLITUS 1 (1967). In doing so, it follows the definition also utilised by Lyria Bennett Moses at various points. See for instance Bennett Moses, *supra* note 14 at 591–592; Also see Lyria Bennett Moses, *Recurring Dilemmas: The Law’s Race to Keep Up With Technological Change*, 21 UNIV. NEW SOUTH WALES FAC. LAW RES. SER. (2007), <http://www.austlii.edu.au/journals/UNSWLRS/2007/21.html> (last visited Jul 3, 2018). Note, one might object that this definition of ‘technology’ is anthropocentric because it refers to *human* capability as its benchmark, and that technology would not cease to be such if it expanded the capabilities of animals, or indeed of robots or AI systems. However, such extensions risk confusion. Particularly in the latter case, for instance, it is unclear if we should speak of a robot having its capability extended by (other) technologies, or variably that this ought to be considered simply a case of one technology integrating with another (“getting an upgrade”, or ‘interfacing with a new API or module’). As such, in this project I bracket these debates and will retain the anthropocentric definition.

²¹ Bennett Moses, *supra* note 14 at 592 (“This essay concentrates on technology as being that which overcomes physical, as opposed to legal, normative or self-imposed constraint.”). See also Crootof, *supra* note 12 at 9.

²² Excepting ALLAN DAFOE, *AI Governance: A Research Agenda* 52 15–33 (2018), <https://www.fhi.ox.ac.uk/govaiagenda/>; Ross Gruetzmacher et al., *Forecasting AI Progress: A Research Agenda*, ARXIV200801848 CS (2020), <http://arxiv.org/abs/2008.01848> (last visited Aug 11, 2020).

²³ See also the discussion on ‘technology and society’ in section 3.4.3.

law and legal systems in shaping the economic incentives, social norms or regulatory environment that affected the developmental pathway of the technology.²⁴

While this question is a critical one from a legal angle, within this chapter and framework, the emphasis is on the second face of ‘technological change’—as *change from technology*. This aspect examines technology not as an effect, but rather as one source of societal changes that are relevant to governance.

4.1.3 The concept of sociotechnical change

This functional account of technology and technological change also underpins Lyria Bennett Moses’s ‘Theory of Law and Technological Change’.²⁵ Bennett Moses notes that questions of ‘law and technology’ are rarely if ever directly about technological progress itself (whether incremental or far-reaching). Instead, she argues that lawyers and legal scholars are primarily focused on questions of “how the law ought to relate to activities, entities, and relationships made possible by a new technology.”²⁶ This brings the focus on patterns of ‘socio-technical change’²⁷—that is, instances where changes in certain technologies ultimately translate to new ways of carrying out old conduct, or create entirely new forms of conduct, entities, or new ways of being or connecting to others.

As such, the question of governing new technologies is articulated not with reference to a pre-set list of technologies,²⁸ but is itself relatively ‘technology-neutral’. It is this functional understanding of ‘socio-technological change’ that potentially informs more fruitful analysis of when and why we require regulation for new technological or scientific progress. It can also underlie a more systematic examination of which developments in AI technology are relevant for the purposes of law and governance. In particular, such a perspective can contribute insights to all three dimensions of technological ‘disruption’ highlighted by Brownsword, Scotford, and Yeung—“legal disruption, regulatory disruption, and the challenge of constructing regulatory environments that are fit for purpose in light of technological disruption.”²⁹

However, a number of ambiguities remain: specifically, what types of sociotechnical changes (e.g. new possible behaviours or states of being) actually give rise to a rationale for a governance response? When a technology creates a new opportunity for behaviour, is it the mere possibility that constitutes a rationale for regulation? Or is it the actuality of such behaviour? What happens if the sociotechnical change produced by a technology is not (just) directed at society, but also at law or legal process directly? Can sociotechnical changes be anticipated? We will require a more granular understanding of the dynamics of sociotechnical change, exploring

²⁴ Crootof and Ard, *supra* note 1 at 19.

²⁵ Bennett Moses, *supra* note 14.

²⁶ *Id.* at 591.

²⁷ *Id.* at 591–592.; Bennett Moses, *supra* note 20.

²⁸ Lyria Bennett Moses, *Regulating in the Face of Sociotechnical Change*, in THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 573–596, 576 (Roger Brownsword, Eloise Scotford, & Karen Yeung eds., 2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49> (last visited May 13, 2017). (“This paper does not play the dangerous game of articulating a finite list of important technological fields that are likely to influence the future, likely to enable further technological innovation, or likely to pose the challenges to law and regulation as any such list would change over time.”).

²⁹ Brownsword, Scotford, and Yeung, *supra* note 19 at 7.

how, when and why such changes constitute a ‘rationale for governance’, and how they may then be approached as ‘targets for governance’.

4.2 Sociotechnical change as Governance Rationale

For the purposes of governing AI, it is key to understand the conditions under which specific advances or uses in AI give rise to sociotechnical changes of a sufficiently disruptive character that they create a governance *rationale*.

4.2.1 Thresholds: sociotechnical change and affordances

When and why do AI capabilities become a governance problem that warrants legal or regulatory solutions? It is important to recognize that not all new breakthroughs, capabilities or even use cases of a new technology will by definition produce the sort of ‘sociotechnical change’ that requires changes in-, or responses by governance systems. Indeed, Lyria Bennett-Moses has argued that, from a regulatory perspective, “[t]echnology is rarely the only ‘thing’ that is regulated and the presence of technology or even new technology alone does not justify a call for new regulation.”³⁰ In doing so, she calls for a shift in approach from ‘regulating technology’ to ‘adjusting law and regulation for sociotechnical change.’³¹

In practical terms, this relates to the observations that new social (and therefore governance) opportunities or challenges are not created by the mere fact of a technology being conceived, or even prototyped, but rather by them being translated into new ‘affordances’ for some actor.³² As per the above, such affordances can be new types of behaviour, entities, or relationships that were never previously possible, but which now in principle are available to actors.³³ Even at the international level, and even in the domain of security, not quite everything that can be made, is made: there are examples of powerful technologies that have been theorised or even demonstrated, but which were never seriously pursued or widely deployed, for various reasons.³⁴ As such, it is only if and when these affordances are both recognised and widely seized upon, that one might speak of sociotechnical change.

³⁰ Bennett Moses, *supra* note 28 at 575.

³¹ *Id.* at 574.

³² The original concept of ‘affordances’ derived from the work of ecological psychologist James G. Gibson, in the context of the relationship between an animal and its environment. James J. Gibson, *The Ecological Approach to the Visual Perception of Pictures*, 11 LEONARDO 227 (1978). However, in the context of technology, an ‘affordance’ is understood as “a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used.” DONALD A. NORMAN, *THE DESIGN OF EVERYDAY THINGS* 11 (Revised and expanded edition ed. 2013). This concept of affordances, and its link to ‘legal disruption’, are more fully explored in Liu et al., *supra* note 1.

³³ Bennett Moses, *supra* note 14 at 591: (“When lawyers claim to be interested in issues surrounding law and technology, it is usually related to questions of how the law ought to relate to activities, entities, and relationships made possible by a new technology. As technology changes, we can do things, make things, and form connections that were not previously practicable.”).

³⁴ See for instance, in a military context, the creative deployment configurations for strategic nuclear missiles surveyed in: OFFICE OF THE DEPUTY UNDER SECRETARY OF DEFENSE FOR RESEARCH AND ENGINEERING (STRATEGIC AND SPACE SYSTEMS), *ICBM Basing Options: A Summary of Major Studies to Define A Survivable Basing Concept for ICBMs* (1980), <http://www.dtic.mil/dtic/tr/fulltext/u2/a956443.pdf> (last visited Oct 3, 2018).

A similar functional argument has been made directly in the context of regulatory questions around AI and robotics. For instance, in the debate over ‘robolaw’, Balkin emphasised the need for examining the societal ‘salience’ of a new technology, rather than trying to discern any ‘essential features’.³⁵ This approach emphasizes the *usage* of a technology—not just the creation of a new affordance which makes a certain capability-space accessible, but the actual seizing of this new opportunity, by sufficient actors, to drive certain changes in society.³⁶

The insight is that for the purposes of law and governance, relevant technological changes are not *prima facie* about changes to ‘inputs’ or ‘core science’ (such as ‘fundamental theory’) alone, except insofar as these translate to (anticipated) changes to actors’ capabilities that would result in new societal changes. Technologies pose a challenge for societies (and therefore governance) not because they are intrinsically possible, nor even because they are discovered, but specifically because they expand the action-space for some or all actors, into directions that are harmful or of concern in some way or on some grounds,³⁷ or in directions that would be beneficial if and only if a certain degree of cooperation or coordination were assured.³⁸

4.2.2 Magnitude of ‘disruptive’ sociotechnical change

How does technological change translate into sociotechnical change? When would this be disruptive to law? The clearest category of ‘disruptive’ sociotechnical change may involve *absolute* and *categorical* changes: when progress ‘unlocks’ certain new capabilities or behaviour which

³⁵ Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. LAW REV. CIRCUIT 17 (2015). This approach stands in contrast to that of Ryan Calo, who seeks to distil ‘essential qualities’ of robotics as regulatory target; Calo, *supra* note 9.

³⁶ One caveat; it should be recognised that, especially in an international context, a technology can drive considerable change and political disruption even if it is not (yet) actually deployed, if it is (mistakenly) *perceived* to already be in (wide) usage. Consider for instance the US misperceptions, during the early Cold War, of a looming ‘Missile Gap’ with the Soviet Union. DANIEL ELLSBERG, THE DOOMSDAY MACHINE: CONFESSIONS OF A NUCLEAR WAR PLANNER (2017).

³⁷ See also Luciano Floridi, *Energy, Risks, and Metatechnology*, 24 PHILOS. TECHNOL. 89–94, 89 (2011). (arguing that “technologies lower constraints and expand affordances [and as such] continuously redesign the feasibility space of the agents who enjoy them”). For a macrostrategic exploration that takes such concepts to the limit, see also Nick Bostrom, *The Vulnerable World Hypothesis*, GLOB. POLICY 1758–5899.12718 (2019). (exploring the hypothesis that societies may sooner or later hit upon a technological capability that expands the feasibility-space into a particular direction that is so catastrophic and uncontrollable that the default outcome would be terminal global disaster), as well as the commentary and expansion by Manheim. David Manheim, *The Fragile World Hypothesis: Complexity, Fragility, and Systemic Existential Risk*, FUTURES (2020), <http://www.sciencedirect.com/science/article/pii/S0016328720300604> (last visited May 29, 2020). See in general the emerging literature on ‘global catastrophic and existential risks’. NICK BOSTROM & MILAN M. CIRKOVIC, GLOBAL CATASTROPHIC RISKS (1 edition ed. 2008); Nick Bostrom, *Existential Risk Prevention as a Global Priority*, 4 GLOB. POLICY 15–31 (2013); PHIL TORRES, MORALITY, FORESIGHT, AND HUMAN FLOURISHING: AN INTRODUCTION TO EXISTENTIAL RISKS (2017); TOBY ORD, THE PRECIPICE: EXISTENTIAL RISK AND THE FUTURE OF HUMANITY (2020). For interesting recent discussions of the geopolitical implications of such risks, see also Nathan Alexander Sears, *Existential Security: Towards a Security Framework for the Survival of Humanity*, 11 GLOB. POLICY 255–266 (2020); Nathan Alexander Sears, *International Politics in the Age of Existential Threats*, J. GLOB. SECUR. STUD. 1–23 (2020). For a discussion of a disaster risk reduction approach to such risks (emphasizing the role of systemic vulnerabilities or exposures, and not just hazards), see Hin-Yan Liu, Kristian Cedervall Lauta & Matthijs M. Maas, *Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research*, 102 FUTURES 6–19 (2018).

³⁸ This point also highlights how new technological change is of course not categorically or exclusively problematic. This also underlies the discussion, in 4.4.5, of ‘common benefits’.

were previously simply impossible for anyone, but which now enable behaviour that is harmful or of concern for one or more reasons.

More commonly, there might be situations with technological change driving more *relative* and ambiguous sociotechnical changes. In this way, new technologies can be disruptive not just in the absolute sense of unlocking unprecedented capabilities, but rather in a ‘relative’ sense, compared to a preceding status quo.³⁹ One version of this would involve situations where technological progress *lowers thresholds* or use preconditions for a certain capability (e.g. advanced video editing; online disinformation campaigns; cryptographic tools) which was previously already in reach for some actors, but which consequently becomes achievable for a much larger or much more diverse set of actors. Another version of this could involve developments that *scale up* certain existing capabilities.⁴⁰ Likewise, we can anticipate technologies driving *positional* changes. This might be seen clearly on the international level, where certain uses of AI could drive shifts in which particular actors (e.g. the US, EU or China) are dominant or exert the most power within the international system while leaving the ‘rules’ or dynamics of that system more or less unaltered.⁴¹

Disruption can also manifest through changes in the ‘nature’ or *structural dynamics* of that (international) society. For instance, AI could enable certain uses that shift the relative balance of power or influence between *types* of actors (e.g. away from states and towards non-state actors). Significantly, technologies do not need to be qualitatively novel in order to drive such structural shifts.⁴²

Finally, disruption can also manifest through shifts in the *means* by which certain actors seek to exercise ‘power’—shifting the emphasis from ‘hard’ military force to computational propaganda, or from multilateralism to increasing ‘lawfare’.⁴³ It could even alter the principal *terms* by which actors come to conceive of, measure and pursue their power or (national) interest (for example, data, global public opinion, oil reserves, population, soft power or territory).⁴⁴

³⁹ Matthijs M. Maas, *International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order*, 20 MELB. J. INT. LAW 29–56, 33 (2019) (reviewing various types of global disruptions that AI could create, in terms of positional changes, structural changes amongst types of actors, or changes to how actors conceive of, or pursue their various interests). Paper [III].

⁴⁰ See also Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV180207228 Cs (2018), <http://arxiv.org/abs/1802.07228> (last visited Feb 21, 2018).

⁴¹ For instance, see Alina Polyakova, *Weapons of the weak: Russia and AI-driven asymmetric warfare*, BROOKINGS (2018), <https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/> (last visited Jan 13, 2019). Elsewhere, Schneier and Farrell have argued that democracies, considered as information systems, are structurally more vulnerable to disinformation attacks than are autocracies. HENRY FARRELL & BRUCE SCHNEIER, *Common-Knowledge Attacks on Democracy* (2018), <https://papers.ssrn.com/abstract=3273111> (last visited Jan 13, 2019).

⁴² CUMMINGS ET AL., *supra* note 6 at iv (“[t]echnological change does not have to be dramatic or sudden to create meaningful shifts in power balances or social structures”).

⁴³ Charles Dunlap, *Lawfare Today: A Perspective*, YALE J. INT. AFF. 146–154, 146–148 (2008) (crediting new communications technologies and media with increasing the scope, velocity and effectiveness of “lawfare” efforts aimed at eroding public support for conflicts). Discussed in Crootof, *supra* note 12 at 6–7.

⁴⁴ Maas, *supra* note 39 at 32. Paper [III].

To be certain, any definition of ‘relative’ disruption is leaky, and contestable. Establishing an appropriate baseline against which to measure or mark whether an impact of AI has been ‘disruptive’ to existing societal dynamics remains a critical analytical question.⁴⁵

4.2.3 Governance rationales

More to the point, which sociotechnical changes are sufficiently disruptive that they constitute governance rationales? To be certain, many technological innovations do not clear this bar, even if from a scientific or engineering standpoint they involve considerable material or technological alterations to the state of the art, because they do not yield large sociotechnical changes.⁴⁶ In other cases, technologies do produce large sociotechnical changes, but these are not relevant per one of the regulatory criteria discussed later. Nonetheless, as the history of law and technology illustrates, there are many types of technological change that do pose new problems and so spur regulation. This turns on an examination of the general rationales for new regulatory interventions.

There are various accounts for when and why regulatory intervention is warranted by the introduction of new technologies. For instance, Beyleveld & Brownsword argue that emerging technologies generally give rise to two kinds of concerns: “one is that the application of a particular technology might present risks to human health and safety, or to the environment [...] and the other is that the technology might be applied in ways that are harmful to moral interests”.⁴⁷ However, while these may be the most prominent rationales, the full scope of reasons for regulation may extend further.

Tony Prosser has argued that, in general, regulation has four grounds: “(1) regulation for economic efficiency and market choice, (2) regulation to protect rights, (3) regulation for social solidarity, and (4) regulation as deliberation.”⁴⁸ As Lyria Bennett Moses notes, all four of these rationales may certainly become engaged by new technologies.⁴⁹ After all, technology can certainly create a site for *new market failures* which warrant regulatory intervention to ensure economic efficiency and market choice, such as in cases where there is a need for the articulation of technical standards to ensure interoperability between new products, or to remedy information inadequacies for consumers. Likewise, new technologies can generate many *new risks or harms*—

⁴⁵ Moreover, it may also be difficult to draw a clear line between disruption that is unambiguously ‘technological’ (directly driven by AI capabilities), and indirect disruption that appears ‘non-technological’ (for example, gradual changes to public values such as the appreciation of data privacy), or at most as a side effect of the use of AI. This of course pertains to the more general difficulty of pinning down a definition of ‘technology’, or drawing a clear line to distinguish ‘technology-driven change’ from more general societal change. *Id.* at 32.

⁴⁶ There is an additional distinction that could be drawn, between cases where the regulatory concern is over the novelty of a particular *process* (e.g. synthetic biology) involved in the technological capability, or whether it is about the novelty of the resulting *outputs* or *applications* (e.g. 3D printing). This may have implications for whether we consider the technology to be uncovered by existing regulation and give rise to a governance rationale. I thank Roger Brownsword for this observation.

⁴⁷ Deryck Beyleveld & Roger Brownsword, *Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning*, 4 LAW INNOV. TECHNOL. 35–65, 35 (2012).

⁴⁸ TONY PROSSER, THE REGULATORY ENTERPRISE: GOVERNMENT, REGULATION, AND LEGITIMACY 18 (2010), <https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199579839.001.0001/acprof-9780199579839> (last visited Jun 23, 2020). See also Bennett Moses, *supra* note 28 at 578.

⁴⁹ Bennett Moses, *supra* note 28 at 578.

either to human health or the environment, or to moral interests—which create a need for regulation to protect the rights of certain parties (e.g. restrictions of new weapons; the ban on human cloning).⁵⁰ Thirdly, new technologies can create concern about *social solidarity*, as seen in concerns over the ‘digital divide’ at both a national and international level, creating a need for regulation to ensure adequate inclusion. Finally, new technologies can create sites or pressures for the exertion of *proper democratic deliberation* over the design or development pathways of technologies.⁵¹

Yet again, as Bennett Moses notes, in each of these cases, it is not ‘technology’ per se that provides a special rationale for governance, above and beyond the basic social changes (e.g. potential market failures; risks to rights; threats to solidarity; or democratic deficits) that are at stake.⁵² The primary regulatory concern is over the emergence of the ‘sociotechnical’ effects that occur. This conceptual shift can help address one limit that regulatory or governance strategies encounter if they focus too much or too narrowly on technology. As she argues:

“...treating technology as the object of regulation can lead to undesirable technology specificity in the formulation of rules or regulatory regimes. If regulators ask how to regulate a specific technology, the result will be a regulatory regime targeting that particular technology. This can be inefficient because of the focus on a subset of a broader problem and the tendency towards obsolescence.”⁵³

As such, this lens helps keep into explicit focus specific *rationales* for governance in each use case of AI: on what grounds and when regulation is needed and justified?

These four accounts of rationales are valuable as a foundation. However, it may be analytically valuable to draw a more granular distinction between harms to human health, or harms to moral interests.⁵⁴ Finally, along with those five direct rationales, I suggest there may be one indirect category. That is, these first categories all concern rationales for governance to step in in response to sociotechnical changes affecting society (i.e. the regulatees) directly. However, there might also be cases where some sociotechnical change presents some risks directly to the existing legal order charged with mitigating the prior risks. In such cases, sociotechnical change can produce a threat to regulation, spurring an indirect governance rationale. In the context of AI, for instance, this governance rationale would focus in on the regulation of AI systems that are particularly disruptive to legal categories (in order to maintain coherence), or on the regulation of AI systems that are themselves used in or for regulatory purposes.⁵⁵

⁵⁰ *Id.* at 579–580.

⁵¹ *Id.* at 581–583.

⁵² *Id.* at 583.

⁵³ *Id.* at 584.

⁵⁴ Following Beyleveld and Brownsword, *supra* note 47.

⁵⁵ What, in Chapter 5, I call ‘governance displacement’ (and the sub-types of legal *automation* in rule-making, in compliance enforcement, and in the *substitution* of regulatory modalities). See Section 5.4. Note, there is hypothetically a third level, which focuses on the (recursive) regulation of hypothetical AI ‘safety-‘ or ‘audit’ systems which could be used to regulate (or check) AI systems deployed for primary regulatory purposes (e.g. adjudication). This would constitute the use of regulatory AI to regulate regulatory AI systems. I thank Roger Brownsword for this suggestion.

Drawing together the above accounts, one might then speak of a governance rationale for a certain new AI system or application whenever the use of these capabilities drives sociotechnical changes (new ways of carrying out old behaviour, or new behaviours, relations or entities) which result in one or more of the following situations: (1) possible market failures; (2) new risks to human health or safety; (3) new risks to moral interests, rights, or values; (4); new threats to social solidarity; (5) new threats to democratic process; (6) new threats directly to the coherence, efficacy or integrity of the existing governance system charged with mitigating these prior risks.

This is not to say that these rationales apply the same way in all contexts. Indeed, they may be weighted differently across distinct legal systems. For instance, all six of these challenges surely appear at the international level—diverse developments both technological and social can produce international market failures, risks to human health or the environment, challenges to rights, threats to (global) social solidarity, democratic process, or direct threats to the international legal order. However, depending on one's assumptions, that need not mean that all six rationales might be equally relevant or in scope for international norms. For instance, it has been argued that the state-centric international law system “is primarily preoccupied with finding the answers that are required to separate the powers of the sovereign states and to ensure the maintenance of order and stability.”⁵⁶ That means that traditional international law might be particularly attuned to AI-driven sociotechnical changes that result in risks to human life or challenges to human rights. Conversely, the international law system might be less intrinsically concerned with addressing international market failures.⁵⁷ Nonetheless, from a broader perspective of global governance—especially one that emphasizes the need for international legal systems to evolve in order to better address commons problems⁵⁸—all six rationales can and should be considered at the global level.⁵⁹ As such, while taking into consideration the different legal contexts, this taxonomy can still inform the potential grounds for AI governance globally.

4.2.4 The epistemic limits of anticipating sociotechnical change

The ‘sociotechnical change’ perspective also throws into stark relief the epistemic challenges around anticipating or identifying the full set of governance rationales in advance.

It may be comparatively easy to identify new AI capabilities demonstrated in a lab, or applications that are seeing deployment, and attempt to reason forward to various societal challenges. However, while this is necessary and legitimate, it can only take us so far. This is

⁵⁶ ANDERS HENRIKSEN, INTERNATIONAL LAW 17–18.

⁵⁷ That is not to say that it lacks regimes focused at addressing such potential international market failures—nothing could be further from the truth. However, the mere possibility of a market failure may not offer the same strong rationale for intervention under international law as would a violation of human rights.

⁵⁸ See for instance Nico Krisch, *The Decay of Consent: International Law in an Age of Global Public Goods*, 108 AM. J. INT. LAW 1–40 (2014); Anne van Aaken, *Is International Law Conducive To Preventing Looming Disasters?*, 7 GLOB. POLICY 81–96 (2016); J. Pauwelyn, R. A. Wessel & J. Wouters, *When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking*, 25 EUR. J. INT. LAW 733–763 (2014).

⁵⁹ This can be seen in various soft law instruments; see for instance the emphasis on crossing the digital divide by the UN Secretary-General’s High-Level Panel on Digital Cooperation. United Nations General Assembly, *Road map for digital cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation* 14 (2020), <https://undocs.org/pdf?symbol=en/A/74/821> (last visited Sep 1, 2020). HIGH-LEVEL PANEL ON DIGITAL COOPERATION, *Report of the Secretary-General: Roadmap for Digital Cooperation* 39 (2020), https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/Roadmap_for_Digital_Cooperation_EN.pdf.

because sociotechnical impacts can take longer to manifest, and will realize themselves in unexpected ways. Moreover, there may be contestation over the exact societal implications of a technology's deployment, especially where it concerns core ethical questions.

As such, there are considerable challenges to assessing, in a timely fashion and to a sufficient level of reliability, the relative effects or prominence of distinct types of sociotechnical change, and the governance rationales they might raise. In fact, there can often remain pervasive uncertainty over a technology's societal effects, even years after its initial development and deployment. Take the example of cyberwar. While some early accounts anticipated catastrophic terrorist cyber-attacks on state infrastructures with kinetic effects,⁶⁰ many scholars have since concluded that such analyses overstated the (kinetic) threat from 'cyberwar', as well as the lasting ability of small actors (the 'lone hacker') to carry out such attacks.⁶¹ Nonetheless, this view of cyberwar continues to dominate amongst some policymakers even today.⁶²

The same dynamics can be seen in debates around emerging AI applications, where there remains considerable uncertainties around the relative plausibility or imminence of various threats that are often proposed. Indeed, there can be surprising uncertainty even around some challenges that are widely perceived to be well established or underway. Consider debates around technological unemployment. While some scholars have warned of far-reaching economic and societal dislocation as an outcome of increasing technological unemployment,⁶³ it remains unclear to what extend such effects are likely in the near- to medium-term.⁶⁴

Likewise, while there is a growing body of scholarship on the challenges of 'computational propaganda',⁶⁵ there is disagreement over whether various incidents are the early warning shots of a development that will erode the very epistemic foundations of societies,⁶⁶ or whether these are high-profile but ultimately peripheral incidents. For instance, some scholars have suggested that public discourse may be overstating the current democratic threats from 'filter bubbles',⁶⁷

⁶⁰ John Arquilla & David Ronfeldt, *Cyberwar is Coming!*, ATHENAS CAMP PREP. CONFL. INF. AGE 1995–1996 (1992). Though for a more nuanced analysis, see John Stone, *Cyber War Will Take Place!*, 36 J. STRATEG. STUD. 101–108 (2013).

⁶¹ THOMAS RID, CYBER WAR WILL NOT TAKE PLACE (1 edition ed. 2013). As Horowitz also notes, "it is no longer true, and probably never was, that anyone with a computer can successfully launch a cyber attack against a state or sophisticated organization." Michael C. Horowitz, *Do Emerging Military Technologies Matter for International Politics?*, 23 ANNU. REV. POLIT. SCI. 385–400, 394 (2020).

⁶² Horowitz, *supra* note 61 at 393.

⁶³ Carl Benedikt Frey & Michael A. Osborne, *The future of employment: How susceptible are jobs to computerisation?*, 114 TECHNOL. FORECAST. SOC. CHANGE 254–280 (2017). See also JOHN DANAHER, AUTOMATION AND UTOPIA: HUMAN FLOURISHING IN A WORLD WITHOUT WORK 1 (2019).

⁶⁴ For instance, Scholl and Hanson have reviewed patterns in job vulnerability metrics between 1999 and 2019, and found "no evidence yet of a revolution in the patterns or quantity of automation", and where automation did occur, did not find this to predict changes in either pay or employment. Keller Scholl & Robin Hanson, *Testing the automation revolution hypothesis*, 193 ECON. LETT. 109287 (2020).

⁶⁵ Gillian Bolsover & Philip Howard, *Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda*, 5 BIG DATA 273–276 (2017); Bruce Schneier, *Bots Are Destroying Political Discourse As We Know It*, THE ATLANTIC, 2020, <https://www.theatlantic.com/technology/archive/2020/01/future-politics-bots-drowning-out-humans/604489/> (last visited Jan 15, 2020).

⁶⁶ Herbert Lin, *The existential threat from cyber-enabled information warfare*, 75 BULL. AT. SCI. 187–196 (2019).

⁶⁷ See for instance AXEL BRUNS, ARE FILTER BUBBLES REAL? (1 edition ed. 2019).

'fake news',⁶⁸ or 'bots' on social media.⁶⁹ More recently, while much coverage of 'DeepFakes' has touted their potential misuse for political manipulation,⁷⁰ the vast majority of DeepFakes today are instead still used for (gendered) harassment,⁷¹ a challenge that might be missed by regulatory initiatives focused on the spectre of political manipulation. Indeed, it may be that from the perspective of propaganda, these techniques do not, ultimately, offer much marginal benefit over already-existing 'cheapfakes' avenues to manipulate media content.⁷²

That is not to suggest that the political threat from DeepFakes has been dismissed, and we can rest easy.⁷³ Rather, it shows some of the complexities involved in reasoning from a demonstrated capability or affordance ('the ability to forge media') to an ecology of downstream uses. Rather, it highlights a set of complicating dynamics.

For one, there is no reason to expect the sociotechnical impacts of a given technology to increase *evenly in all sectors*. As such, while we might overstate some challenges, there is also a risk we will underestimate many others. For instance, it might be the case that the 'democracy-eroding' effects of DeepFakes will remain relatively modest or marginal, but that they will see widespread use in conducting various types of criminal fraud.⁷⁴

Moreover, the sociotechnical impacts of a given technology *may not increase linearly or monotonically*. It could be the case that the 'democracy-eroding' effects of DeepFakes technology

⁶⁸ For instance, contrary to widespread perception, computational propaganda likely did not play a large role in swaying the 2016 US election or Brexit vote. YOCHAI BENKLER, ROBERT FARIS & HAL ROBERTS, NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION, AND RADICALIZATION IN AMERICAN POLITICS (2018). Indeed, 'fake news' likely made up a small minority (6%) of all news consumed in that US election, although it did target small sub-communities intensely, which might produce political effects. Jacob L Nelson & Harsh Taneja, *The small, disloyal fake news audience: The role of audience availability in fake news consumption*, 20 NEW MEDIA SOC. 3720–3737 (2018).

⁶⁹ Likewise, while bots are certainly active on social media, some reviews have found that they "accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it." Soroush Vosoughi, Deb Roy & Sinan Aral, *The spread of true and false news online*, 359 SCIENCE 1146–1151 (2018); Brendan Nyhan, *Fake News and Bots May Be Worrisome, but Their Political Power Is Overblown*, THE NEW YORK TIMES, February 16, 2018, <https://www.nytimes.com/2018/02/13/upshot/fake-news-and-bots-may-be-worrisome-but-their-political-power-is-overblown.html> (last visited Apr 16, 2019). There are also considerable methodological challenges in studying online bots. See also Michael Kreil, *The Army That Never Existed: The Failure of Social Bots Research* (2019), <https://michaelkreil.github.io/openbots/> (last visited Aug 31, 2020). And also Siobhan Roberts, *Who's a Bot? Who's Not?*, THE NEW YORK TIMES, June 16, 2020, <https://www.nytimes.com/2020/06/16/science/social-media-bots-kazemi.html> (last visited Aug 31, 2020).

⁷⁰ Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. LAW REV. 1753–1820 (2019).

⁷¹ HENRY AJDER ET AL., *The State of DeepFakes: Landscape, Threats, and Impact* 27 (2019), <https://deeptracelabs.com/mapping-the-deepfake-landscape/>. See also Samantha Cole, *For the Love of God, Not Everything Is a Deepfake*, VICE (2020), https://www.vice.com/en_us/article/7kzgg9/joe-biden-tongue-gif-twitter-deepfake (last visited May 5, 2020).

⁷² BRITT PARIS & JOAN DONOVAN, *DeepFakes and Cheap Fakes: The Manipulation of Audio & Visual Evidence* 50 (2019), <https://datasociety.net/output/deepfakes-and-cheap-fakes/>.

⁷³ And it certainly does not suggest that this technology does not create pressing problems, as shown by the pervasive misuse in gendered harassment.

⁷⁴ M. Caldwell et al., *AI-enabled future crime*, 9 CRIME SCI. 14 (2020); Keith J Hayward & Matthijs M Maas, *Artificial intelligence and crime: A primer for criminologists*, CRIME MEDIA CULT. 1741659020917434 (2020).

are overstated for the near-term, but that they will be extensive in the longer term, once a certain threshold of performance is reached.⁷⁵

Furthermore, especially in cases involving a general-purpose technology, the *pathways of sociotechnical change* will be complex and unanticipated. It could be that DeepFakes technology will indeed have democracy-eroding effects, but in unexpected domains—not in the form of faux political smear videos, but rather, perhaps, through their effects on the evidential value of traditional categories of (video) evidence in a court of law.⁷⁶ Alternatively, it might be the generation of human-like text, rather than faked videos, that have the most far-reaching political impact. It could also prove the case that AI's role cannot be easily isolated within any one election, but that in aggregate, they do erode the 'epistemic backstop' of society.⁷⁷ The point is that we do not know, and may not actually know with reasonable certainty until a point in time surprisingly far along into the technology's development and deployment curve.

This difficulty complicates assessments of a technology's sociotechnical impacts, including which governance rationales, if any, it might raise. What does that suggest? Critically, this difficulty in accurately predicting a new technology's trajectory or societal impacts may be taken by some as an argument in favour of a reactive 'wait-and-see' regulatory approach, which simply argues we should only take action if and when problems do arise. Yet that may be the wrong lesson to draw. Indeed, as illustrated by the evolving 'nature' of cyberspace or the ongoing debates over cyberwarfare, attempting to 'wait out' the effects of technological developments—the often proposed 'sober' alternative to 'speculation'—need not provide the presumed level of clarity in time.

Indeed, the risk with a 'wait-and-see approach' is that for many technologies, there will not come a 'just right' moment of sufficient certainty to act, until it is far too late along into its development. Moreover, as argued by Rebecca Crootof, this reactive approach may have two drawbacks, in that it might foster "hypersensitive rulemaking" in response to unusual accidents; and that it foregoes a valuable opportunity to channel the direction of technological development.⁷⁸

This may suggest that rather than aiming at specific anticipated technologies, governance systems would do better to adopt a general assumption of further or ongoing change, and adopt policies or governance approaches that are resilient or adaptive across distinct scenarios. In preparing sets of policies that might be up to such tasks, we will need to take stock of how this technology might be approached as 'governance target', whichever governance rationales it creates. Moreover, given these situations of uncertainty, regulatory responses to a new technology

⁷⁵ For instance, once quality improves to such a level that effective detection becomes functionally impossible. Alex Engler, *Fighting deepfakes when detection fails*, BROOKINGS (2019), <https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/> (last visited Dec 11, 2019).

⁷⁶ Hin-Yan Liu & Andrew Mazibrada, *Artificial Intelligence Affordances: Deep-fakes as Exemplars of AI Challenges to Criminal Justice Systems*, UNICRI SPEC. COLLECT. ARTIF. INTELL. (2020), <http://unicri.it/towards-responsible-artificial-intelligence-innovation> (last visited Jul 13, 2020); See also Alexa Koenig, "*Half the Truth is Often a Great Lie*": Deep Fakes, Open Source Information, and International Criminal Law, 113 AJIL UNBOUND 250–255 (2019).

⁷⁷ Regina Rini, *Deepfakes and the Epistemic Backstop* (2019), <https://philpapers.org/archive/RINDAT.pdf>; Don Fallis, *The Epistemic Threat of Deepfakes*, PHILOS. TECHNOL. (2020), <https://doi.org/10.1007/s13347-020-00419-2> (last visited Aug 10, 2020).

⁷⁸ Crootof, *supra* note 12 at 21–22.

need not only be motivated by retrospectively confirmed impacts, but can still be grounded on concerns about the ‘envisioned purposes’ of a technology (e.g. ‘AI surveillance’), or the potential side-effects (e.g. ‘algorithmic discrimination’). Such a precautionary or anticipatory approach may be particularly key in the context of particularly high-stakes risks or scenarios.⁷⁹

4.3 Sociotechnical change as Governance Target

Along with providing a greater grounding for understanding whether, when and why to regulate new (AI) capabilities, a consideration of sociotechnical change can also shed light on the regulatory ‘texture’ or features of the underlying AI capabilities: its constitution as ‘governance target’.⁸⁰ This has two components: (4.3.1) a baseline consideration of the particular *material features*, and (4.3.2) a more conceptual consideration of the general *problem logic*.

4.3.1 Regulatory textures and the role of materiality

In one sense, insofar as the lens of sociotechnical change shifts the emphasis away from the ‘newness’ of certain technologies, in order to focus on disruptive sociotechnical changes (new behaviours, relations, or ways of being), it is an account that moderately de-emphasizes the artefactual or material dimension of technology governance. Nonetheless, once we have answered the question of whether certain sociotechnical changes create a *rationale* for governance, we face the question of how to design governance appropriately.

In answering this question—of how to approach certain technologies as a governance *target*—it can certainly be valuable to re-open considerations of the technology itself. For one, this is because ongoing developments or breakthroughs in the underlying scientific state-of-the-art can shape perceptions of imminent applications, which may have a large effect on both public debate as well as state calculations around global regimes, even if the technology is not fully realised.⁸¹

More importantly, the material ‘features’ of a given technology can be highly relevant for understanding *how* to regulate. They can clearly illuminate the relative purchase or efficacy of different regulatory strategies at controlling, regulating or mitigating the risks of concern. For instance, the rate at which the costs of computing hardware will fall in coming years, will affect the rate and speed of diffusion of military AI applications,⁸² as well as the ease with which such

⁷⁹ *Id.* at 23–25. There is of course extensive debate around the general role and usefulness of various formulations of the precautionary principle, as well as precautionary approaches around new technologies. This literature is too extensive, but see Beyleveld and Brownsword, *supra* note 47.

⁸⁰ Cf. the discussion of AI as ‘regulatory target’ in Miriam C. Buiten, *Towards Intelligent Regulation of Artificial Intelligence*, 10 EUR. J. RISK REGUL. 41–59, 46–48 (2019).

⁸¹ See also Picker’s discussion of international regimes pursued for putative or non-existing technologies. Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDOZO LAW REV. 151–219, 194–196 (2001).

⁸² Michael C. Horowitz, *Artificial Intelligence, International Competition, and the Balance of Power*, TEXAS NATIONAL SECURITY REVIEW, 2018, <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/> (last visited May 17, 2018).

proliferation might be halted through export control regimes.⁸³ More subtly, the trajectory and process of weapon design and procurement can be significantly affected by material factors.⁸⁴

Specifically, with regards to AI challenges, a consideration of technical features of specific *techniques* (e.g. algorithmic approaches),⁸⁵ may by itself not be sufficient in scoping the full range of sociotechnical changes, or in establishing a rationale for governance; however, it can certainly help in tailoring or focusing subsequent regulatory efforts. Such a detailed ‘engineers’ perspective helps in a number of ways: (1) it helps mitigate undue anthropomorphisation of AI; (2) it allows a more granular mappings of the state of the technology, rates of progress in different approaches,⁸⁶ and preconditions for its dissemination to various actors; (3) it helps map how different AI paradigms or techniques can be linked to distinct safety or ethics issues.⁸⁷ After all, while AI techniques are generally quite transferable across domains or contexts, they are of course very distinct in their workings. In a number of contexts, considering which particular contemporary AI approaches are used can help shed light on at least some of the risks which are or are not associated with those approaches. For instance, we can consider reinforcement learning algorithms’ propensity to raise safety problems involving reward hacking or interruptibility of the algorithm;⁸⁸ supervised and unsupervised learning algorithms’ susceptibility to reproducing data biases either directly or by proxy (when certain attributes are protected); or neural networks’ lack of explainability relative to far more interpretable AI approaches such as decision trees.⁸⁹ This is valuable in highlighting some of the heterogeneity of approaches, as well as in spurring consideration of the degree to which a problematic situation involving ‘AI’ can actually be addressed through technical interventions aimed at the AI system, or whether this is better served by broader institutional interventions.

Of course, the question can be asked: after all the emphasis on shifting governance focus away from ‘technology’ or its ‘essential features’, does this not simply smuggle ‘technology’ directly back front and center to legal analysis? Not necessarily. The point here is that while it is sociotechnical change which determines whether there is a governance *rationale*—i.e. what are (and whether there are) societal problems that need to be addressed—local material ‘features’ can

⁸³ Jade Leung, Sophie-Charlotte Fischer & Allan Dafoe, *Export Controls in the Age of AI*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/> (last visited Sep 2, 2019).

⁸⁴ Maaike Verbruggen, *In Defense of Technological Determinism* (2020).

⁸⁵ What Schuett calls ‘design’: Jonas Schuett, *A Legal Definition of AI*, ARXIV190901095 Cs, 4 (2019), <http://arxiv.org/abs/1909.01095> (last visited Jan 6, 2020).. See also the discussion around ways of (and purposes in) defining AI, in 2.1.1.

⁸⁶ For one insightful mapping, see also Francesco Corea, *AI Knowledge Map: how to classify AI technologies*, MEDIUM (2020), https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020 (last visited Jul 8, 2020).

⁸⁷ See for instance the mapping in Jose Hernandez-Orallo et al., *AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues* (2020), http://ecai2020.eu/papers/1364_paper.pdf. (distinguishing also between ‘artefacts’ and ‘techniques’).

⁸⁸ Dario Amodei et al., *Concrete Problems in AI Safety*, ARXIV160606565 Cs (2016), <http://arxiv.org/abs/1606.06565> (last visited May 13, 2017).

⁸⁹ Schuett, *supra* note 85 at 4. Note, Schuett calls these ‘inherent risks’. As per the discussion in the next paragraph, the debate over whether these are truly ‘inherent’ to various approaches may turn on a somewhat philosophical or semantic debate: given future progress, certain approaches that are today ‘opaque’ may become more explainable. Nonetheless, for the purposes of governance analysis in the near term, generalizing over the risks or features ‘inherent’ to certain approaches can be pragmatically functional ‘legal fictions’.

matter in crafting adequate governance solutions. However, that analysis of a given technology's 'architectural texture' as a means of assessing its regulatory surface *at that moment in time* does not amount to assuming that these features are essential or even necessarily very stable.⁹⁰

Rather, this argument admits that many features of a technology that appear critical or fundamental to it in one moment can themselves be altered by ongoing technological progress, or by shifting social, regulatory or even (geo)political contexts. After all, just as early 'features' of the internet (such as the assumed anonymity) faded or evolved over time, many material and operational features of AI technology may develop over time, in ways that overturn widely shared assumptions. For instance, in recent years, public debates on AI have often been informed by a presumption that many machine learning algorithms are intrinsically 'unexplainable',⁹¹ or that training AI systems requires massive datasets and computing power.⁹² There may certainly be some truth to such characterisations today, and for the coming years. Yet these are not essential features of all AI technology: advances in algorithmic explainability research or in data efficiency might render either premise (or both) less salient over time.⁹³ The point is therefore to consider a technology's features or parameters at a specific moment, as a snapshot in that technology's temporary, dynamic regulatory texture.

Indeed, considerations of a technology's features are relevant not just in terms of determining the most optimal governance levers, but also in evaluating the (political) *viability* of achieving such regulation in the first place. This is especially the case at the international level. For instance, Rosert & Sauer have argued that two previously successful global weapons ban campaigns, against blinding lasers and anti-personnel mines, owed their success in part to the fact that both technologies produced specific types of injuries (blinding and maiming) that were both particularly horrifying as well as particularly distinctive to each technology.⁹⁴ This is a specific illustration of a more general point: that along with diverse historical, political, and social factors, a technology's architectural characteristics can play a significant role in rendering some (weapons) technologies more 'regulation-resistant' than others.⁹⁵

⁹⁰ Cf. Calo, *supra* note 9.

⁹¹ Jenna Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, 3 BIG DATA SOC. 2053951715622512 (2016). Although for a counterargument, see Joshua A. Kroll, *The fallacy of inscrutability*, 376 PHIL TRANS R SOC A 20180084 (2018).

⁹² Dario Amodei & Danny Hernandez, *AI and Compute*, OPENAI BLOG (2018), <https://blog.openai.com/ai-and-compute/> (last visited May 22, 2018); BEN BUCHANAN, *The AI Triad and What It Means for National Security Strategy* (2020), <https://cset.georgetown.edu/research/the-ai-triad-and-what-it-means-for-national-security-strategy/> (last visited Aug 19, 2020).

⁹³ Aaron D. Tucker, Markus Anderljung & Allan Dafoe, *Social and Governance Implications of Improved Data Efficiency*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 378–384 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375863> (last visited Feb 12, 2020).

⁹⁴ Significantly, they argue that these traits are not shared by the current line of LAWS; Elvira Rosert & Frank Sauer, *How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies*, 0 CONTEMP. SECUR. POLICY 1–26 (2020). For a similar argument, see also Rebecca Crootof, *Why the Prohibition on Permanently Blinding Lasers is Poor Precedent for a Ban on Autonomous Weapon Systems*, LAWFARE (2015), <https://www.lawfareblog.com/why-prohibition-permanently-blinding-lasers-poor-precedent-ban-autonomous-weapon-systems> (last visited Apr 17, 2019).

⁹⁵ Sean Watts, *Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War*, 91 INT. LAW STUD. 83 (2015). Rebecca Crootof, *Jurisprudential Space Junk: Treaties and New Technologies*, in RESOLVING CONFLICTS IN THE LAW 106–129, 115–116 (Chiara Giorgetti & Natalie Klein eds., 2019),

As such, this argument amounts to an ‘asymmetric’ account of socio-technical change—or the assumption of ‘asymmetric governance rationale/target materiality’.⁹⁶ That is, in considering when AI capabilities create a ‘governance rationale’, this lens looks less at ‘inputs’ than at ‘outputs’. It focuses less on incremental progress in a certain technology, or its ‘essential features’, and more on the sociotechnical changes in the form of new (or anticipated) sociotechnical changes.⁹⁷ At the same time, in considering how these AI capabilities should be approached as a ‘governance target’, it does consider the technology’s ‘inputs’ (its temporary and changeable texture that determines its regulatory surface) as relevant considerations in considering distinct governance arrangements.

4.3.2 Regulatory surfaces and problem logics

Moreover, along with technical or material features, we should consider the particular regulatory surfaces of an AI technology, in terms of distinct *problem logics*.

As such, it is important to clarify that the ‘sociotechnical change’ perspective does not imply that the underlying technological progress in AI, or the deployment of specific *applications* in ever more sectors, are irrelevant and should not be paid attention. If anything, the opposite is the case: tracking and extrapolating patterns and rates in AI development, as well as the proliferation and latency of specific AI applications, remain critical in order to map the ‘regulatory surface’ of these challenges as governance targets. Individual applications can provide important signals, including about the urgency of regulation. This should be explored in greater detail, and as such the next section will explore a detailed taxonomy of problem logics.

4.4 Governing AI Sociotechnical Change: a taxonomy of Problem Logics

AI governance regimes might be designed with greater leverage over their governance ‘targets’ if they considered more explicitly the types of sociotechnical changes that are at issue, along with the material texture as well as problem logics.

One way of framing governance regimes could be to bucket AI applications with reference to the type of ‘governance rationale’ they give rise to. There is something to this approach, and yet it is incomplete in at least one way: just because two different AI use cases give rise to a similar governance rationale (e.g. they both create problems for ‘rights’), does not necessarily mean that they have a similar regulatory texture or problem logic. Accordingly, bucketing them together might not always be insightful, productive or actionable.

As such, whereas the question of *whether* a governance regime is needed should depart from an examination of governance rationales, the subsequent question of *how* a governance

<https://brill.com/view/book/edcoll/9789004316539/BP000015.xml> (last visited Mar 15, 2019). Crootof, *supra* note 12 at 23.

⁹⁶ See also the discussion in Chapter 3.4.3. (on ontological assumptions around the interaction between technology and society).

⁹⁷ This is explored, in the context of security regimes for military AI, in Chapter 7.1.1.

regime ought to be organised may better depart from an examination of these AI capabilities as ‘governance targets’, as discussed above.

This raises the question of how to fruitfully cluster types of sociotechnical change, in ways that facilitate common policies or insights. As previously noted,⁹⁸ scholars in AI ethics and AI governance have articulated diverse ways to segment the challenges of AI. These taxonomies are all valuable. However, at least in some cases, these frameworks may overemphasize pre-existing ‘thematic’ bodies or disciplines of law, rather than clusters of structured problems in the space of sociotechnical change.⁹⁹ Accordingly, to better grapple with the sociotechnical change driven by these technologies in a manner more sensitive to the different regulatory textures and problem logics, I suggest a more productive account could draw different distinctions.¹⁰⁰ Specifically, I suggest that AI systems¹⁰¹ can create sociotechnical changes in six ways that can be clustered as: ‘ethical challenges’, ‘security threats’, ‘safety risks’, ‘structural shifts’, ‘common benefits’, and ‘governance disruption’. Distinguishing between these six ‘ideal-types’ can be valuable, as they come with distinct regulatory surfaces, and might therefore be amenable to distinct governance logics or levers (See Table 4.1).

To paraphrase a witticism from science: all concepts are wrong—some concepts are useful.¹⁰² As such, it should be clarified that this taxonomy of six types of problem logics is not meant to be mutually exclusive: indeed, in practice, certain AI capabilities can give rise to a bucket of problems across these categories. In other cases, one of these dimensions might dominate the others. In either case, it is still valuable to disaggregate the distinct components or logics of these problems. Moreover, this taxonomy is also not meant to be exhaustive. It aims to capture certain regularities which help ask productive governance questions. For each category, we can ask—how does the AI capability create a governance *rationale*? How should this be approached as governance *target*? What are the barriers, and what governance solutions are foregrounded?

⁹⁸ In the introduction of this chapter.

⁹⁹ In this sense, they invoke what Nicolas Petit has called a ‘legalistic’ approach, which ‘consists in starting from the legal system, and proceed by drawing lists of legal fields or issues affected by AIs and robots.’ PETIT, *supra* note 8 at 2. He contrasts this to the ‘technological’ approach, which involves charting “legal issues from the bottom-up standpoint of each class of technological application” (*ibid*), such as focusing on driverless cars or social robots (what I would call an application-centric approach; see also the discussion of AI definitions in Chapter 2.1.1).

¹⁰⁰ Another promising typology (taking the perspective of AI researchers) is offered by Ashurt and colleagues, who, under ‘types of societal impact’, make a thematic distinction between ‘Economic’, ‘Environmental’, ‘Equality’, ‘Development’, ‘Political’, ‘Products and Services’, and ‘Science and Technology’ impacts. Carolyn Ashurt et al., *A Guide to Writing the NeurIPS Impact Statement*, MEDIUM (2020), https://medium.com/@operations_18894/a-guide-to-writing-the-neurips-impact-statement-4293b723f832 (last visited May 14, 2020).

¹⁰¹ This model could in principle be adapted and applied to the governance of many other technologies, though an exploration of this remains beyond the scope of this current project. I thank Jeffrey Ding for raising this possibility.

¹⁰² Cf. George E. P. Box, *Science and Statistics*, 71 J. AM. STAT. ASSOC. 791–799, 792 (1976).

Type and questions	Corresponding governance rationales	Examples in AI (selected)	Governance Surface (origin / barriers to resolution)	Governance Logics (selected)
Ethical challenges <i>What rights, values or interests does this threaten?</i>	<ul style="list-style-type: none"> New risks to moral interests, rights or values New threats to social solidarity Threats to democratic process 	<ul style="list-style-type: none"> <i>Justice</i>: bias; explainability <i>Power</i>: facial recognition <i>Democracy</i>: AI propaganda <i>Freedom</i>: ‘Code as Law’; ‘algocracy’ 	<ul style="list-style-type: none"> Actor <i>apathy</i> (to certain values) Underlying <i>societal disagreement</i> (culturally and over time) over how to weigh the values, interests or rights at stake 	<ul style="list-style-type: none"> Bans (mend—or end?) Oversight & accountability mechanisms; auditing ‘Machine ethics’ Ethics education Value-Sensitive Design
Security threats <i>How is this vulnerable to misuse or attack?</i>	<ul style="list-style-type: none"> New risks to moral interests, rights or values New risks to human health or safety 	<ul style="list-style-type: none"> <i>AI as tool</i>: DeepFakes; <i>AI as attack surface</i>: adversarial input <i>AI as shield</i>: fraudulent trading agents; UAV smuggling 	<ul style="list-style-type: none"> Actor <i>malice</i> (various motives) ‘Offense-defense balance’ of AI knowledge Intrinsic vulnerability of human social institutions to automated social engineering attacks. 	<ul style="list-style-type: none"> <i>Perpetrator-focused</i>: change norms, prevent access; improve detection & forensics capabilities to ensure attribution and deterrence <i>Target-focused</i>: reduce exposure; red-teaming; ‘security mindset’
Safety risks <i>Can we rely on and control this?</i>	<ul style="list-style-type: none"> New risks to human health or safety 	<ul style="list-style-type: none"> Unpredictability and opacity Environmental interactions Automation bias and ‘normal accidents’ ‘Value misalignment’ 	<ul style="list-style-type: none"> Actor <i>negligence</i>, overtrust and automation bias ‘Many hands’ problem—long and discrete supply chains Behavioural features of AI systems (opacity; unpredictability; optimisation failures; specification gaming) 	<ul style="list-style-type: none"> Relinquishment (of usage in extreme-risk domains) ‘Meaningful Human Control’ (various forms) Safety engineering (e.g. reliability; corrigibility; interpretability; limiting capability or deployment; formal verification) Liability mechanisms & tort law; open development
Structural shifts <i>How does this shape our decisions?</i>	<ul style="list-style-type: none"> (all, indirectly) 	<ul style="list-style-type: none"> Change calculations: LAWS lower perceived costs of conflict Increased scope for miscalculation: e.g. attack prediction systems 	<ul style="list-style-type: none"> Systemic <i>incentives</i> for actors. (alters choice architectures; increases uncertainty & complexity; competitive value erosion) Exacerbates other challenges 	<ul style="list-style-type: none"> Arms control (mutual restraint) Confidence-Building Measures (increase trust or transparency)
Common Benefits <i>How can we realize opportunities for good with this?</i>	<ul style="list-style-type: none"> Possible market failures 	<ul style="list-style-type: none"> Ensure AI interoperability ‘AI for global good’ projects Distributing benefits of AI 	<ul style="list-style-type: none"> Systemic <i>incentives</i> for actors (Coordination challenges around cost-sharing, free-riding) Overcoming loss aversion 	<ul style="list-style-type: none"> (Global) standards ‘Public interest’ regulation and subsidies ‘Windfall clause’ & redistributive guarantees
Governance Disruption <i>How does this change how we regulate?</i>	<ul style="list-style-type: none"> New risks directly to existing regulatory order 	<ul style="list-style-type: none"> AI systems creating <i>substantive ambiguity</i> in law Legal automation altering <i>processes</i> of law Erodes political <i>scaffolding</i> 	<ul style="list-style-type: none"> Legal system <i>exposure</i> and dependence on conceptual orders or assumptions 	<ul style="list-style-type: none"> Provisions to render governance ‘innovation-proof’: technological neutrality; authoritative interpreters, sunset clauses; ... Oversight for legal automation; distribution

Table 4.1. Taxonomy of AI problem logics under Sociotechnical Change

4.4.1 Ethical challenges

Ethical challenges around AI revolve around concerns that AI could enable or create new forms of behaviour, entities, or (power) relations that raise problems for fundamental rights or values in a society. This may concern behaviour that is outright illegal, or which is (still) legal but seen, by some or by many, as violating normative values. In most cases, these concerns are anchored in (anticipated) threats to *existing* values; however, in some cases technologies can also reveal or highlight new rights or values which we previously did not realize we valued.¹⁰³

A key feature of AI-driven sociotechnical changes that can be clustered as ‘ethical challenges’, is that the new behaviour, entities or relations commonly raise questions over foundational political theoretical concepts, such as justice, power, democracy and freedom.¹⁰⁴ As such, it should be no surprise that ethical challenges comprise one of the broadest and most ostensibly diverse categories, given a series of high-profile misuses or at least controversial use cases of AI.

For example, concerns around algorithmic *justice* adopt the ethical challenges problem logic. Concerns have been raised over algorithmic bias,¹⁰⁵ both in relatively ‘innocuous’ (but still important) settings such as in search machine prompts,¹⁰⁶ and especially in the context of systems used for high-stakes legal decision-making or policing.¹⁰⁷ More broadly, many have argued that the use of algorithmic decision-making by states or legal systems may affect important (human) rights,¹⁰⁸ especially when they are used in the context of vulnerable populations such as migrants.¹⁰⁹ Likewise, the capability of ‘using machine learning systems to make sophisticated but opaque decisions’ creates various ethical concerns around explainability.

Concerns over a loss of privacy also play into questions over changes in power relations.¹¹⁰ This is particularly spurred on by increasing use of AI in surveillance. AI can identify human

¹⁰³ See generally Jack Parker & David Danks, *How Technological Advances Can Reveal Rights?* (2019), http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_129.pdf. One example could be the ‘right to be forgotten’.

¹⁰⁴ JAMIE SUSSKIND, *FUTURE POLITICS: LIVING TOGETHER IN A WORLD TRANSFORMED BY TECH* (2018).

¹⁰⁵ AI Now Institute, *AI IN 2018: A YEAR IN REVIEW*, AI Now INSTITUTE (2018), <https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e> (last visited Feb 26, 2019); KATE CRAWFORD ET AL., *AI Now 2019 Report* 100 (2019), https://ainowinstitute.org/AI_Now_2019_Report.pdf; Kate Crawford & Ryan Calo, *There is a blind spot in AI research*, 538 NAT. NEWS 311 (2016).

¹⁰⁶ SAFIYA NOBLE, *ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM* (1 edition ed. 2018).

¹⁰⁷ Mareile Kaufmann, Simon Egbert & Matthias Leese, *Predictive Policing and the Politics of Patterns*, 59 BR. J. CRIMINOL. 674–692 (2018).

¹⁰⁸ Q. C. VAN EST, J. GERRITSEN & L. KOOL, *Human rights in the robot age: challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality* (2017), <https://research.tue.nl/en/publications/human-rights-in-the-robot-age-challenges-arising-from-the-use-of-> (last visited May 22, 2019).

¹⁰⁹ PETRA MOLNAR & LEX GILL, *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada’s Immigration and Refugee System* (2018), <https://citizenlab.ca/wp-content/uploads/2018/09/IHRP-Automated-Systems-Report-Web-V2.pdf> (last visited Apr 6, 2019); Petra Molnar, *Technology on the margins: AI and global migration management from a human rights perspective*, 8 CAMB. INT. LAW J. 305–330 (2019).

¹¹⁰ Ryan Calo, *Peeping HALs: Making Sense of Artificial Intelligence and Privacy*, 2 EJLS - EUR. J. LEG. STUD. (2010), <http://www.ejls.eu/6/83UK.htm> (last visited May 13, 2017). However, some have also suggested AI systems can be used in defence or protection of privacy. Andrea Scripa Els, *Artificial Intelligence as a Digital Privacy Protector*, 31 HARV. J. LAW TECHNOL. 19 (2017); Urs Gasser, *Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy*, 130 HARV. LAW REV. FORUM 10 (2016); Andrew Trask, *Safe*

faces, even masked ones in blurred pictures;¹¹¹ and can enable the remote identification of individuals not just by face but even by gait, (remotely laser-measured) heartbeat signature,¹¹² and even by bioacoustic signature (measuring how sound waves pass through human bodies).¹¹³

More generally, AI systems can change power structures,¹¹⁴ or drive growing inequality and societal instability, as a result of technological unemployment and a global oligopolistic or mercantilist market structure.¹¹⁵ Domestically, there are concerns that increasing algorithmic ‘perception control’, the dual disciplinary functions of pervasive (and even predictive) surveillance systems, and opaque algorithmic governance can further embed unequal and uncontestable power relations.¹¹⁶ AI systems might also strengthen authoritarian states in their surveillance and repression capabilities;¹¹⁷ or such capabilities could allow these actors to increasingly challenge the global liberal order.¹¹⁸ Finally, not all ethical concerns (directly) concern human interests or rights. For instance, some have charted the potential negative impacts of AI technology on sustainability and the natural world, either directly,¹¹⁹ or indirectly, because of the technology’s political effects on the dynamics and viability of effective environmentalist movements.¹²⁰

Considered in terms of governance *rationales*, ethical challenges implicate new risks to moral interests, rights or values; new threats to social solidarity, or threats to democratic process. Considered as a governance *target*, AI-linked sociotechnical changes which invoke ethical challenges have certain salient features. In weak cases, the problems cannot be traced back to AI

Crime Prediction: Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance (2017), <https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/> (last visited Jun 8, 2017).

¹¹¹ Matt Reynolds, *Even a mask won’t hide you from the latest face recognition tech*, NEW SCIENTIST, 2017, <https://www.newscientist.com/article/2146703-even-a-mask-wont-hide-you-from-the-latest-face-recognition-tech/> (last visited Jun 18, 2019); Amarjot Singh et al., *Disguised Face Identification (DFI) with Facial KeyPoints using Spatial Fusion Convolutional Network*, ARXIV170809317 CS (2017), <http://arxiv.org/abs/1708.09317> (last visited Jun 18, 2019); LilHay Newman, *AI Can Recognize Your Face Even If You’re Pixelated*, WIRED (2016), <https://www.wired.com/2016/09/machine-learning-can-identify-pixelated-faces-researchers-show/> (last visited Feb 26, 2017).

¹¹² David Hambling, *The Pentagon has a laser that can identify people from a distance—by their heartbeat*, MIT TECHNOLOGY REVIEW, 2019, <https://www.technologyreview.com/s/613891/the-pentagon-has-a-laser-that-can-identify-people-from-a-distance-by-their-heartbeat/> (last visited Jul 1, 2019). See broadly Hayward and Maas, *supra* note 74.

¹¹³ Joo Yong Sim et al., *Identity Recognition Based on Bioacoustics of Human Body*, IEEE TRANS. CYBERN. 1–12 (2019). (demonstrating an approach that achieved an accuracy of 97% at identifying individuals, nearly as accurate as fingerprints or iris scans, persisting over at least two months).

¹¹⁴ Hin-Yan Liu, *The power structure of artificial intelligence*, 10 LAW INNOV. TECHNOL. 197–229 (2018).

¹¹⁵ DAFOE, *supra* note 22 at 39–42.

¹¹⁶ John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 PHILOS. TECHNOL. 245–268 (2016).

¹¹⁷ Steven Feldstein, *The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression*, 30 J. DEMOCR. 40–52 (2019). Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCIENTIFIC AMERICAN, 2017, <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/> (last visited May 29, 2017).

¹¹⁸ Richard Danzig, *An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order*, 5 LAWFARE RES. PAP. SER. (2017), <https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf> (last visited Sep 1, 2017).

¹¹⁹ Roy Schwartz et al., *Green AI*, ARXIV190710597 CS STAT (2019), <http://arxiv.org/abs/1907.10597> (last visited Sep 2, 2019); KATE CRAWFORD & VLADAN JOLER, *Anatomy of an AI System* (2018), <http://www.anatomyof.ai> (last visited Jun 22, 2020); BIRD ET AL., *supra* note 5 at 28–29.

¹²⁰ Peter Dauvergne, *The globalization of artificial intelligence: consequences for the politics of environmentalism*, 0 GLOBALIZATIONS 1–15 (2020).

principals' hostility to certain rights. Tech developers, it might be assumed, do not generally set out to produce biased or opaque algorithms.¹²¹ However, the source of the problem could be found in the relative *apathy* to these values (at least in comparison with other—potentially legitimate—goals which are perceived to stand in tension to these), or an *ignorance* that their activities could impinge on these values.

In such cases, governance logics can highlight the ‘mend or end’ debates of algorithmic accountability. That is, they may attempt to ‘mend’ the AI application under scrutiny, either within the system in the form of machine ethics, or through organisational accountability or algorithmic auditing mechanisms.¹²² Alternatively, they may seek to outlaw the application entirely, as seen in recent US debates over facial recognition technology.¹²³

In hard cases, however, ethical challenges can be especially difficult to address, because they reflect underlying and live societal disagreements over the values, interests or rights at stake, or how these should be weighted. In those cases, AI’s sociotechnical changes may prove especially intractable. As such, these challenges will face particular contestation and differences amongst parties, potentially foregrounding the value of negotiating general consensus. Alternatively, and on the international level, small and informal clubs of like-minded actors might be able to move ahead and set certain standards.¹²⁴

4.4.2 Security threats

In other cases, AI capabilities enable sociotechnical changes that can be clustered as *security threats*. It is useful to clarify, first, what is meant by ‘security’, given the term’s diverse usage. For instance, Fjeld and others have argued that “security addresses external threats to an AI system”, as opposed to “internal” threats or flaws that threaten its performance.¹²⁵ This definition focusses on AI technology’s susceptibility to (external) hacking or spoofing, treating the AI as an attack surface. Here, technical work has focused on security threats to AI systems—as in the literature in data poisoning and ‘adversarial inputs’,¹²⁶ of the type which allowed diverse researchers to ‘spoof’ sensitive AI systems.¹²⁷

¹²¹ Though this may not always be the case, unfortunately. In at least some cases, clients or customers may in fact configure algorithms in biased ways, as with the algorithm that was rigged by the US ICE to recommend detention in all cases. Sam Biddle, *ICE’s New York Office Uses a Rigged Algorithm to Keep Virtually All Arrestees in Detention. The ACLU Says It’s Unconstitutional.*, THE INTERCEPT (2020), <https://theintercept.com/2020/03/02/ice-algorithm-bias-detention-aclu-lawsuit/> (last visited Sep 7, 2020).

¹²² Inioluwa Deborah Raji et al., *Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing*, in PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 33–44 (2020), <https://doi.org/10.1145/3351095.3372873> (last visited Feb 23, 2020).

¹²³ Luke Stark, *Facial Recognition is the Plutonium of AI*, 25 XRDS 50–55 (2019).

¹²⁴ Cf. Jean-Frédéric Morin et al., *How Informality Can Address Emerging Issues: Making the Most of the G7*, 10 GLOB. POLICY 267–273 (2019). This anticipates the discussion in, Chapter 7.4.6, on the ‘breadth vs. depth’ dilemma facing AI governance initiatives.

¹²⁵ Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, 37 (2020), <https://dash.harvard.edu/handle/1/42160420> (last visited Jan 16, 2020).

¹²⁶ Alexey Kurakin, Ian Goodfellow & Samy Bengio, *Adversarial Machine Learning at Scale*, ARXIV161101236 CS STAT (2017), <http://arxiv.org/abs/1611.01236> (last visited Jun 16, 2020).

¹²⁷ Including those in self-driving cars, or in volumetric medical data. TENCENT KEEN SECURITY LAB, *Experimental Security Research of Tesla Autopilot* (2019), [https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf.](https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf); Samuel G. Finlayson et al., *Adversarial attacks on medical machine*

However, in a broader perspective, security challenges should be considered as relating to some actor's actual and knowing misuse or exploitation of an AI system (or its vulnerabilities), in the pursuit of goals that are widely recognised as illegitimate, and often formally recognised as illegal.¹²⁸ More operationally, AI gives rise to distinct security challenges because the technology does not function just as a potential attack surface, but can also be used as a tool or even 'shield' for malicious actors.¹²⁹

Accordingly, other work has explored the potential for AI capabilities to be misused to scale up old security risks, or even to create entirely new attack vectors, in areas such as cybersecurity, crime, and terrorism.¹³⁰ Other security problem logics revolve around 'political security', as seen in concerns over how democracy and the rule of law will be eroded,¹³¹ especially in the face of 'computational propaganda', 'neural fake news',¹³² or 'deep fakes'.¹³³ There is some uncertainty over how imminent some of these attacks are, since some of the more exotic attacks or use cases may not transfer easily outside labs, or find at most limited use amongst only the most high-capacity actors.¹³⁴

Nonetheless, there are clearly a diversity of security threats around, to, and through AI systems.¹³⁵ As a governance *rationale*, most of these raise new risks to human health or safety, though they can also pertain to risks to moral interests, rights or values. As a 'governance target', AI sociotechnical changes that centre on *security* have distinct features. In the first place, they derive not from actor apathy or ethical laxness, but from 'malice'. In the second place, they depend not just on the sophistication of the AI capability in question, but also on the 'vulnerability' and 'exposure' of the target. This is key, because there is at least one reason to assume that AI-enabled attacks might pose a structurally larger security problem than conventional 'hacks'. This is because, as argued by Shevlane and Dafoe, AI is a technology that aims at automating activities that normally require human intelligence, which means that it is particularly suitable for

¹²⁸ learning, 363 SCIENCE 1287–1289 (2019). For a discussion, see also Ryan Calo et al., *Is Tricking a Robot Hacking?*, UNIV. WASH. SCH. LAW RES. PAP. (2018), <https://papers.ssrn.com/abstract=3150530> (last visited Aug 2, 2018).

¹²⁹ To clarify, the way the term is used here is closer to the notion of 'security' in 'cybersecurity', than to 'security' in traditional international law debates (e.g. relating to the use of kinetic military force).

¹³⁰ Hayward and Maas, *supra* note 74 at 6–9. We also identify a third category, of AI being used as a new potential criminal 'intermediary' or 'shield'. *Id.* at 9–10. On which, see also Lynn M. LoPucki, *Algorithmic Entities*, UCLA SCH. LAW LAW-ECON RES. PAP. (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2954173 (last visited May 19, 2017).

¹³¹ Brundage et al., *supra* note 40; Thomas C. King et al., *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, SCI. ENG. ETHICS (2019), <https://doi.org/10.1007/s11948-018-00081-0> (last visited Feb 21, 2019); Hayward and Maas, *supra* note 74.

¹³² Helbing et al., *supra* note 117; Paul Nemitz, *Constitutional democracy and technology in the age of artificial intelligence*, 376 PHIL TRANS R SOC A 20180089 (2018).

¹³³ Rowan Zellers, *Grover: A State-of-the-Art Defense against Neural Fake News* (2019), <https://rowanzellers.com/grover/?domain=nytimes.com&date=May+29%2C+2019&authors=&title=Answers+to+Where+to+Watch+the+Pope%E2%80%99s+U.S.+Visit&article=> (last visited Jun 13, 2019); Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models*, ARXIV190809203 Cs (2019), <http://arxiv.org/abs/1908.09203> (last visited Nov 18, 2019).

¹³⁴ Bolsover and Howard, *supra* note 65; Chesney and Citron, *supra* note 70; Katarina Kertysova, *Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Counteracted*, 29 SECUR. HUM. RIGHTS 55–81 (2018).

¹³⁵ The difficulty of anticipating or predicting these problems was discussed in section 4.2.4.

¹³⁶ See also Nicolas Papernot, *A Marauder's Map of Security and Privacy in Machine Learning*, ARXIV181101134 Cs (2018), <http://arxiv.org/abs/1811.01134> (last visited Jun 16, 2020).

interfering in human social systems—or exploiting features of a social system as vulnerability. The ‘exploits’ such attacks rely on are often features not of digital architectures or code, but of human social systems, which are far less amenable to quick or painless ‘patches’.¹³⁶ Accordingly, whereas software vulnerabilities in conventional cyber security can often be addressed by the dissemination and implementation of a software fix, addressing ‘vulnerabilities’ of the human social system might require far-reaching and prohibitively costly changes to ordinary human social practices or habits (such as our faith in the fidelity of voices).

AI security challenges therefore foreground some governance responses or governance logics. Some of these are perpetrator-focused: attempts to change norms, prevent access; improve detection & forensics capabilities to ensure attribution and deterrence. Others are aimed at reducing the vulnerability or exposure of targets, such as by red-teaming, and adopting a ‘security mindset’.¹³⁷

4.4.3 Safety risks

Other AI challenges can be characterised as *safety risks*.¹³⁸ AI systems, perhaps more than most or any other technology, introduce an unusually diverse palette of failure modes.

Much work to date has focused on the safety and reliability of individual ‘cyber-physical systems’—various autonomous vehicles, drones, delivery robots or other AI products that operate in the public or private space.¹³⁹ Kunz & Ó hÉigearthaigh have suggested that AI accident risks

¹³⁶ Toby Shevlane & Allan Dafoe, *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*, in PROCEEDINGS OF THE 2020 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (AIES '20), 177 (2020), <http://arxiv.org/abs/2001.00463> (last visited Jan 9, 2020) (“[d]igital systems are in general the easiest to patch at scale, whereas physical and social systems are often much more costly to change, especially at scale.”).

¹³⁷ C. Severance, *Bruce Schneier: The Security Mindset*, 49 COMPUTER 7–8 (2016).

¹³⁸ Again, there are various ways to define safety and security. For instance, Fjeld et al. argue that “The principle of safety generally refers to proper internal functioning of an AI system and the avoidance of unintended harms.” Fjeld et al., *supra* note 125 at 37. They note that an overarching term here is ‘reliability’, which connotes both safety as well as security, in that a system that is ‘reliable’ “is safe, in that it performs as intended, and also secure, in that it is not vulnerable to being compromised by unauthorized third parties” *Id.* at 37. In this taxonomy, however, I draw slightly different distinctions; ‘safety’ problems need not only come from inside an AI system, but can also be caused by its interaction and emergence with other systems or humans. Secondly, I use a broader conception of ‘security’, which also covers misuse of AI systems, in the service of attacks traditionally considered ‘security threats’.

¹³⁹ JAMES M. ANDERSON ET AL., *Autonomous Vehicle Technology: a Guide for Policymakers* (2016), https://www.rand.org/pubs/research_reports/RR443-2.html (last visited Oct 17, 2017). Note, there has been much work in this area which focused on the exploration of ‘trolley problems’. Sven Nyholm & Jilles Smids, *The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?*, 19 ETHICAL THEORY MORAL PRACT. 1275–1289 (2016); Edmond Awad et al., *The Moral Machine experiment*, 563 NATURE 59 (2018). Other scholars have critiqued this framing. Cf. Martin Cunneen et al., *Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics*, 51 CYBERN. SYST. 59–80 (2020); Johannes Himmelreich, *Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations*, 21 ETHICAL THEORY MORAL PRACT. 669–684 (2018); Abby Everett Jaques, *Why the moral machine is a monster* 10 (2019). Conversely, others have defended their relevance: Geoff Keeling, *Why Trolley Problems Matter for the Ethics of Automated Vehicles*, 26 SCI. ENG. ETHICS 293–307 (2020); Jean-François Bonnefon, Azim Shariff & Iyad Rahwan, *The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars*, 107 PROC. IEEE 502–504 (2019). For an exploration of structural discrimination around such vehicles, see also Hin-Yan Liu, *Three Types of Structural Discrimination Introduced by Autonomous Vehicles*, 51 UC DAVIS LAW REV. 32 (2018).

can emerge from four sources of harm: “(i) hardware faults, such as exploding batteries; (ii) data-related harm including data corruption, theft and manipulation, (iii) software malfunction or misuse, including AI software components, and (iv) human operator mistakes or deliberate attacks.”¹⁴⁰

AI’s safety risks can be exacerbated by the speed with which AI solutions can be rolled out, but also by the *general reliability challenges of all software development*: for instance, past studies have estimated that the software industry produces an average rate of 15-50 errors per 1,000 lines of code.¹⁴¹ However, in addition to facing reliability and safety challenges held in common with all software, AI systems are also vulnerable to additional sets of failures.¹⁴²

One concern is that many autonomous AI systems or applications can be *highly unpredictable*, at times intrinsically so. Indeed, depending on the specific architecture or techniques underlying an application, AI systems can be ‘brittle’, reacting unexpectedly when encountering unforeseen new situations.¹⁴³ In part, this derives from the limitations of current approaches, and the degree to which even systems that perform at super-human levels on narrow tasks often do so through decision-making architectures that are dissimilar from those used by humans.¹⁴⁴ As a result, as noted by Paul Scharre, “[i]n the wrong situation, AI systems go from supersmart to superdumb in an instant.”¹⁴⁵

Moreover, accidents can derive not just from the behaviour of individual systems, but also through an AI system’s *interactions with the environment*. For instance, the interaction of one algorithm with another can give rise to cascading interactions and system accidents, such as the 2010 algorithmic stock market ‘flash crash’, triggered by high-frequency trading algorithms.¹⁴⁶ Alternately, an AI system’s interaction with humans can produce unanticipated reactions. This was illustrated by Microsoft’s failure to anticipate the risks of its “Tay” chatbot learning from

¹⁴⁰ Martina Kunz & Seán Ó hÉigearaigh, *Artificial Intelligence and Robotization*, in OXFORD HANDBOOK ON THE INTERNATIONAL LAW OF GLOBAL SECURITY, 2 (Robin Geiss & Nils Melzer eds., 2020), <https://papers.ssrn.com/abstract=3310421> (last visited Jan 30, 2019).

¹⁴¹ STEVE MCCONNELL, CODE COMPLETE: A PRACTICAL HANDBOOK OF SOFTWARE CONSTRUCTION, SECOND EDITION (2nd edition ed. 2004); As discussed in PAUL SCHARRE, *Autonomous Weapons and Operational Risk* 13 (2016), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf. Indeed, a 2018 Gartner report predicted that, “through 2022, 85 percent of AI projects will deliver erroneous outcomes due to bias in data, algorithms or the teams responsible for managing them.” Gartner, *Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence*, GARTNER (2018), <https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence> (last visited Sep 7, 2020).

¹⁴² While the below is one way of bucketing issues, there are broader AI safety issues. For instance, in one recent review, the ‘AI safety issue groups’ are disambiguated into 21 distinct categories. Hernandez-Orallo et al., *supra* note 87 at 7.

¹⁴³ Christian Szegedy et al., *Intriguing properties of neural networks*, ARXIV13126199 Cs (2014), <http://arxiv.org/abs/1312.6199> (last visited Jun 22, 2020).

¹⁴⁴ CUMMINGS ET AL., *supra* note 6 at 2. (“Machines and humans have different capabilities, and, equally importantly, make different mistakes based on fundamentally divergent decision-making architectures”).

¹⁴⁵ Paul Scharre, *Killer Apps: The Real Danger of an AI Arms Race*, FOREIGN AFFAIRS, 2019, <https://www.foreignaffairs.com/articles/2019-04-16/killer-apps> (last visited Apr 17, 2019).

¹⁴⁶ Andrei A. Kirilenko et al., *The Flash Crash: The Impact of High Frequency Trading on an Electronic Market*, SSRN ELECTRON. J. (2014), https://www.cftc.gov/sites/default/files/idc/groups/public/@economicanalysis/documents/file/oce_flashcrash0314.pdf (last visited Sep 8, 2020). It should be noted that later investigations suggest the initial cause of the crash was in fact malicious human error.

Twitter users, who rapidly turned it into a bot reciting the racist content it was fed.¹⁴⁷ Indeed, recent interdisciplinary work into the study of ‘machine behaviour’ has drawn on existing frameworks from the field of ethology (the science of animal behaviour), which emphasises parallel examinations of the *function, mechanism, development, or evolutionary history* of a given AI system’s behaviour, and emphasizing how these factors should in turn be explored not just at the level of individual machines, but should also reckon with collective behaviour by networks of interacting systems, or hybrid interactions between systems and humans.¹⁴⁸

In a similar vein, it has been argued that some AI systems may even be categorically *susceptible to emergent and cascading ‘normal accidents’*.¹⁴⁹ ‘Normal Accidents’ are emergent system accidents that inevitably—as a ‘normal’ result of the way the system has been set up or is operated—emerge in specific organisations working with specific sociotechnical systems under particular operational settings.¹⁵⁰ Such accidents are particularly challenging to mitigate in AI systems, because keeping a ‘human-in-the-loop’ is not always viable in all applications, given the speed or scale of some AI operations. However, normal accident theory suggests that implementing automated safety measures may perversely enable normal accidents, by increasing a system’s ‘interactive complexity’ further. Moreover, such ‘fail-safes’ instil further automation bias, as operators may trust safety systems overmuch, resulting in ‘risk homeostasis’ and creating new avenues for error.¹⁵¹ Accordingly, automation bias and inappropriate human ‘overtrust’ of AI systems may ensure that a human that is nominally ‘on the loop’ in practice functions less like a guardian and more like a ‘moral crumple zone’.¹⁵²

Beyond these specific safety concerns, there are also more general concerns around *peculiar behaviours and failure modes* displayed by AI systems.¹⁵³ Even relatively simple AI agents can solve existing bounded problems in ways unanticipated by their designers, producing strategies that technically solve the given problem, but not in a manner that was ever intended.¹⁵⁴

¹⁴⁷ Dave Gershgorn, *Here’s How We Prevent The Next Racist Chatbot*, POPULAR SCIENCE (2016), <https://www.popsci.com/heres-how-we-prevent-next-racist-chatbot> (last visited Feb 21, 2019).

¹⁴⁸ Iyad Rahwan et al., *Machine behaviour*, 568 NATURE 477 (2019)..

¹⁴⁹ Matthijs M. Maas, *Regulating for “Normal AI Accidents”: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment*, in PROCEEDINGS OF THE 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 223–228 (2018), <https://doi.org/10.1145/3278721.3278766> (last visited Sep 8, 2020). See also SCHARRE, *supra* note 141; STEPHANIE CARVIN, *Normal Autonomous Accidents: What Happens When Killer Robots Fail?* (2017), <https://papers.ssrn.com/abstract=3161446> (last visited Dec 17, 2019).

¹⁵⁰ See classically Charles Perrow, *Normal Accidents: Living with High Risk Technologies* (1984), <http://press.princeton.edu/titles/6596.html> (last visited Mar 4, 2017).

¹⁵¹ Maas, *supra* note 149 at 225. There is in general an extensive literature on automation bias. See also Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, *Automation bias: a systematic review of frequency, effect mediators, and mitigators*, 19 J. AM. MED. INFORM. ASSOC. JAMIA 121–127 (2012).

¹⁵² M. C. ELISH, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (We Robot 2016)* (2016), <https://papers.ssrn.com/abstract=2757236> (last visited Nov 14, 2017).

¹⁵³ Amodei et al., *supra* note 88; Ram Shankar Siva Kumar et al., *Failure Modes in Machine Learning Systems*, ARXIV191111034 CS STAT (2019), <http://arxiv.org/abs/1911.11034> (last visited Jan 7, 2020); Jan Leike et al., *AI Safety Gridworlds*, ARXIV171109883 CS (2017), <http://arxiv.org/abs/1711.09883> (last visited Dec 1, 2017).

¹⁵⁴ For a series of amusing and revealing examples, see Joel Lehman et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, ARXIV180303453 CS (2018), <http://arxiv.org/abs/1803.03453> (last visited Apr 2, 2018). A classic example also includes a genetic algorithm intended to evolve a new design for an oscillating circuit, which hit on a way to hack its hardware to function as a receiver for radio waves emitted from computers in the area. See J. Bird & P. Layzell, *The evolved radio and its implications for modelling the evolution of novel sensors*, 2 in

While the resulting behaviour may at times be amusing, at other times, situations of ‘perverse instantiation’ or ‘reward hacking’ results in hazardous (or potentially illegal) behaviour.¹⁵⁵ Indeed, such failures are but one of several distinct but generalizable failure modes that have been identified in AI systems.¹⁵⁶

One significant concern is that, rather than the early ‘growing pains’ of an immature technology, such incidents may reflect a core problem with the way AI systems are commonly used in ‘optimizing metrics’ in overtly narrow ways;¹⁵⁷ or that these failures demonstrate intrinsic behavioural risks that arise around certain AI systems that explore the landscape of potential strategies in order to pursue or optimize certain narrow goals.¹⁵⁸ If that is the case, some of the risks from such systems may increase even as they get more sophisticated. As Dafoe has suggested;

“previous kinds of accidents arise because the AI is “too dumb”. More advanced AI systems will overcome some of these risks, but gain a new kind of accident risk from being “too clever”. In these cases a powerful optimization process finds “solutions” that the researchers did not intend, and that may be harmful.”¹⁵⁹

Indeed, this has fed concerns that if continued technological progress eventually is to yield more capable AI systems, such systems might pose extreme risks to human welfare if they are not properly ‘aligned’.¹⁶⁰ While cultural depictions of AI have fed pervasive misperceptions about the nature of such risks,¹⁶¹ there are reasons why we might expect that the ‘alignment’ of- or control over advanced AI systems with human values or interests would not by default be an easy task.

Curiously, the difficulty of such ‘alignment’ is reflected not just by work from AI research or philosophy, but can also be derived from extensive scholarship in law and economics on various

PROCEEDINGS OF THE 2002 CONGRESS ON EVOLUTIONARY COMPUTATION. CEC'02 (CAT. No.02TH8600) 1836–1841 (2002), <http://ieeexplore.ieee.org/document/1004522/> (last visited Jul 10, 2020). See also Dario Amodei & Jack Clark, *Faulty Reward Functions in the Wild*, OPENAI (2016), <https://openai.com/blog/faulty-reward-functions/> (last visited Feb 18, 2017). And see generally Victoria Krakovna et al., *Specification gaming: the flip side of AI ingenuity*, DEEPMIND (2020), <https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity> (last visited May 12, 2020).

¹⁵⁵ In a simulated market, trading algorithms discovered and settled upon tacit price fixing and collusion strategies to maximize profit. Enrique Martinez Miranda, Peter McBurney & Matthew J. W. Howard, *Learning Unfair Trading: A Market Manipulation Analysis from the Reinforcement Learning Perspective*, in PROCEEDINGS OF THE 2016 IEEE CONFERENCE ON EVOLVING AND ADAPTIVE INTELLIGENT SYSTEMS, EAIS 2016 103–109 (2016), [https://kclpure.kcl.ac.uk/portal/en/publications/learning-unfair-trading\(4a13e3a7-0af5-48b5-898a-582bcb51e9af\)/export.html](https://kclpure.kcl.ac.uk/portal/en/publications/learning-unfair-trading(4a13e3a7-0af5-48b5-898a-582bcb51e9af)/export.html) (last visited Feb 20, 2019); King et al., *supra* note 130 at 10–12.

¹⁵⁶ Amodei et al., *supra* note 88 at 3.

¹⁵⁷ Rachel Thomas & David Uminsky, *The Problem with Metrics is a Fundamental Problem for AI*, ARXIV200208512 CS (2020), <http://arxiv.org/abs/2002.08512> (last visited Aug 27, 2020).

¹⁵⁸ Cf. Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk.*, in GLOBAL CATASTROPHIC RISKS 308–345 (Nick Bostrom & Milan M. Cirkovic eds., 2008).

¹⁵⁹ DAFOE, *supra* note 22 at 26.

¹⁶⁰ STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019), <https://www.amazon.com/Human-Compatible-Artificial-Intelligence-Problem-ebook/dp/B07N5J5FTS> (last visited Dec 4, 2019). For a survey of approaches, see also KAJ SOTALA & ROMAN V. YAMPOLSKIY, *Responses to Catastrophic AGI Risk: A Survey*. (2013), <https://intelligence.org/files/ResponsesAGIRisk.pdf>.

¹⁶¹ Future of Life Institute, *AI Safety Myths*, FUTURE OF LIFE INSTITUTE (2016), <https://futureoflife.org/background/aimyths/> (last visited Oct 26, 2017).

principal-agent problems, and the challenges around ‘incomplete contracting’ (the impossibility to write a ‘complete contingent contract’ that specifies all possible situations, creating the need to rely on ex post adjudication mechanisms);¹⁶² it is reflected in work on human psychology;¹⁶³ and it can be derived from various studies of optimal behaviour in existing reinforcement learning systems.¹⁶⁴ There is therefore active work on various dimensions of ‘scalable’ AI safety,¹⁶⁵ including ‘interruptibility’,¹⁶⁶ ‘safe exploration’;¹⁶⁷ and multi-agent dynamics between different algorithms,¹⁶⁸ amongst others.¹⁶⁹ Of course, much uncertainty remains. It is far from obvious how great such risks will ultimately be, nor on what timeframe they might manifest, nor—assuming the possibility of eventual catastrophic risks—the extent to which meaningful technical ‘safety’ research can be undertaken in the near term.¹⁷⁰ However, thoughtful people have at least entertained positions on various sides of these debates. That should not be cause for panic, but it might be for further investigation and strategy formulation.

¹⁶² Dylan Hadfield-Menell & Gillian Hadfield, *Incomplete Contracting and AI Alignment*, ARXIV180404268 Cs (2018), <http://arxiv.org/abs/1804.04268> (last visited Aug 22, 2018).

¹⁶³ For instance, psychological research suggests that ‘dark triad’ individuals are able to understand others, without necessarily empathizing. Petri J. Kajonius & Therese Björkman, *Individuals with dark traits have the ability but not the disposition to empathize*, 155 PERSONAL. INDIVID. DIFFER. 109716 (2020). This weakly suggests that one common assurance (‘if an AI system were ever truly intelligent, it would know what we meant by our requests, and care about achieving that’) need not hold even amongst humans.

¹⁶⁴ Illustrating for instance the ‘instrumental convergence’ thesis. Stephen M. Omohundro, *The Basic AI Drives*, 171 FRONT. ARTIF. INTELL. APPL. 483–492 (2008). For a recent demonstration of this, see Alexander Matt Turner, *Optimal Farsighted Agents Tend to Seek Power*, ARXIV191201683 Cs (2019), <http://arxiv.org/abs/1912.01683> (last visited Jan 6, 2020).

¹⁶⁵ Jose Hernandez-Orallo, Fernando Martinez-Plumed & Shahar Avin, *Surveying Safety-relevant AI Characteristics*, in PROCEEDINGS OF 1ST AAAI’S WORKSHOP ON ARTIFICIAL INTELLIGENCE SAFETY (SAFEAI) 9 (2019), http://ceur-ws.org/Vol-2301/paper_22.pdf. Note, this work spans both ‘near-term’ concerns over AI alignment as well as longer-term concerns over possible catastrophic risks from enormously capable future AI systems. The dialogue between these communities is at times adversarial, but needs not be so, given many of the common themes they explore. For more nuanced explorations on these disagreements, see Seth D. Baum, *Reconciliation between factions focused on near-term and long-term artificial intelligence*, 33 AI SOC. 565–572 (2018); Stephen Cave & Seán S. Ó hÉigearthaigh, *Bridging near- and long-term concerns about AI*, 1 NAT. MACH. INTELL. 5 (2019); Carina Prunkl & Jess Whittlestone, *Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 138–143 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375803> (last visited Feb 12, 2020); Edward Parson et al., *Artificial Intelligence in Strategic Context: An Introduction*, AI PULSE (2019), <https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/> (last visited Feb 26, 2019); Seth D. Baum, *Medium-Term Artificial Intelligence and Society*, 11 INFORMATION 290 (2020).

¹⁶⁶ Laurent Orseau & Stuart Armstrong, *Safely Interruptible Agents*10 (2016).

¹⁶⁷ Joshua Achiam & Dario Amodei, *Benchmarking Safe Exploration in Deep Reinforcement Learning* (2019), <https://pdfs.semanticscholar.org/4d0f/6a6ffcd6ab04732ff76420fd9f8a7bb649c3.pdf>.

¹⁶⁸ David Manheim, *Overoptimization Failures and Specification Gaming in Multi-agent Systems*, ARXIV181010862 Cs (2018), <http://arxiv.org/abs/1810.10862> (last visited Jan 14, 2019); David Manheim, *Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence*, 3 BIG DATA COGN. COMPUT. 21 (2019); JESSE CLIFTON, *Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda* (2020), <https://longtermrisk.org/research-agenda?fbclid=IwAR0TeDk9PfRzCOi4v7t2RcSwmb0xhC-I4mEYQodMAyBnle2pmfBFOMJylo>.

¹⁶⁹ See also a list of sub-projects in DAFOE, *supra* note 22 at 29–30. See also Tom Everitt, Gary Lea & Marcus Hutter, *AGI Safety Literature Review*, ARXIV180501109 Cs (2018), <http://arxiv.org/abs/1805.01109> (last visited May 7, 2018). For updates on this rapidly evolving field, one can follow <https://rohinshah.com/alignment-newsletter/>

¹⁷⁰ Prunkl and Whittlestone, *supra* note 165.

In sum, however, it should be clear that in the context of AI technology, ‘safety’ becomes a more nuanced, refined and tricky topic than many conventional approaches or doctrines in either ‘safety regulation’ or ‘safety engineering’ have conventionally reckoned with. As a governance *rationale*, these safety risks most obviously involve new risks to human health and safety. As a governance *target*, these risks have distinct features that are relevant to appropriate regulation. Rather than be the fault of human actors’ *apathy* or malice, they may derive from their *ignorance* or *negligence*, coupled with the effects of overtrust and automation bias. This is exacerbated, at the industrial level, by the ‘many hands’ problem. However, at the artefactual level, this problem is also exacerbated by a range of behavioural features of some AI approaches.

Accordingly, governance logics for AI safety problems may commonly be closely tied to questions of safety engineering (to ensure e.g. reliability; corrigibility; interpretability; limiting capabilities or deployment spaces; formal verification). However, it can also be grounded in operational principles such as ‘Meaningful Human Control’—formulated in a broad way. At the institutional level, safety challenges foreground *ex post* liability mechanisms & tort law.¹⁷¹ At the very limit, safety concerns that cannot be adequately secured might warrant (temporary) relinquishment of certain AI capabilities, either categorically or at least their deployment in certain high-stakes domains.¹⁷²

4.4.4 Structural shifts

In terms of *structure*, scholars increasingly examine the ways various technologies, including AI, could (inadvertently) alter decision environments for human actors at the systemic level, in ways that could increase the risk of miscalculation or accident, or lead to greater tensions, instability, or conflict, or pressure various parties to incrementally surrender or compromise on certain values.¹⁷³ Notably, Remco Zwetsloot and Allan Dafoe have pointed out how, in examining risks from AI, many implicitly or explicitly bucket problems as coming from either ‘accident’ or ‘misuse’—and have argued that this “accident-misuse dichotomy obscures how technologies, including AI, often create risk by shaping the environment and incentives”.¹⁷⁴ They accordingly

¹⁷¹ Although for some of the shortfalls of these, see also Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, HARV. J. LAW TECHNOL., 388–392 (2016), <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf> (last visited Mar 5, 2018).

¹⁷² Such, for instance, was one conclusion of Charles Perrow with regards to the ‘normal accident risks’ from nuclear power. Attributed in: Nick Pidgeon, *In retrospect: Normal Accidents*, 477 NATURE 404–405, 404 (2011). Interestingly, during the Cold War, computer scientists made the argument that, given the inevitability of computer bugs, and the impossibility of achieving flawless software performance, we ought not to integrate computer systems in nuclear command and control systems, because these could never be made sufficiently reliable. Alan Borning, *Computer system reliability and nuclear war*, 30 COMMUN. ACM 112–131 (1987). For a recent argument making this link to the integration of machine learning in nuclear command and control, see also Shahar Avin & S. M. Amadae, *Autonomy and machine learning at the interface of nuclear weapons, computers and people, in THE IMPACT OF ARTIFICIAL INTELLIGENCE ON STRATEGIC STABILITY AND NUCLEAR RISK* (V. Boulain ed., 2019), <https://www.repository.cam.ac.uk/handle/1810/297703> (last visited Oct 16, 2019).

¹⁷³ Remco Zwetsloot & Allan Dafoe, *Thinking About Risks From AI: Accidents, Misuse and Structure*, LAWFARE (2019), <https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure> (last visited Feb 12, 2019); Agnes Schim van der Loeff et al., *AI Ethics for Systemic Issues: A Structural Approach* (2019), <http://arxiv.org/abs/1911.03216> (last visited Jan 13, 2020).

¹⁷⁴ Zwetsloot and Dafoe, *supra* note 173.

suggest that this dichotomy should be expanded to also take stock of a ‘structural perspective’ on AI risks.¹⁷⁵

Rather than just examining how new technology can afford agents with new capabilities—that is, new opportunities for (mis)use—this perspective asks us to consider the broader effects on society and social institutions, including how the introduction of AI systems may unwittingly shape the environment and incentives (the ‘structure’) in which decision-makers operate. While technologies do not shape society deterministically,¹⁷⁶ they may still reshape the strategic environment and structural incentives under which actors operate (especially in competitive contexts), creating pressures that facilitate—or even appear to ‘compel’—other first-order effects in the domains of ethics, safety, or security.¹⁷⁷

It has long been recognised that prevailing technological conditions can often provide—or at least contribute to ‘security dilemmas’ in which states might find themselves locked.¹⁷⁸ Indeed, Albert Stein has noted that rather than an ‘inherent’ feature of international politics, the security dilemma is predicated on structural features of the international system, which include the state of knowledge, science, and technology. As such, he argues that “[...] the security dilemma presumes either that offensive weapons exist and are superior to defensive ones, or that weapons systems are not easily distinguishable”.¹⁷⁹

The same could apply to the use of AI. Specifically, new AI capabilities or use cases might unintentionally—and, at times ‘invisibly’—shift actors’ incentives in hazardous ways, for various reasons, and at both the global and local levels (see Table 4.2). Such sociotechnical change would be all the harder to anticipate and mitigate, because it might not be intended by any one actor.¹⁸⁰

¹⁷⁵ *Id.*

¹⁷⁶ As discussed in Section 3.4.3.

¹⁷⁷ For instance, one influential account of the origins of the First World War argues that the specific socio-technological features of the contemporary European railroad system required (or at least put an operational premium on) certain logistical choices, such as tight and interlocking schedules, which in turn required (or at least put a strategic premium on) rapid, all-or-nothing mass mobilization decisions. This reduced states’ manoeuvre room to signal to one another through more muted and less escalatory troop movements, and pitched the dominoes of general war. That is not to say that this technological configuration either compelled or necessitated war, but rather that it was a contributing factor that structured and constrained each party’s decision space (and time) in making choices around the outbreak and scope of war. Stephen Van Evera, *The Cult of the Offensive and the Origins of the First World War*, 9 INT. SECUR. 58–107 (1984). However, it should be noted that this interpretation remains contested. See for instance the critique by KEIR A. LIEBER, WAR AND THE ENGINEERS: THE PRIMACY OF POLITICS OVER TECHNOLOGY (1 edition ed. 2008). And for a discussion, see Jack Snyder & Keir A. Lieber, *Defensive Realism and the “New” History of World War I*, 33 INT. SECUR. 174–194 (2008).

¹⁷⁸ For the foundational work on the security dilemma, see John H. Herz, *Idealist Internationalism and the Security Dilemma*, 2 WORLD POLIT. 157–180 (1950); For a modern restatement, see Shiping Tang, *The Security Dilemma: A Conceptual Analysis*, 18 SECUR. STUD. 587–623 (2009).

¹⁷⁹ Arthur A. Stein, *Coordination and Collaboration: Regimes in an Anarchic World*, 36 INT. ORGAN. 299–324, 320 (1982).

¹⁸⁰ It is useful to note that, because these involve ways in which AI systems affect the decisions of many interrelated actors, the problems they create are particularly acute in competitive contexts, but they could also apply more broadly, even in non-zero-sum contexts.

Diffusion of new AI capabilities...	... Alters choice architecture (change calculations and decisions of principals)	... Drives uncertainty & complexity (increase room for errors and mistakes)	... Forces value erosion from competition (as each reacts to pressures of other's AI choices)
Macro-level (Strategic, global decisions)	<ul style="list-style-type: none"> Makes geopolitical competition appear more as a 'winner-take-all' or zero-sum competition Computational propaganda facilitates greater 'lawfare', reduces reputational costs of misbehaviour 	<ul style="list-style-type: none"> Introduces general <i>uncertainty</i> over relative capabilities, enabling strategic miscalculation 	<ul style="list-style-type: none"> Perceived pressures to develop AI weapons, even if it increases proliferation risk to malicious actors
Micro-level (Tactical, local decisions)	<ul style="list-style-type: none"> Lowers prospective (reputational) costs of behaviour (e.g. less kinetic military conflicts; less blatant forms of intrusive surveillance) 	<ul style="list-style-type: none"> Narrow time windows for decisions, increasing potential for accidental escalation Predictive algorithms might create <i>false certainty</i>, driving adverse self-fulfilling cycles 	<ul style="list-style-type: none"> Perceived pressures to deploy AI weapons, even if it increases 'flash war' risk
Introduces or exacerbates...	AI ethics and security challenges; general conflict	AI safety risks	All other issues; erodes conditions for cooperation

Table 4.2. Types of AI-induced structural shifts (macro vs. micro), with examples

In the first place, new AI applications could *alter choice architectures*, changing the calculations or deliberate decisions of principals, in ways that either outright privilege choices leading to AI ethics or security challenges, or indirectly lead to general conflict.

At the *macro-level*, for instance, certain AI capabilities could make relations amongst actors appear *more zero-sum*, or create the impression of a 'winner-take-all' competition.¹⁸¹ More loosely, AI applications in speeding up scientific research or various industries could theoretically reduce the economic or scientific interdependence of states, in ways that afford states a greater sense of self-sufficiency. Finally, improvements in the capability of AI systems to shape public perceptions or debate could increase the willingness, on behalf of unscrupulous actors, to engage in more exploitative practices, knowing that the global (or at least domestic) reputational fallout might be more easily contained.¹⁸²

AI capabilities could also alter incentives and choices at the *meso- or micro-level*. They could lower the threshold to a specific conflict or attack, for instance by lowering the perceived risks or costs of such actions. This could happen because the new AI capability sufficiently increases the expected probability of rapid victory; reduces the direct danger to human soldiers, or reduces the perceived intensity or visibility of conflict in the public eye.¹⁸³ Such dynamics need not remain contained to solely military contexts, however. Technological advances in remote

¹⁸¹ Zwetsloot and Dafoe, *supra* note 173.

¹⁸² On the other hand, technologies could also certainly be used by other actors to unveil abuses, and to hold parties to greater account. See briefly Lorna McGregor, *Are New Technologies an Aid to Reputation as a Disciplinarian?*, 113 AJIL UNBOUND 238–241 (2019). This is further taken up in chapter 5.5.2.

¹⁸³ Arthur Holland Michel, *The Killer Algorithms Nobody's Talking About*, FOREIGN POLICY (2020), <https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/> (last visited Jan 21, 2020). See also Matthijs M. Maas, *Innovation-Proof Governance for Military AI? How I learned to stop worrying and love the bot*, 10 J. INT. HUMANIT. LEG. STUD. 129–157, 144–146 (2019). (discussing how certain types of military AI might reduce the reputational costs of noncompliance with bans). This also anticipates the discussions, in section 5.6.2.1 ('increasing the spoils of noncompliance').

biometric identification technologies could well enable new vectors of surveillance. These might be as granular as studding public infrastructure with facial-recognition cameras, yet be much less visible and therefore perceived as much less intrusive by the public by comparison.¹⁸⁴

In the second place, AI capabilities might result in behaviours that drive increased *uncertainty and complexity*, increasing the room for errors and mistakes, potentially producing safety challenges.

At the *macro-level*, new AI use cases could drive *general strategic uncertainty* or lack of clarity over relative (military or cyber) capabilities, or over the balance of power, opening up scope for miscalculation, mutual misperceptions, and misunderstandings.¹⁸⁵ More specifically, they could create overlap between offensive and defensive uses or actions, driving security dilemmas. More generally, uncertainty could be fed as a result of AI applications in ‘neural fake news’ and computational propaganda driving a decay in the ‘epistemic security’ landscape.¹⁸⁶

At the *micro-level*, new AI capabilities might *narrow time-windows* for decision makers to make critical choices, reducing the ability of policymakers to resolve or reduce high levels of uncertainty. For instance, systems that would put a strong tactical or strategic premium on offense over defense,¹⁸⁷ might be perceived to put such parties in a ‘use it or lose it’ situation. Indeed, Schneider has previously identified a military ‘capability-vulnerability paradox’ in ‘digitally-enabled warfare’, whereby the same digital capabilities that make the (US) military more effective on the battlefield, also render them more vulnerable to pre-emptive attacks on the underlying digital networks and technologies.¹⁸⁸ As such, the introduction of certain AI capabilities by any one party, might put pressure on decision-makers—even well-intentioned ones—on both sides to make risky decisions, or decisions under imperfect information.¹⁸⁹

Moreover, at the micro-level, the predictions of AI models could also produce *false certainty* amongst actors, for instance, by creating adverse self-fulfilling predictions or systemic pressures: the use of machine learning for pre-attack ‘early warning’ system could inflame tense geopolitical situations and lower the threshold to the use of force.¹⁹⁰ Even AI models that are used ‘for good’ might create such perverse systemic externalities if deployed carelessly: for instance, Van der Loeff and colleagues have discussed the hypothetical case of an AI system used to predict the impact of climate on land and agricultural production, in order to predict and manage food

¹⁸⁴ We refer to this as the ‘ubiquitous yet tacit surveillance’ of the ‘hidden state’, in Hayward and Maas, *supra* note 74 at 12–13. Others have described such trends as the ‘Disneyfication’ of smart city policing. Elizabeth E. Joh, *Policing the smart city*, 15 INT. J. LAW CONTEXT 177–182 (2019).

¹⁸⁵ Matthew Kroenig & Bharath Gopalaswamy, *Will disruptive technology cause nuclear war?*, BULLETIN OF THE ATOMIC SCIENTISTS (2018), <https://thebulletin.org/2018/11/will-disruptive-technology-cause-nuclear-war/> (last visited Nov 22, 2018).

¹⁸⁶ Lin, *supra* note 66.

¹⁸⁷ On the scaling of the offense-defense balance, see Ben Garfinkel & Allan Dafoe, *How does the offense-defense balance scale?*, 42 J. STRATEG. STUD. 736–763 (2019).

¹⁸⁸ JACQUELYN SCHNEIDER, *Digitally-Enabled Warfare: The Capability-Vulnerability Paradox* 15 (2016), <https://www.cnas.org/publications/reports/digitally-enabled-warfare-the-capability-vulnerability-paradox>; Jacquelyn Schneider, *The capability/vulnerability paradox and military revolutions: Implications for computing, cyber, and the onset of war*, 42 J. STRATEG. STUD. 841–863 (2019).

¹⁸⁹ Zwetsloot and Dafoe, *supra* note 173.

¹⁹⁰ Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NATL. SECUR. LAW POLICY 1–25, 6–7 (2019).

supplies. Such a system would be very valuable for ensuring food security, but it might also spark adverse systemic effects—“price hikes, hoarding of food supplies, environmentally unsustainable practices, conflict over arable land and increased inequality”—if its information was shared freely or publicly.¹⁹¹

Thirdly, structural shifts produced by AI systems could result in *value erosion from competition*.¹⁹² These are situations where actors individually would prefer to maintain certain values or principles (e.g. spending public funds on education rather than on defence; securing full data privacy; maintaining a ‘human-in-the-loop’ in autonomous weapons systems; exhaustively testing autonomous vehicles in safe circumstances). However, each actor perceives it is confronted by steep trade-offs between maintaining those values or compromising on them in order to remain (economically or strategically) competitive with rivals who, it is feared, are more willing to make that trade.¹⁹³ While it is not true that such strategic pressures are invariably strong or irreversible, there are many cases in which they manifest as harmful feedback cycles that generate trade-offs between private gains and public harms, or various races to the bottom.

At the *macro level*, this could result in development decisions that create downstream security or ethics problems. For instance, this dynamic could result in parties developing certain AI applications (either explicitly military, or dual-use) which are hard to control and might subsequently diffuse to malicious actors. This would echo the experience with intelligence community hacking tools or ‘backdoors’ in digital infrastructure, which were developed and retained because of strategic considerations, but which eventually proliferated into the hands of malicious hacker groups, proving powerful tools for misuse.¹⁹⁴

Likewise, at a more ‘micro’ level, AI applications could drive ‘safety erosion from competition’. This could occur when the pursuit by one party of certain competitive AI applications

¹⁹¹ van der Loeff et al., *supra* note 173 at 3–4.

¹⁹² DAFOE, *supra* note 22 at 7. The term is also discussed in Allan Dafoe, *AI Governance: Opportunity and Theory of Impact*, EA FORUM (2020), <https://forum.effectivealtruism.org/posts/42reWn0TEhFqu6T8/ai-governance-opportunity-and-theory-of-impact> (last visited Sep 20, 2020). (“Just as a safety-performance trade-off, in the presence of intense competition, pushes decision-makers to cut corners on safety, so can a trade-off between any human value and competitive performance incentivize decision makers to sacrifice that value. Contemporary examples of values being eroded by global economic competition could include non-monopolistic markets, privacy, and relative equality”).

¹⁹³ DAFOE, *supra* note 22 at 7.

¹⁹⁴ For instance, in 2013, a group of hackers calling themselves the Shadow Brokers stole a few disks of hacking tools developed by the US National Security Agency (NSA). By sharing these secrets on the internet, the group exposed major security vulnerabilities in internet infrastructures, and moreover put sophisticated cyberweapons widely available, amongst others providing the key exploit underpinning the 2017 global WannaCry ransomware attack, as well as the ‘EternalBlue’ tool. Bruce Schneier, *Who Are the Shadow Brokers?*, THE ATLANTIC, 2017, <https://www.theatlantic.com/technology/archive/2017/05/shadow-brokers/527778/> (last visited Sep 8, 2020); Scott Shane, *Malware Case Is Major Blow for the N.S.A.*, THE NEW YORK TIMES, May 16, 2017, <https://www.nytimes.com/2017/05/16/us/nsa-malware-case-shadow-brokers.html> (last visited Sep 8, 2020). Lily Hay Newman, *How Leaked NSA Spy Tool “EternalBlue” Became a Hacker Favorite*, WIRED, 2018, <https://www.wired.com/story/eternalblue-leaked-nsa-spy-tool-hacked-world/> (last visited Sep 8, 2020). Similarly, one can compare the proliferation of the ‘Blackshades Remote Access Tool’: described as a ‘criminal franchise in a box’, and sold via PayPal for as little as US \$40, Blackshades allowed users without technical skills to effectively deploy ransomware and conduct eavesdropping operations. See John Markoff, *As Artificial Intelligence Evolves, So Does Its Criminal Potential*, THE NEW YORK TIMES, October 23, 2016, <https://www.nytimes.com/2016/10/24/technology/artificial-intelligence-evolves-with-its-criminal-potential.html> (last visited Nov 26, 2018); As well as the discussion of lessons for AI in Hayward and Maas, *supra* note 74 at 10.

compels others parties to rush along development or deployment with less safety precautions, or with less time for safety testing.¹⁹⁵ Indeed, strategic logics could even suggest that certain systems (e.g. cyberwarfare defence uses) could open up avenues for accident cascades. As such, Paul Scharre has argued that while there are ways in which military AI systems could contribute to strategic stability, if they cannot be kept under reliable control their deployment would create situations of “fragile stability”, where crises rapidly escalate out of the control of human commanders, producing what Scharre has called ‘flash wars’.¹⁹⁶

In terms of sociotechnical change and problem logics, AI’s structural impacts are interestingly distinct from the first three categories (‘ethics’, ‘security’, ‘safety’). This is because structure is rarely a first-order challenge or rationale, but instead catalyses many first-order problems in these other areas. It both increases the risks of ethical, safety or security challenges in AI systems themselves, but also highlights their broader societal effects and spillover into apparently distinct domains. As such, structural shifts may indirectly give rise to the full gamut of rationales for regulation and governance discussed previously.

As such, considering *structure* also highlights a distinct perspective on needed regulatory solutions—to address this *target*. Notably, an *ethics* (‘problematic use’) perspective puts the focus on measures that improve the norms of developers, engineers and operators, or which institutionalize oversight mechanisms to hold these parties to account *ex post*. A *security* (‘misuse’) perspective puts focus on changing either the motivations or opportunities for malicious individuals, or on ensuring deterrence through denial, traceability or response. A *safety* (‘accident’) perspective focuses on improving the patience, caution, or competence of the producers, engineers, or operators. By contrast, a *structural* perspective puts less emphasis on various actors’ passive ethical laxness (ethics), active malice (security), or inadvertent neglect (safety). Rather it looks at the effects of technologies on structuring actors’ decisions and choices in ways that generate, enable, or undercut efforts to address first-order ethical, security or safety issues. These indirect effects can be indeliberate, but no or few individual actors have the ability to unilaterally prevent or resolve them, or even to foresee their emergence in the first place. Because they do not derive from any one individual actor’s active intent, structural challenges can be particularly severe to address or govern. In some sense, they emerge from the level of systemic incentives that actors collectively may find themselves locked into.¹⁹⁷ Accordingly, addressing challenges from ‘structure’ often cannot be achieved by targeting individual actors, or by any one actor’s unilateral action. Rather, more than any other types of AI sociotechnical change, addressing structural challenges of AI may require some measure of collective action, especially at the international level. This faces the broad set of problems common to securing ‘mutual restraint’ global public goods. Practically, governance can also be supported or enabled

¹⁹⁵ RICHARD DANZIG, *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority* 40 (2018), <https://www.cnas.org/publications/reports/technology-roulette>.

¹⁹⁶ Paul Scharre, *Autonomous Weapons and Stability*, March, 2020.

¹⁹⁷ See also Dafoe, *supra* note 192. (“When we think about the risks arising from the combustion engine---such as urban sprawl, blitzkrieg offensive warfare, strategic bombers, and climate change---we see that it is hard to fault any one individual or group for negligence or malign intent. It is harder to see a single agent whose behaviour we could change to avert the harm, or a causally proximate opportunity to intervene. Rather, we see that technology can produce social harms, or fail to have its benefits realized, because of a host of structural dynamics. The impacts from technology may be diffuse, uncertain, delayed, and hard to contract over”).

through various transparency- and trust promoting agreements, such as Confidence-Building Measures (CBMs).

4.4.5 Common benefits

The above has been a survey of many mostly threatening developments. However, along with the many risks, it remains important to keep in mind the upsides of AI technology. Indeed, while much of the discussion so far has foregrounded risks more than benefits, this is not to conclude that the emergence of AI technology necessarily produces more harm than good.¹⁹⁸ While it may be the case that AI introduces foundational challenges, it is also possible that the technology is stuck in a ‘technology trap’. This is a phenomenon described by Carl Frey, whereby certain technological transformations which eventually could bring broad prosperity will, in an early stage of development, deliver only meagre results while disproportionately producing societal dislocation. This in turn could create the chance that some technologies become rejected in their early stages, before they have had a chance to ‘prove their worth’.¹⁹⁹ While it remains unclear whether AI is one such technology,²⁰⁰ unless we wish to commit to far-reaching opposition to the technology, we should ensure that its benefits are realised as fully, as widely, and—critically—as soon as possible. This requires considering strategic challenges around realizing or accelerating opportunities for common goods through and in AI. This has three components.

In the first place, as previously noted, there is a need for various forms of global coordination to *avert market failures or inefficiencies*, to achieve the full economic and scientific fruits of AI. Accordingly, some have called for a ‘new technology diplomacy’ to avert the fragmentation of global regimes,²⁰¹ and to limit protectionism or ‘tech nationalism’.

Secondly, there are many areas where AI could be used *pro-actively in support of various social or global goods*.²⁰² Some have proposed it can support of global commons such as the

¹⁹⁸ Indeed, as Parson et al. note, ‘[a]lthough it may be tempting to view any large change as harmful, at least initially, we do not presume that the societal impacts of AI will be entirely, or even predominantly, harmful. In fact, there are strong grounds to anticipate large benefits.’ Edward Parson et al., *Could AI drive transformative social progress? What would this require?*, AI PULSE, 8 (2019), <https://aipulse.org/could-ai-drive-transformative-social-progress-what-would-this-require/> (last visited Sep 28, 2019).

¹⁹⁹ CARL BENEDIKT FREY, THE TECHNOLOGY TRAP: CAPITAL, LABOR, AND POWER IN THE AGE OF AUTOMATION (2019).

²⁰⁰ Interestingly, and more provocatively, there may be certain technologies which could show the reverse of Frey’s technology trap: that is, they produce considerable and appealing upsides in an early stage of development, and so raise little opposition. Yet it is only in their later stages, once societies have grown much more dependent on them, that the technology’s far-reaching downsides begin to crystallize. Fossil fuel industry may well prove one such technology. A more controversial example might be found in spaceflight: the technology has historically been broadly perceived as a lofty ambition, and has produced tremendous spin-off innovations that has benefited many aspects of society. However, Daniel Deudney has argued that on longer time horizons, expansive space exploration opens up a series of new catastrophic threats to humanity. DANIEL DEUDNEY, DARK SKIES: SPACE EXPANSIONISM, PLANETARY GEOPOLITICS, AND THE ENDS OF HUMANITY (2020).

²⁰¹ Claudio Feijoo et al., *Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy*, TELECOMMUN. POLICY 101988 (2020).

²⁰² ITU, *AI for Good Global Summit 2019 Insights* (2019), <https://itu.foleon.com/itu/aiforgood2019/home/> (last visited Oct 6, 2019). See also Alexandra Luccioni & Yoshua Bengio, *On the Morality of Artificial Intelligence*, ARXIV191211945 Cs (2019), <http://arxiv.org/abs/1912.11945> (last visited Jan 20, 2020); Luciano Floridi et al., *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 MINDS MACH. 689–707 (2018).

Sustainable Development Goals²⁰³ or the fight against climate change.²⁰⁴ Others have identified positive uses in disaster relief coordination. For instance, as argued by Heather Roff, in complex humanitarian crises, AI systems could help “examine vast amounts of data relating to available resources, use satellite imagery or images from surveillance aircraft to map affected terrain, as well as find survivors, and thereby estimate the necessary resource requirements given limitations on time, goods and manpower.”²⁰⁵ AI systems also could see use in strengthening international law; they could help monitor and enforce human rights,²⁰⁶ aid diplomatic processes,²⁰⁷ and help resolve and arbitrate international conflicts.²⁰⁸

Thirdly, governance regimes might be necessary to ensure that the considerable benefits of *AI technology are distributed* in a broad and inclusive manner, especially as or if the technology proves increasingly transformative.²⁰⁹

While there are certainly critiques of the various ‘AI for Good’ frameworks,²¹⁰ ensuring that societies seize the positive benefits from AI systems—particularly global public goods, or benefits which would not naturally be supplied by markets—must be a core goal. Much of this would need to confront underlying systemic incentives, such as coordination challenges around cost-sharing or free-riding. This problem logic however foregrounds the role of various governance systems aimed at securing greater public interest regulation.

4.4.6 Governance Disruption

Finally, AI systems may enable new types of behaviour or relations that pose a direct challenge to existing *legal systems*. While many of the above involve challenges for governance, changes to the *legal* processes can be more ambiguous. Some have hailed them as creating

²⁰³ Ricardo Vinuesa et al., *The role of artificial intelligence in achieving the Sustainable Development Goals*, 11 NAT. COMMUN. (2020), <https://www.nature.com/articles/s41467-019-14108-y> (last visited May 8, 2019). But for a cautious approach, see Jon Truby, *Governing Artificial Intelligence to benefit the UN Sustainable Development Goals*, n/a SUSTAIN. DEV. (2020), <https://onlinelibrary.wiley.com/doi/abs/10.1002/sd.2048> (last visited Mar 13, 2020).

²⁰⁴ David Rolnick et al., *Tackling Climate Change with Machine Learning*, ARXIV190605433 CS STAT (2019), <http://arxiv.org/abs/1906.05433> (last visited Jul 23, 2019).

²⁰⁵ Heather M. Roff, *Advancing Human Security Through Artificial Intelligence*, in ARTIFICIAL INTELLIGENCE AND INTERNATIONAL AFFAIRS: DISRUPTION ANTICIPATED, 25 (2018), <https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>.

²⁰⁶ Steven Livingston & Mathias Risso, *The Future Impact of Artificial Intelligence on Humans and Human Rights*, 33 ETHICS INT. AFF. 141–158 (2019).

²⁰⁷ KATHARINA E. HÖNE, *Mapping the challenges and opportunities of artificial intelligence for the conduct of diplomacy* (2019), <https://www.diplomacy.edu/sites/default/files/AI-diplo-report.pdf> (last visited Mar 14, 2019).

²⁰⁸ Ashley Deeks, *High-Tech International Law*, 88 GEORGE WASH. LAW REV. 575–653, 633–637 (2020). This is discussed in more detail in Section 5.5. (on ‘displacement’).

²⁰⁹ For a discussion of this principle of ‘magnanimity’ and ‘broad benefit’ in the context of future, more advanced AI systems, see also Bostrom, Dafoe, and Flynn, *supra* note 4. Elsewhere, Dafoe has referred to this as the problem of ‘constitution design’ for advanced AI Dafoe, *supra* note 192.. For an interesting practical recent application, see also the proposal for a ‘Windfall Clause’, an ‘ex ante commitment by AI firms to donate a significant amount of any eventual extremely large profits [...] that a firm could not earn without achieving fundamental, economically transformative breakthroughs in AI capabilities.’ Cullen O’Keefe et al., *The Windfall Clause: Distributing the Benefits of AI for the Common Good*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 327–331, 327 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375842> (last visited Feb 12, 2020).

²¹⁰ See Jared Moore, *AI for Not Bad*, 2 FRONT. BIG DATA 32 (2019).

opportunities. At the same time, there are concerns over the rule of law. AI can impact on legal systems in three ways. In the first place, it can affect the *substance* of law, requiring a need for new regulation to clarify existing laws or create new laws. In the second place, it can affect the *processes* of law, altering the ways in which those laws are produced, monitored or enforced, with effects that can range from strengthening the rule of law, to instead ‘replacing’ it. In the third place, it can affect and potentially *erode the political scaffolding* of legal systems, risking their efficacy or legitimacy.

AI-driven sociotechnical change that produces governance disruption is a special case. As discussed, in terms of governance *rationale*, it invokes a distinct challenge—not a risk to regulatees, but a potential threat directly to the existing regulatory order. This results in a need for intervention to ensure the continuing or renewed adequacy and accountability of the governance system. This highlights a diverse set of strategies, from those aimed at rendering legal systems more ‘innovation-proof’ (e.g. technological neutrality; the designation of authoritative interpreters; sunset clauses), to those that aim to ensure adequate oversight systems for the deployment of automation in support of legal and governance systems.

Governance disruption is a special case, and remains underexplored in much discussions of AI’s global impact. As such, we will spend significant time at exploring this dynamic in the next chapter. However, we first finish with some reflections on the applications, strengths or limits of the sociotechnical change lens for AI governance.

4.5 Evaluating Sociotechnical Change for AI governance

How useful is the lens of sociotechnical change? As argued by Bennett Moses, switching the conceptual focus from ‘regulating for technology’ to ‘regulating for changes in the sociotechnical landscape’ may help provide valuable principles or insights for regulatory design, choice of institution, as well as for regulatory timing and responsiveness.²¹¹ However, while the lens has strengths as an analytical tool, these should not blind us to its limits and use conditions. As such, we will conclude by briefly reviewing some of these.

4.5.1 Uses: regulatory triage, tailoring, timing, & design

In terms of benefits, we can identify advantages in terms of *triage, tailoring, timing and responsiveness, and regulatory design*.

Firstly, the lens can help in carrying out governance *triage* on AI governance challenges. This model can help focus attention on the most socially disruptive outputs of AI, and as such helps re-focus scarce regulatory attention. In doing so, it can reduce the risk that debate is hijacked by attention-grabbing demonstrations of problems that may not be scalable (or of interest) outside labs. Conversely, it allows one to focus also on widespread but indirect second-order challenges. For instance, beyond enabling an analysis of the direct challenges of AI (in the areas of *ethics, security, and safety*), this framework also enables one to consider three governance ‘problem types’ which remain less widely featured in such debates, such as the ways AI systems

²¹¹ Bennett Moses, *supra* note 28 at 585–591.

can shift incentive *structures*, how to *realize beneficial opportunities*, or how the technology will disrupt the legal systems we would use to regulate it.

Secondly, the lens helps in *tailoring* governance solutions to the challenge at hand. Rather than consign regulators to confront self-similar challenges (around ethics; security; safety, ...) many times in siloed legal domains,²¹² this taxonomy helps highlight common features or themes of such challenges, as they are mirrored across domains wherever specific AI capabilities (e.g. ‘predictive profiling’; ‘computer vision’) are deployed.

Thirdly, in terms of regulatory *timing and responsiveness*, the emphasis on sociotechnical change highlights the inadequacies of governance strategies that are grounded either in an attempt to predict sociotechnical changes in detail, or those reactive responses which prefer to ‘wait out’ technological change. Rather, it emphasises the importance of anticipation and adaptive or ‘innovation-proof’ governance approaches.²¹³

Similarly, and fourthly, in terms of *regulatory design*, the sociotechnical change lens highlights when and why governance should adopt technology-specific rules, and where it might be technology-neutral.²¹⁴ By considering the specific regulatory or governance rationale in play, we may understand when or whether technological neutrality is to be preferred. In this view, the point is not to find a regulatory strategy that is optimal for present-day facial recognition algorithms, or which already outlaws long lists of anticipated future developments. The idea is rather to develop governance tools that are up to the task of managing distinct problem logics—new ethical challenges, security threats, safety risks, structural shifts, opportunities for benefit, or governance disruptions—in a way that can be relatively transferable across- or agnostic to the specific AI techniques used to achieve those affects.

4.5.2 Limits: operationalisability, scale, prediction

Nonetheless, the sociotechnical change lens is not an analytical ‘silver bullet’ that improves any AI governance analysis or initiative anywhere. There are certainly limits to this approach, in terms of operationalisability, scalability, or prediction.

In the first place, it may not be easily *operationalisable*. Considerations of ‘sociotechnical change’ can be abstract, especially in comparison to analyses that focus on concrete use cases of AI in specific sectors (e.g. facial recognition). As such, this framework may not always allow for an easy translation or aggregation of diverse AI capabilities into a general ‘rationale’ (e.g. ‘market failure’), nor allow a straightforward mapping of challenges. In many practical contexts, policymakers require concrete, intuitive, and stable categories and definitions.²¹⁵ However, the point of the taxonomy is not to provide a new blueprint for (international) regulation structured along ‘AI capability problem-logics’ rather than ‘AI applications’. The point is simply to provide a thinking tool for disaggregating the spectrum of a particular AI capability’s sociotechnical impact.

²¹² See also Crootof and Ard, *supra* note 1 at 1. (“The fundamental challenge of techlaw is not how to best regulate novel technologies, but rather how to best address familiar forms of uncertainty in new contexts”).

²¹³ Cf. Maas, *supra* note 183.

²¹⁴ Bennett Moses, *supra* note 28 at 586. (“regulatory regimes should be technology- neutral to the extent that the regulatory rationale is similarly neutral”).

²¹⁵ Schuett, *supra* note 85; TURNER, *supra* note 3 at 8–9.

In the second place, it might be objected that this lens, even if useful at the level of national AI policy, is not *scalable* (or transferable) to the global stage. To be certain, the concept and scholarship on ‘sociotechnical change’ has been mostly derived from various domestic law contexts. As such, some of the rationales for regulation may not transfer directly or fully to the international level. Yet while it might be true that not all of these rationales for regulation are explicitly recognised in international law, they can still serve as policy goals or global public goods to be secured.

In the third place, this lens does not overcome the general epistemic limits around *prediction*. It allows us to interpret past or clearly ongoing societal impacts by AI technologies, and to speculate in structured ways about anticipated future sociotechnical impacts of certain capabilities. Yet it would be presumptuous to expect that this approach is able to avoid the deep challenges around prediction. In fact, it may be even harder to anticipate sociotechnical change than ‘merely’ predicting technological trends. This is simply to say that this lens does not necessarily do better at forecasting issues than many existing approaches. But then it is not meant to. The point is perhaps less to enable foresight and anticipation, and instead to support forms of resilience or adaptability.

In sum, the lens of sociotechnical change should be considered not a roadmap for AI governance, but a thinking tool or compass. The distinction between six types of sociotechnical changes created by AI is not meant to be exhaustive, predictive, or prescriptive in terms of regulatory responses. Instead, it is meant as a tool for organizing thought processes around governing AI in diverse sectors. It is aimed at enabling a structured consideration of key steps, such as: (1) when, why and how a given AI capability gives rise to sociotechnical changes; (2) When and why these changes rise to create a rationale for governance; (3) How to approach the target of regulation, in a way that reckons not just with the material features of the technology, but also with common problem structures or logics. In this way, it is hoped, this lens might help provide loose heuristics or templates for identifying and thinking through patterns of AI challenges and response strategies.

4.6 Conclusion: AI governance for sociotechnical change

This chapter has sketched the first of three perspectives on ‘change’ in AI governance. It explored how, when, and why law and regulation for AI ought to tailor themselves to broad *sociotechnical change* more than local *technological change*. Accordingly, global AI governance regimes should, first, focus less on isolated AI capabilities or use cases in isolated domains, and rather should depart from an understanding of when and how shifts in AI capabilities drive cross-domain patterns of socio-technical change. Second, governance approaches for AI should seek to understand how socio-technical changes created by AI applications can be disambiguated into specific types of challenges—ethical challenges, security threats, safety risks, structural shifts, common benefits, and governance disruption—each of which can come with distinct problem features, and may demand different governance logics.

Of course, the emphasis on sociotechnical change is not a new insight within the scholarship on law, regulation and new technologies. However, in a fragmented AI governance

landscape, it remains a valuable corrective to approaches that select, organize or prioritize AI policy issues based on high-profile but non-representative incidents, popular-cultural resonance, or ‘fit’ to pre-existing legal domains. In sum, ‘sociotechnical change’ should be considered not an entirely new paradigm for AI governance, but rather a complementary perspective. Such a lens is subject to its own conditions and limits, but when used cautiously, can offer a fuller and more considered mapping of which challenges are possible, plausible, or already-pervasive—and how these might be best met.

Finally, of the diverse types of sociotechnical change produced by AI systems, one which stood out was ‘governance disruption’. This is a key dynamic, which is particularly important to understanding changes in the tools of governance. Accordingly, we next turn to this lens.

Chapter 5. Technology-Driven Legal Change: AI as Governance Disruptor

The governance disruption lens explores how and why the use of AI capabilities might produce intended or unintended change in the global legal order itself. This discussion proceeds as follows: after introducing the exploration of ‘governance disruption’ at the international legal level, I (5.1) provide an abridged review of the historical connection between technological innovation and global legal change, and (5.2) sketch the governance disruption framework. I then explore the three varieties of disruption. Under (5.3) *Development*, the sociotechnical change produced by AI capabilities results in situations of substantive legal uncertainty or ambiguity. This creates a need for change in governance in order to resolve these tensions, and ensure the international legal system can remain fit to address the challenges. Under (5.4) *Displacement*, AI technologies are used to support, or to substitute, key processes of the global legal order, with distinct effects. In cases of (5.5) *Destruction*, AI affects and erodes the effectiveness and coherence of the global legal order, either because it proves conceptually or politically incapable of carrying through certain required developments, or because AI applications directly or indirectly erode the political scaffolding of international regimes. Finally, (5.6) I conclude.

So far, we have discussed AI as *site* of change, and—more importantly—as a *source* of sociotechnical change. As noted, such a lens is valuable in framing AI governance regimes, because it enables a more systematic comparison and prioritisation of which AI issues rise to create a most urgent *rationales* for governance, and because it then allows regimes to consider the texture of these uses as distinct *targets* for governance.

There is, however, one special case of AI-produced sociotechnical change: this is governance disruption. This concerns situations where AI capabilities creates new types of behaviour, entities, or relations which affect not just (global) society, but also or especially impact on the concepts and tools of governance itself. Examining the effects that emerging AI capabilities could exert on the global governance architecture surely matters for understanding the trajectories of component regimes within that system, including regimes which are in turn focused on the regulation of AI itself (see Figure 5.1).

However, there remains only a relatively small and fragmented literature exploring how AI will drive changes in the fabric or formation of international law and global governance. There are a few exceptions, however, which this chapter will draw on.

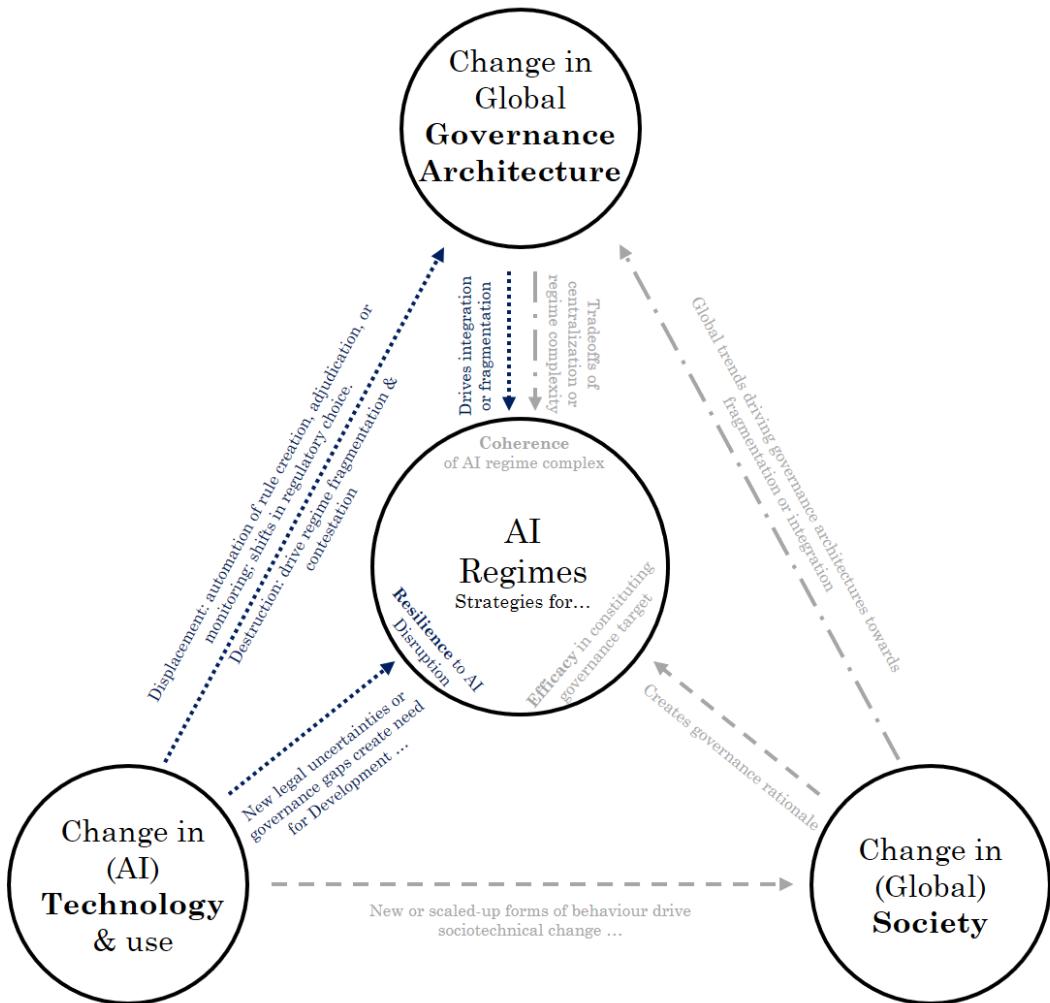


Figure 5.1. Conceptual sketch: Governance Disruption

In the first place, Thomas Burri, Berenice Boutin, and Ashley Deeks have provided general treatments of the ways in which AI technology could affect or play a role in international law.¹ Megiddo has explored the potential contribution of big data to customary international law.² Alschner and colleagues have explored the production of international legal materials (such as draft treaties) by algorithms,³ and have explore some of the implications of the ‘data-driven future’

¹ Thomas Burri, *International Law and Artificial Intelligence*, 60 GER. YEARB. INT. LAW 91–108 (2017); Berenice Boutin, *Technologies for International Law & International Law for Technologies*, GRONINGEN JOURNAL OF INTERNATIONAL LAW (2018), <https://grojil.org/2018/10/22/technologies-for-international-law-international-law-for-technologies/> (last visited Oct 31, 2018); Ashley Deeks, *High-Tech International Law*, 88 GEORGE WASH. LAW REV. 575–653 (2020); Ashley Deeks, *Introduction to the Symposium: How Will Artificial Intelligence Affect International Law?*, 114 AJIL UNBOUND 138–140 (2020).

² Tamar Megiddo, *Knowledge Production, Big Data and Data-Driven Customary International Law* (2019), <https://papers.ssrn.com/abstract=3497477> (last visited Jan 21, 2020).

³ Wolfgang Alschner & Dmitriy Skougarevskiy, *Towards an automated production of legal texts using recurrent neural networks*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTICIAL INTELLIGENCE AND LAW 229–232 (2017), <https://doi.org/10.1145/3086512.3086536> (last visited May 14, 2020); Wolfgang Alschner & Dmitriy Skougarevskiy, *Can Robots Write Treaties? Using Recurrent Neural Networks to*

of the (text-corpus-rich) international economic law.⁴ Deeks and others explore the specific implications of conflict prediction algorithms on the international law on the use of force.⁵ Livingston & Risso have discussed the implications of AI for potential monitoring of human rights.⁶ More generally, Nelson has evaluated the effects of digitisation (including the spread of AI) on existing arms control and export control regimes.⁷

However, this literature is still small compared to the broader body of work that has examined the phenomena of 'law and technology' and 'legal automation' in a domestic law context. Over the past decade, such scholarship has studied, through both concrete cases and at a conceptual level, how the use of algorithms and AI systems might alter the coherence, form, and practices of legal systems, or even the underlying values of regulators.⁸

Significantly, if one expects AI technologies to have such a disruptive impact on domestic legal systems, one should arguably expect their disruptive effects to be even steeper in the international domain. This is because in comparison to national law, the global legal order appears less well equipped to anticipate or resolve situations of legal uncertainty, or to keep accountable the insertion of technology in legal processes. After all, at the international level there is no authoritative final legislator to compel *ex ante* consideration of a technology, to guarantee the coherence of legal responses across domains, or to coordinate (and critically scrutinize) the integration of new technologies into law-making, -adjudication, or -enforcement practices. Moreover, amending or altering multilateral treaty regimes to take stock of subsequent technological developments can often prove a more painful, drawn-out, and contested affair than the revision of domestic laws.⁹ Furthermore, governance disruption can pose a challenge that is not just legal or procedural, but also political. After all, while international relations scholars are

Draft International Investment Agreements, in JURIX: LEGAL KNOWLEDGE AND INFORMATION SYSTEMS 114–119 (F. Bex & S. Villata eds., 2016), <https://papers.ssrn.com/abstract=2984935> (last visited May 14, 2020). As also discussed in Deeks, *supra* note 1 at 604–606, 620.

⁴ Wolfgang Alschner, Joost Pauwelyn & Sergio Puig, *The Data-Driven Future of International Economic Law*, 20 J. INT. ECON. LAW 217–231 (2017).

⁵ Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NATL. SECUR. LAW POLICY 1–25 (2019).

⁶ Steven Livingston & Mathias Risso, *The Future Impact of Artificial Intelligence on Humans and Human Rights*, 33 ETHICS INT. AFF. 141–158 (2019).

⁷ AMY J. NELSON, *Innovation Acceleration, Digitization, and the Arms Control Imperative* (2019), <https://papers.ssrn.com/abstract=3382956> (last visited May 29, 2020).

⁸ The literature is vast, and no exhaustive review is here attempted. However, some contributions over the past years include: Roger Brownsword, *Technological management and the Rule of Law*, 8 LAW INNOV. TECHNOL. 100–140 (2016); Benjamin Alarie, Anthony Niblett & Albert H. Yoon, *Law in the future*, UNIV. TOR. LAW J. (2016), <https://www.utpjournals.press/doi/abs/10.3138/UTLJ.4005> (last visited Jan 28, 2019); Benjamin Alarie, *The path of the law: Towards legal singularity*, 66 UNIV. TOR. LAW J. 443–455 (2016); Anthony J Casey & Anthony Niblett, *Self-driving laws*, 66 UNIV. TOR. LAW J. 429–442 (2016); Brian Sheppard, *Warming up to inscrutability: How technology could challenge our concept of law*, 68 UNIV. TOR. LAW J. 36–62 (2018); Woodrow Hartzog et al., *Inefficiently Automated Law Enforcement*, 2015 MICH. STATE LAW REV. 1763 (2016); Rebecca Crootof, “*Cyborg Justice*” and the Risk of Technological-Legal Lock-In, COLUMBIA LAW REV. FORUM (2019), <https://papers.ssrn.com/abstract=3464724> (last visited Nov 18, 2019); John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 PHILOS. TECHNOL. 245–268 (2016); Karen Yeung, ‘*Hypernudge*’: Big Data as a mode of regulation by design, 20 INF. COMMUN. SOC. 118–136 (2017).

⁹ Rebecca Crootof, *Jurisprudential Space Junk: Treaties and New Technologies*, in RESOLVING CONFLICTS IN THE LAW 106–129 (Chiara Giorgetti & Natalie Klein eds., 2019), <https://brill.com/view/book/edcoll/9789004316539/BP000015.xml> (last visited Mar 15, 2019). See also section 5.3.6 (on carrying out development).

still weighing the precise impact of AI,¹⁰ there is a general appreciation that new innovations can redraw the global political map.¹¹ It would be very surprising if such far-reaching (geo)-political shocks did not have their echoes at the international legal level.

As such, this chapter will explore the dynamics of AI-driven governance disruption at the global level. I (5.1) review the history of technological innovation and global legal change, and (5.2) sketch the governance disruption framework, after which I discuss the dynamics of AI-driven (5.3) development, (5.4) displacement, and (5.5) Destruction.

5.1 International law and technology: an abridged history

While the ‘governance disruption’ impacts of AI at the global level remain relatively understudied, that is not to say there is no basis to work from. In fact, legal scholars have elsewhere examined in detail the ways in which other technological innovations have historically spurred or shaped the development of international law. For instance, Colin Picker has reviewed the interrelation between technological change and change in international law, tracking these dynamics across diverse historical eras (from the pre-modern era to the birth of modern international law to the 21st Century), and across substantive domains (ranging from space law to the law of the sea, and from the international law of fisheries to the nuclear non-proliferation regime). On this basis, he notes how:

“Technology changes international law in a number of ways, by forcing states to either: (1) agree to modify their behaviour (usually through the device of a treaty); (2) abandon previously agreed behaviour (abandonment of treaties); (3) abandon the effort to agree on new behaviour (abandonment of treaty formation); (4) engage in new practices that eventually are accepted by the global community (creating customary international law); (5) abandon previously widely accepted customs (abandonment of customary international law); or (6) accept peremptory obligations (creating *ius cogens*).”¹²

This pattern appears to be the case for many technologies, though it appears to be especially strong around military innovation. After all, the management or control of new military technologies has been a core component throughout the history of international law and global governance, spurring landmark legal innovations and change.¹³

¹⁰ Michael C. Horowitz, *Artificial Intelligence, International Competition, and the Balance of Power*, TEXAS NATIONAL SECURITY REVIEW, 2018, <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/> (last visited May 17, 2018); Michael C. Horowitz, *Do Emerging Military Technologies Matter for International Politics?*, 23 ANNU. REV. POLIT. SCI. 385–400 (2020).

¹¹ Daniel W Drezner, *Technological change and international relations*, 33 INT. RELAT. 286–303, 287 (2019). (“[a]ny technological change is also an exercise in redistribution. It can create new winners and losers, alter actor preferences, and allow the strategic construction of new norms”).

¹² Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDOZO LAW REV. 151–219, 156 (2001).

¹³ Picker, *supra* note 12; Rosemary Rayfuse, *Public International Law and the Regulation of Emerging Technologies*, in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY, 502 (2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22> (last visited Jan 3, 2019). See also generally Matthijs M. Maas, *International Law Does Not Compute: Artificial*

Critically, many of these cases show how new technologies can alter international law directly—by serving as a new concern or object of existing governance arrangements—but also indirectly, through far-reaching second- and third-order cultural, social, and political shifts.¹⁴ The resulting legal innovations in response to these technologies’ sociotechnical changes have at times extend far beyond the narrow remit of the particular crises or wars that ostensibly sparked them. Inventions such as cannons have as such induced not just positional shifts in relative power, but also structural changes in the logistics, scale, and economy of military activity, for instance, having contributed to the rise of nation-states.¹⁵ Given its generality, conceptual complexity, and its ability to be integrated within existing weapons systems, AI may stake a claim to shaping the next chapter in this core governance domain.¹⁶

Likewise, Rebecca Crootof has discussed a general pattern by which new technological capabilities can at times enable or create new patterns of state behaviour which, once widespread, create new customary international law which, under the *lex posterior* principle, functionally supersede any older treaty commitments.¹⁷ This leaves into its wake ‘jurisprudential space junk’: collections of fragmented, hard-to-amend treaty regimes which “are theoretically in force but actually simply clutter and confuse the relevant legal regime.”¹⁸ As one illustration of this, Crootof discusses the evolution of customary norms around submarine warfare, and how these eroded and undercut previous treaty regimes.¹⁹ Specifically, the 1930 London Naval Treaty and the 1936

Intelligence and The Development, Displacement or Destruction of the Global Legal Order, 20 MELB. J. INT. LAW 29–56 (2019). Paper [III].

¹⁴ See also Matthijs M. Maas, *Innovation-Proof Governance for Military AI? How I learned to stop worrying and love the bot*, 10 J. INT. HUMANIT. LEG. STUD. 129–157 (2019). Paper [II].

¹⁵ As Allenby claims; “[c]hariots, which required supporting personnel and money to operate, privileged aristocratic warriors; stirrups elevated the mounted horseman and made the steppe warrior with his composite bow feared for centuries; the rise of gunpowder and cannon profoundly changed the logistics necessary to field an army, and thereby created economies of scale in military activities that encouraged the rise of nation-states”. Braden Allenby, *Are new technologies undermining the laws of war?*, 70 BULL. AT. SCI. 21–31, 21 (2014). However, while these are popular accounts, it should be noted that the military role of technological supremacy in history—at least during the pre-modern era, and potentially since—remains highly contested. See for instance: Kelly DeVries, *Catapults are Still not Atomic Bombs: Effectiveness and Determinism in Premodern Military Technology*, 7 VULCAN 34–44 (2019); David Zimmerman, *Neither Catapults nor Atomic Bombs: Technological Determinism and Military History from a Post-Industrial Revolution Perspective*, 7 VULCAN 45–61 (2019).

¹⁶ Note that Braden Allenby is more sceptical about the novelty of LAWS with regards to the challenge they pose to the laws of war, arguing that this question mostly comes down to their ability to discriminate between targets (as well as or better than human soldiers), which he holds to be a factual question. Allenby, *supra* note 15 at 27. Conversely, he argues that cyber weapons, by mixing together civilian and military activity, pose a far larger conceptual and doctrinal challenge. *Id.* at 27. Without taking issue with this characterization of cyber weapons, I suggest that ultimately military uses of AI technology (a far broader category than LAWS) are likely to generate a wider range of issues for more distinct regimes and norms, than a narrow focus on LAWS’ ability to discriminate would suggest. Maas, *supra* note 14. See also Ben Koppelman, *How Would Future Autonomous Weapon Systems Challenge Current Governance Norms?*, 164 RUSI J. 98–109 (2019).

¹⁷ Crootof, *supra* note 9 at 113; See also Rebecca Crootof, *Change Without Consent: How Customary International Law Modifies Treaties*, 41 YALE J. INT. LAW 65, 284–288 (2016). (“*Lex posterior derogat legi priori* is the principle that the most recently developed rule prevails over prior law”).

¹⁸ Crootof, *supra* note 9 at 107.

¹⁹ In Paper [III], I also illustrated these disruptive effects through a related discussion of how submarine warfare upset the assumptions of the ‘International Prize Court’, proposed in 1907 to hear cases regarding the capturing of prizes (mostly ships) during wartime. This was the first treaty proposal for a truly international court, and while it soon proved moribund for political reasons, rapid technological change in submarine warfare would have made its mission and role moot before long. Maas, *supra* note 13 at 35–36. See generally Janet Manson,

Limitation of Naval Armament Treaty held that submarines were not distinct from surface warships, and as such were also bound by the prohibition against attacking enemy merchant vessels without first ensuring the safety of their sailors. However, Crootof argues that widespread state practice in violation of these norms, especially during the Second World War, functionally created a new and less restrictive customary international law of submarine warfare.²⁰ This was in part because a range of ‘architectural’ tactical and logistical constraints on submarine operation came to be far more determinative of actual practices during wartime than the treaty instruments that were nominally in force. As a result, this new customary norm functionally replaced the older treaty law in international law, and effectively expanded states’ rights regarding the lawful use of submarines, rendering the previous treaties ‘dead letter’.²¹

More generally, we can distinguish how new technologies or innovations can shape international law in a variety of ways. New innovations can affect international law *doctrinally*, because the sociotechnical change they enable or create creates a new rationale for regulating certain ‘targets’ (artefacts or capabilities). Technologies can also change the *processes* by which international law is created or enforced. In more extreme cases, new technologies can change international law *politically*, by ‘changing the stakes’ of previously more moderate problems, in a way that provides the impetus for key broader legal innovations, or undercuts the political scaffolding of previous orders.

This illustrates how in extreme cases, new technologies can pose a ‘full-spectrum’ challenge—or opportunity—to international law, impacting at not only the legal or normative level, but simultaneously at the operational and political ones. Given the extreme versatility of AI technology, I argue that AI systems will be similarly likely to affect the doctrine, politics, process, enforcement and efficacy of international law—and with it, the likely texture and trajectory of regimes and regime complexes within it. In order to better chart and analyse such effects in the case of AI, this chapter therefore sets out a taxonomy of governance disruption.²²

5.2 An Overview of AI Governance Disruption

Paper [III] sketched an analytical framework through which to explore AI’s effects on the general international law system. Specifically, drawing together several concepts and frameworks, I distinguished three types of possible global legal impacts of AI: ‘development’, ‘displacement’, and ‘destruction’. These are sketched below (see Table 5.1).²³

International law, German Submarines and American Policy, 1977, <http://archives.pdx.edu/ds/psu/15921> (last visited Oct 23, 2018).

²⁰ LIMITATION AND REDUCTION OF NAVAL ARMAMENT (LONDON NAVAL TREATY), 112 LNTS 65 (1930); LIMITATION OF NAVAL ARMAMENT (SECOND LONDON NAVAL TREATY), 197 LNTS 387 (1937). See also Crootof, *supra* note 9 at 113–114.

²¹ Crootof, *supra* note 9 at 114.

²² Maas, *supra* note 13. In other recent work, we have called this ‘legal disruption’. Hin-Yan Liu et al., *Artificial intelligence and legal disruption: a new model for analysis*, 0 LAW INNOV. TECHNOL. 1–54 (2020). However, as this aims to consider the broader context of governance, ‘governance disruption’ appears more suitable.

²³ It should be noted that many of the examples presented in the sections below should be considered not as predictions, but as conditional examples or thought experiments that highlight the legal effects or implications of various scenarios. See also Section 3.3.2.

	Type	Example	Outcome
Need for Development	New governance gaps	AI-enabled swarm warfare (possibly) not covered by existing international regimes	Need for new law.
	Conceptual uncertainty or ambiguity	LAWS highlight potential ambiguity or inadequacy of concepts such as 'intent', 'effective control', etc.	Need for new law or adaptation of law, to sharpen existing rules or clarify concepts.
	Incorrect scope of application	Underinclusive application of Convention Against Torture to use of autonomous robots for interrogation.	Need for new law or adaptation of law, to demarcate scope and applicability of existing instruments.
		Overinclusive applicability of company law enabling incorporation of 'algorithmic entities' with corporate legal personhood.	
	Obsolescence	Behaviour obsolete (necessity)	New types of AI-supported remote biometric surveillance (gait or heartbeat identification) replace face recognition.
		Justifying assumptions no longer valid (adequacy)	Structural unemployability through technological unemployment puts pressure on right to work, ILO regimes.
		No longer cost-effective (enforceability)	Use of DeepFakes or computational propaganda raises monitoring and compliance enforcement costs for various regimes.
	Altered problem portfolio		Military AI regime tailored to respond to <i>ethical challenges</i> of LAWS (e.g. maintaining meaningful human control over lethal force) might not be oriented to address risks of later adjacent AI capabilities (e.g. cyberwarfare) creating <i>structural shifts</i> .
	Automation	International Law Creation & Adjudication	Use of AI text-as-data tools to generate draft treaties, predict arbitral panel rulings, identify state practice, identify treaty conflicts.
		Monitoring & enforcement	Use of various AI tools in monitoring and verifying treaty compliance.
Displacement	Replacement	Changes in regulatory modality	Use of AI tools such as emotion-recognition, social media sentiment analysis, or computational propaganda by states, resulting in increased state preference to resolve disputes in diplomatic channels.
		Conceptual friction	Attempted extension of existing regimes or norms to new technology cannot pass 'laugh test'.
	Erosion	Political 'knots'	Attempted extension of existing regimes or creation of new law, intractable because of political gridlock.
		Increasing the spoils of noncompliance	Innovations increase strategic stakes or ability to bypass monitoring, or lower proliferation thresholds or (political) noncompliance costs.
Destruction	Decline	Active weapon	AI-enabled computational propaganda enables contestation of international law; Suspected use of AI negotiation tools subverts legitimacy of resulting agreements.
		Shift of values	AI capabilities perceived as enabling unilateralism, alternative to multilateralism
			Additional pressure on global legal order

Table 5.1. A Governance Disruption Framework

5.3 Development

New uses of AI can yield sociotechnical changes which are perceived to no longer be adequately covered or addressed within existing international frameworks or regimes, either legally or operationally. This creates a need for legal or regulatory 'development' in existing areas or doctrines of international law.

Of course, AI technologies are certainly not so anathema to (international) law that they confuse or challenge it at every point of contact. Nonetheless, new AI capabilities can, in some cases, create (or reveal) problematic or uncertain legal situations—situations which pose an open question for existing rules, norms or standards in global governance, especially in hard law regimes. The idea that disruptive patterns of sociotechnical change can drive a need for legal 'development' subsumes (or expands on) previous theoretical accounts of the relationship between technology and new law. For instance, David Friedman previously argued that technological change affects law in three ways;

“(1) by altering the cost of violating and enforcing existing legal rules; (2) by altering the underlying facts that justify legal rules; and (3) by changing the underlying facts implicitly assumed by the law, making existing legal concepts and categories obsolete, even meaningless.”²⁴

A more granular taxonomy of law and technological change has been proposed by Lyria Bennett Moses.²⁵ She argues that while not every technology creates the occasion or need for new litigation or legal scholarship, technological change does often create a ‘recurring dilemma’ for law, by creating new entities or enabling new behaviour.²⁶ In her analysis, this creates four distinct types of new legal situations which call for legal change (1) a need for new, special laws; (2) legal uncertainty; (3) incorrect scope in the form of the under- or over-inclusiveness of laws in new contexts; and (4) obsolescence of some laws as a result of technology reducing the importance of conduct, undermining the justification for certain rules, or reducing the cost-effectiveness of enforcement.²⁷

²⁴ David D Friedman, *Does Technology Require New Law?*, 71 PUBLIC POLICY 16, 71 (2001).

²⁵ Lyria Bennett Moses, *Why Have a Theory of Law and Technological Change?*, 8 MINN. J. LAW SCI. TECHNOL. 589-606., 590, 605–606 (2007). See also Lyria Bennett Moses, *Recurring Dilemmas: The Law’s Race to Keep Up With Technological Change*, 21 UNIV. NEW SOUTH WALES FAC. LAW RES. SER. (2007), <http://www.austlii.edu.au/au/journals/UNSWLRS/2007/21.html> (last visited Jul 3, 2018).

²⁶ Bennett Moses, *supra* note 25 at 4–5.

²⁷ *Id.* at 8. Note, beyond my taxonomy of ‘development’ in Paper [III] and in this chapter, Bennett Moses’s framework has also inspired other analyses of legal disruption at the international level. For instance, Rebecca Crootof has articulated a lucid framework to chart international legal disruption produced by new weapons technology, similarly drawing on Bennett Moses’s framework. Rebecca Crootof, *Regulating New Weapons Technology*, in THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 1–25, 6 (Eric Talbot Jensen & Ronald T.P. Alcala eds., 2019). In Crootof’s version, a new technology can be legally disruptive because it can “(A) alter how rules are created or how law is used; (B) make more salient an ongoing but unresolved issue, gap, or contradiction in the existing rules; (C) introduce new uncertainty, usually regarding the application or scope of existing rules; and (D) upend foundational tenets of a legal regime, necessitating a reconceptualization of its aims and purposes.” *Id.* at 6. Notably, the latter three of these categories map loosely onto the various categories in my analysis of development, whereas ‘Significantly Changing How Law is Created or Used’ maps somewhat to this project’s discussion of displacement.

Of course, as emphasised by the discussion around patterns of sociotechnical change, the precise forms such disruptions may take are very hard to anticipate or predict.²⁸ Indeed, this may be almost intrinsic, for if we could easily anticipate these disruptions, this would mean that we could already discern the fault lines (gaps or ambiguities) in our existing laws, suggesting that these could be fixed ahead of time. Nonetheless, we can discuss illustrative examples of general patterns that match the categories of legal development above. As such, the next few sections will discuss how new AI-enabled behaviours, situations or sociotechnical change might create conditions for legal uncertainty, in the form of (5.3.1) new governance gaps; (5.3.2) conceptual uncertainty and ambiguity; (5.3.3) existing laws that are wrongly scoped, in that they are revealed as inappropriately under- or over-inclusive; (5.3.4) the obsolescence of core assumptions of the governance regime; or (5.3.5) an altering of the problem portfolio beyond the original scope or mandate of early regimes. Finally, we will discuss (5.3.6) nuances and challenges around when such situations become truly ‘disruptive’, and when or how existing governance systems are able to carry out the required governance development to address these situations.

5.3.1 New Governance Gaps

In the first place, a new AI application could simply highlight important *new gaps* within the existing tapestry of international legal instruments.²⁹ This could be the case if the technology, or its uses, are perceived to create problematic sociotechnical changes which entirely and clearly falls outside of the scope of any existing regimes or norms.³⁰ It is important to note that this is not inherent, but rather a matter of perception, depending on whether enough parties perceive that (everyone else also perceives that) existing laws cannot be ‘stretched’ to cover the new case.³¹ If this is so, this situation creates a need for new *sui generis* rules to deal with new situations or forms of conduct, or to ban a particular underlying technology or particular applications.

Historically, various technologies have given rise to a perceived shortfall of existing legal frameworks, resulting in the development of entirely new regimes, such as the emergence of outer

²⁸ Gregory N. Mandel, *Legal Evolution in Response to Technological Change*, OXF. HANDB. LAW REGUL. TECHNOL. (2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-45> (last visited Sep 26, 2018).

²⁹ Of course, the question of whether newly discovered legal gaps are ‘new’ gaps, or always-latent ambiguities or shortfalls in the law, is philosophically interesting, but is a bit of a rabbit hole. On the epistemology and ontology of which, see generally Roberto Casati & Achille Varzi, *Holes*, in THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., Summer 2019 ed. 2019), <https://plato.stanford.edu/archives/sum2019/entries/holes/> (last visited Sep 8, 2020).

³⁰ Crootof, *supra* note 27 at 20. (“[n]ew technology can sometimes permit entirely new forms of conduct or generate new negative externalities that cannot be addressed through applying or revising the existing law”).

³¹ This does not mean that states always seek to work within existing regimes, and only seek the creation of new (treaty) regimes when they truly believe they are unable to house issues within existing international law. For instance, there has been exploration of diverse circumstances under which states may create overlapping or contesting institutions on similar issue areas. Johannes Urpelainen & Thijs Van de Graaf, *Your Place or Mine? Institutional Capture and the Creation of Overlapping International Institutions*, 45 BR. J. POLIT. SCI. 799–827 (2015). Likewise, Surabhi Ranganathan has provided three case studies (on seabed mining, the International Criminal Court, and nuclear governance) of instances where various states engineered treaty conflicts in order to achieve desired changes in multilateral regimes. SURABHI RANGANATHAN, STRATEGICALLY CREATED TREATY CONFLICTS AND THE POLITICS OF INTERNATIONAL LAW (2014), <https://www.cambridge.org/core/books/strategically-created-treaty-conflicts-and-the-politics-of-international-law/55EACC81A929FC19216E3380D2E9DF69> (last visited Jun 18, 2020).

space law or arms control regimes aimed at various weapons of mass destruction,³² to name but a few. While the creation of new technology-specific treaty regimes has not always been politically easy, the international legal system is at least in principle clearly capable of proposing and promulgating new legal regimes in order to address the ‘gaps’ opened up by certain new technologies.³³

In what ways could AI create new governance gaps? We could imagine various AI capabilities that would enable new, morally problematic or politically or strategically disruptive forms of conduct that might be seen as problematic (that is, constitute a governance rationale), yet simultaneously as clearly out of scope of existing (international) law. This could include, *inter alia*, the systematic monitoring and social control of populations through enhanced surveillance; the deployment of fully autonomous weapons in swarm configuration;³⁴ (cyber)warfare systems that might be operationally susceptible to emergent accidents;³⁵ or the use of AI in tracking of rival nuclear assets in ways that threaten deterrence stability.³⁶ Such capabilities create forms of global sociotechnical change which may be considered as a problem (creating a governance rationale) by most or all of the parties involved. If these parties also perceive that the change is insufficiently addressed by existing regimes, this creates a perceived need—if not the conditions—for new legal developments (most commonly through treaties) to explicitly ban or control the deployment or use of these systems.

Curiously, there are several general reasons for expecting that AI-produced ‘new governance gaps’ might prove comparatively rare—or at least rarer than one might expect. This is for both general historical reasons and specific technological ones. Generally, whereas earlier technological breakthroughs such as spaceflight and nuclear weapons occurred in a relatively young and sparse global institutional landscape, with clear thematic space to grow, the pattern of global ‘institutional proliferation’ over the last decades has resulted in a more tightly populated and thematically diverse (if not exhaustive) governance landscape. This ensures that rather than

³² Picker, *supra* note 12 at 170–172, 175–178.

³³ See generally Rayfuse, *supra* note 13.

³⁴ For a discussion, see Paul Scharre, *How swarming will change warfare*, 74 BULL. AT. SCI. 385–389 (2018); Irving Lachow, *The upside and downside of swarming drones*, 73 BULL. AT. SCI. 96–101 (2017); Zachary Kallenborn & Philipp C. Bleek, *Swarming destruction: drone swarms and chemical, biological, radiological, and nuclear weapons*, 25 NONPROLIFERATION REV. 523–543 (2018). For a discussion of control approaches, see MAAIKE VERBRUGGEN, *The Question of Swarms Control: Challenges to Ensuring Human Control over Military Swarms* 16 (2019), https://www.nonproliferation.eu/wp-content/uploads/2019/12/EUNPDC_no-65_031219.pdf. See also Nathan Leys, *Autonomous Weapon Systems, International Crises, and Anticipatory Self-Defense*, 45 YALE J. INT. LAW 377–411 (2020). (setting out ‘rules of the road’). Notably, Zachary Kallenborn has suggested drone swarms could be classified as Weapons of Mass Destruction, in which case the Seabed Treaty and the Outer Space Treaty would ban the deployment of such systems in these ‘commons’. ZACHARY KALLENBORN, *Are Drone Swarms Weapons of Mass Destruction?* 5 (2020), <https://media.defense.gov/2020/Jun/29/2002331131/-1-1/0/60DRONESWARMS-MONOGRAPH.PDF> (last visited Sep 9, 2020).

³⁵ See the discussion of ‘normal accidents’ in Section 4.4.2. As well as in Paper [I]: Matthijs M. Maas, *How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons*, 40 CONTEMP. SECUR. POLICY 285–311 (2019).

³⁶ VINCENT BOULANIN ET AL., *Artificial Intelligence, Strategic Stability and Nuclear Risk* 158 (2020), https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf. Michael C. Horowitz, Paul Scharre & Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence*, ARXIV191205291 CS (2019), <http://arxiv.org/abs/1912.05291> (last visited Dec 18, 2019).

creating a new encompassing regime, global governance initiatives might often default to accommodating new issues within an existing ‘regime complex’.³⁷

In the second place, while much of the general discourse and hype around AI may speak to the technology’s inherent ‘novelty’ (and the according need for new legal scholarship), there are also features of the technology that counter this. Most importantly, whereas previous technological breakthroughs more unambiguously and categorically broached ‘new ground’ (e.g. space; cyberspace) where no previous technology had reached before, AI technology faces many contexts where it may not broach such a similarly untouched domain. Paradoxically, the very ‘generality’ of AI systems, and their ability to be integrated within other technologies or platforms, suggests that in some cases, AI capabilities might at least *appear* to map easily onto the domains or jurisdictions of pre-existing regimes.³⁸ In this light, it is perhaps less surprising that some scholars have judged that many challenges created by AI are not very new, and can be accommodated within existing regimes of international law.³⁹ So long as AI technology is merely used to render other platforms or technologies faster or more accurate, this could suggest that it often does not create sufficiently compelling new governance gaps for development to fill.⁴⁰

That does not mean that there will be no AI capabilities that appear sufficiently new to ground a perceived call for new laws, of course. Indeed, the international movement to ban Lethal Autonomous Weapons Systems offers perhaps the clearest example of a situation where many global actors, including a sizeable coalition of states, believe that these technologies face a clear legal gap, and will require new international regulation or even a ban. However, even in this case, the fact that international debates on this topic have, so far, remained predominantly housed under the Convention on Certain Conventional Weapons—rather than shifting to a separate venue, as happened with the Ottawa Mine Ban Treaty,⁴¹ also suggests that many parties may prefer to accommodate these technologies within existing international frameworks.⁴² As such, the question of whether a certain new AI capability creates a perceived need for ‘new laws’, ultimately depends not just on features of the technology, but also of the perceived reach and flexibility of instruments available in the extant governance landscape.

³⁷ As noted by Karen J. Alter & Kal Raustiala, *The Rise of International Regime Complexity*, 14 ANNU. REV. LAW SOC. SCI. 329–349, 337 (2018). (“[g]lobal governance solutions thus must take one of two approaches: (a) International actors can attempt to create an encompassing regime that can address all dimensions of the problem, or (b) international actors can accept that policy solutions will be crafted, coordinated, and implemented within a larger regime complex. [...] although the first option might be more efficient and effective, it is rarely the solution adopted.”). See also Chapters 6 and 7.3-7.4.

³⁸ See for example the discussion of autonomous maritime vehicles under the Law of the Sea, in Natalie Klein, *Maritime Autonomous Vehicles within the International Law Framework to Enhance Maritime Security*, 95 INT. LAW STUD. 29 (2019).

³⁹ See also Martina Kunz & Seán Ó hÉigeartaigh, *Artificial Intelligence and Robotization*, in OXFORD HANDBOOK ON THE INTERNATIONAL LAW OF GLOBAL SECURITY (Robin Geiss & Nils Melzer eds., 2020), <https://papers.ssrn.com/abstract=3310421> (last visited Jan 30, 2019).

⁴⁰ This may be the case even if effectively mitigating the new challenge would in fact be better served by an entirely new instrument.

⁴¹ Although there have been calls for a similar shift to take place in the regulation of LAWS. See Janosch Delcker, *How killer robots overran the UN*, POLITICO, 2019, <https://www.politico.eu/article/killer-robots-overran-united-nations-lethal-autonomous-weapons-systems/> (last visited Oct 2, 2019). See also the discussion in Section 2.3.2.3.1 (on developments in the regulation of LAWS).

⁴² Maas, *supra* note 14.

Finally, it is important to keep in mind that even if an AI capability is perceived to be capable of accommodation within existing regimes, this does not mean that that is also the most effective governance approach. A novel regime might in fact be more effective at addressing it, if it were possible. Conversely, even if an AI capability does give rise to a perceived governance gap, it is an entirely distinct question as to whether new regulation to ‘patch’ this gap will prove politically viable.⁴³

The key situation here is that the existing governance instruments have recognised a new technology (or its specific usage) as clearly falling out of scope of existing instruments, establishing a need for new international law to adequately deal with it.

5.3.2 Conceptual Uncertainty & Ambiguity

Secondly, a new technology may create or render explicit latent conceptual ambiguities or fault lines within existing legal instruments, specifically over cornerstone legal terms, or over the boundary between two previously-distinct bodies of law.⁴⁴ This can create legal uncertainty, because it becomes unclear over how, or even whether, existing laws apply to certain new activities.⁴⁵ As Bennett Moses has argued, technology can spark a need for legal clarification and change whenever there is uncertainty over how to legally classify the new entities, forms of conduct or relationships. Such situations can occur for various reasons: there may be no adequate classification rubric that exists; the new behaviour fits into more than one existing category and becomes subject to different and conflicting rules; or an existing legal category becomes blurred.⁴⁶

Such conceptual uncertainties have been previously produced by diverse technologies. For instance, Crootof has discussed how the use of drones for targeted killing has “highlighted questions regarding the appropriate geographic and temporal limitations of armed conflicts, as well as the relationship between the law of armed conflict and international human rights law”.⁴⁷ Likewise, AI, it has been suggested, combines certain features (such as autonomy and unpredictability) which blur cornerstone legal categories, such as the distinction between ‘agent’ and ‘object’,⁴⁸ and therefore undercut symbolic orders that are at the heart of settling questions of responsibility and liability.⁴⁹ Likewise, it has been argued that LAWS frustrate some of the

⁴³ Critically, if it is not, this situation, where a key problem remains unaddressed because of political gridlock or failure, may result in governance ‘erosion’ as discussed below, under section (5.3.6), and under (5.5.1).

⁴⁴ Maas, *supra* note 13 at 39.

⁴⁵ In a different distinction, Crootof & Ard distinguish between *application uncertainties* (over whether the new application is covered by existing law), *normative uncertainties* (over whether the new application should be covered by certain law), and *institutional uncertainties* (over which agencies or parties should decide these other questions). Rebecca Crootof & B. J. Ard, *Structuring Techlaw*, 34 HARV. J. LAW TECHNOL., 11, 14–36 (2021), <https://papers.ssrn.com/abstract=3664124> (last visited Aug 28, 2020).

⁴⁶ Bennett Moses, *supra* note 25 at 26.

⁴⁷ Crootof, *supra* note 27 at 7; Similar points are also made by Allenby, *supra* note 15; Denise Garcia, *Future arms, technologies, and international law: Preventive security governance*, 1 EUR. J. INT. SECUR. 94–111 (2016).

⁴⁸ JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE 64–79 (2018).

⁴⁹ Particularly in contexts of international humanitarian law. Cf. Hin-Yan Liu, *Categorization and legality of autonomous and remote weapons systems*, 94 INT. REV. RED CROSS 627–652 (2012).

assumptions about state accountability as well as criminal intent, which are critical to the application of international accountability frameworks.⁵⁰

This resulting legal uncertainty creates the need for development, in the form of the clarification or sharpening of existing rules, or their definitions and conditions. The degree to which this can be achieved in practice, within the confines of an existing treaty or regime, may depend on the ease of formulating a new and plausible interpretation of the technology, behaviour or concept in question.⁵¹ However, it is a distinct question whether there is sufficient conceptual flexibility that the previous interpretation can be recovered or restored. If it cannot, it is a political question whether there is sufficient willingness to explicitly amend existing treaties, or propose new regimes, to carry out the required development.⁵²

5.3.3 Incorrect Scope of Application of Existing Laws

In the third place, new technologies can create new behaviours or situations that blur the scope of application of existing regulation, leading to situations of inappropriate over- or under-inclusion. This essentially reflects the classic legal ‘no vehicles in the park’ dilemma—where a certain rule was once formulated to ban with certain objects (e.g. motor vehicles) from a park, but where it was phrased without awareness of other objects (e.g. bicycles, roller skates, electric wheelchairs; drones; ...) that might fall under this terminology, creating later uncertainty over whether it would—or why it should—apply to these new objects.⁵³

5.3.3.1 Accidental or engineered under-inclusivity

On the one hand, technological change can result in situations of *under-inclusivity*, where new technologies (and the resulting capabilities) slip through the net of the exact phrasing of older regimes, even though they might reasonably be held to be in violation of the law’s spirit. In situations where this happens by accident, this may not be such a problem, and can often be pragmatically resolved through adaptive interpretation or treaty amendment by all the states parties. Such seems to have been the reaction to the inadvertent under-inclusion of autonomous vehicles under the 1949 and 1968 Road Traffic Conventions—a situation that has in the last five years seen remarkably rapid action by the Global Forum for Road Traffic Safety to pragmatically resolve this.⁵⁴

⁵⁰ Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 UNIV. PA. LAW REV. 1347–1402 (2016). However, note that such puzzles may not always be intractable. For instance, Thomas Burri has argued that the case law of international courts in principle includes more than sufficient precedent for resolving questions of state control, attribution and the limits of delegation. Burri, *supra* note 1 at 101–103, 108.

⁵¹ In turn, as Crootof has noted, “the success of a new interpretation will depend on the authority of the interpreter, the specificity of the existing law, and the appropriateness of any analogy employed in the interpretation.” Crootof, *supra* note 27 at 14. We will shortly discuss the opportunities (and risks) of re-interpretation, in section 5.3.6.2.

⁵² These two barriers—of conceptual intractability, or political intractability—underlie the later discussion of governance ‘erosion’, under Section 5.5.1.

⁵³ The original formulation derives from H.L.A. Hart. H. L. A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. LAW REV. 593, 607 (1958). See also Pierre Schlag, *No Vehicles in the Park*, 23 SEATTLE UNIV. LAW REV. 381–389 (1999). The link is also drawn by Crootof, *supra* note 27 at 9.

⁵⁴ Bryant Walker Smith, *New Technologies and Old Treaties*, 114 AJIL UNBOUND 152–157 (2020). See also the discussion, in section 5.3.6, on ‘carrying out development’.

However, at the international level, such under-inclusion can also be the result of active attempts by parties “to respect the letter of existing [...] agreements, while violating their spirit.”⁵⁵ There are various historical examples of such engineered under-inclusivity. For instance, under the 1967 Outer Space Treaty, states pledged “not to place in orbit around the earth any objects carrying nuclear weapons or any other kinds of weapons of mass destruction.”⁵⁶ This provision was meant to outlaw nuclear orbital bombardment systems, and more generally ensure that space would remain the preserve of peaceful exploration. However, shortly after signing the treaty, the Soviet Union tested a Fractional Orbital Bombardment System (FOBS) which launched nuclear weapons into space, but de-orbited them before they had completed a single full ‘orbit’, thus allowing them to stay stringently within the letter of the treaty.⁵⁷ For a more recent example in the same field, one can take debates over a series of upgrades which the US has conducted on the W76 nuclear warheads deployed on their missile submarine fleet. By outfitting these missiles with ‘superfuze’ that can trigger detonation based on proximity to a target (rather than pre-set burst-height), this enables a far more reliable targeting of hardened installations such as missile siloes, without the need to reengineer the ballistic missiles themselves for greater accuracy. Consequently, this modestly-named ‘modernization measure’ has been estimated to have tripled the effective counter-force lethality of the US nuclear missile submarine force.⁵⁸ This shows how even seemingly incremental technological shifts or improvements in capability can produce qualitatively significant sociotechnical shifts (such as in structural incentives around nuclear deterrence), even as states formally comply with caps under technology-specific arms control treaties.

In the case of AI, various innovations might produce specific capabilities that would be considered (by many) in violation of at least the spirit of various arms control treaties, resulting in illegitimate ‘under-inclusivity’ of such regimes.⁵⁹ Elsewhere, Amanda McAllister has argued that advances in the use of autonomous robots for interrogation or torture would likely slip through the language set down in the UN Convention Against Torture and Other Cruel, Inhuman

⁵⁵ Maas, *supra* note 14 at 137.

⁵⁶ TREATY ON PRINCIPLES GOVERNING THE ACTIVITIES OF STATES IN THE EXPLORATION AND USE OF OUTER SPACE, INCLUDING THE MOON AND OTHER CELESTIAL BODIES, 610 UNTS 205 (1967). Article IV.

⁵⁷ Raymond L. Garthoff, *Banning the Bomb in Outer Space*, 5 INT. SECUR. 25–40 (1980). As discussed in DANIEL DEUDNEY, DARK SKIES: SPACE EXPANSIONISM, PLANETARY GEOPOLITICS, AND THE ENDS OF HUMANITY 413 (2020). The incident caused a political uproar in the US, and no further tests of the system were conducted, although the launchers stayed operational. Subsequently, FOBS-type systems were explicitly prohibited by the SALT II agreement of 1979; while the US Senate did not ratify SALT II, the Soviet Union did comply with its terms, decommissioning or converting the remaining FOBS launchers by 1983. MIROSLAV GYÚRÖSI, *The Soviet Fractional Orbital Bombardment System Program* 1–1 (2010), <http://www.ausairpower.net/APA-Sov-FOBS-Program.html> (last visited Sep 21, 2020).

⁵⁸ Hans M. Kristensen, Matthew McKinzie & Theodore A. Postol, *How US nuclear force modernization is undermining strategic stability: The burst-height compensating super-fuze*, BULLETIN OF THE ATOMIC SCIENTISTS (2017), <https://thebulletin.org/how-us-nuclear-force-modernization-undermining-strategic-stability-burst-height-compensating-super10578> (last visited Mar 12, 2018). As discussed in Maas, *supra* note 14 at 137.

⁵⁹ Maas, *supra* note 14 at 137–138. (suggesting examples of “novel human-machine integration dynamics (e.g. crowd-sourced labelling of training data; new ways to integrate human operators with drone swarms, or soldier-platform brain-computer interfaces), which could formally re-insert a human ‘in-the-loop’, but in altered or distributed modes of cognitions that render this arrangement less protective or meaningful”).

or Degrading Treatment or Punishment, even as it would clearly be against the spirit of that regime.⁶⁰

5.3.3.2 Inappropriate over-inclusivity

However, there are also cases of potential *inappropriate over-inclusivity* of existing laws to new AI technology. As one interesting example of this, some scholars have in recent years argued that it might be legally possible to grant certain algorithms a semblance of functional legal personhood. For instance, Shawn Bayern has claimed that certain loopholes in existing US company law might allow for the incorporation of a limited liability company (LLC), with an operating agreement that places it under operational control of an AI; if all human members subsequently withdrew, the LLC would theoretically be left with an algorithm solely in charge, functionally establishing ‘algorithmic entities’ with legal personhood.⁶¹ This argument has of course been contested, with many others observing that courts would likely not interpret the relevant statutes in a manner that obviously appears contrary to legislative intent.⁶²

Nonetheless, stranger rulings have on occasion passed, and assuming such a construction were to be held up in court, what could be the result? Bayern and others have sought to extend this same schema to the legal systems of several EU member states.⁶³ This creates a potential additional overinclusivity loophole: as argued by Thomas Burri, if an ‘algorithmic entity’ were successfully established in any EU member state, the internal market principle of the mutual recognition of national legal personality would suggest that that system should then have to be legally recognised by all EU member states.⁶⁴ Such outcomes however, would—and arguably should—be considered situations of inappropriate over-inclusiveness of existing regulation.⁶⁵

More generally, inadvertently over-inclusive regulation has been linked to a range of problematic first- and second-order effects, from unjustifiable social costs of complying with an overinclusive law, unintended overenforcement, or the risk that the underenforcement of overinclusive treaties ends up contributing to visible noncompliance, eroding the perceived strength or legitimacy of a given legal regime by creating the impression it does not carry the force of the rule of law.⁶⁶

When technologies create a perceived problem in the scope of existing legislation—whether under or overinclusive—this creates a need for development to extrapolate and reaffirm

⁶⁰ Amanda McAllister, *Stranger than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture*, MINN. LAW REV. 47 (2018).

⁶¹ Shawn Bayern, *The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems*, 7 EUR. J. RISK REGUL. 297–309, 300–304 (2016). As discussed in Maas, *supra* note 13 at 40.

⁶² Matthew U. Scherer, *Is AI personhood already possible under U.S. LLC laws? (Part One: New York)*, LAW AND AI (2017), <http://www.lawandai.com/2017/05/14/is-ai-personhood-already-possible-under-current-u-s-laws-dont-count-on-it-part-one/> (last visited May 4, 2019); TURNER, *supra* note 48 at 177.

⁶³ Specifically Germany, Switzerland, and the UK (at the time still a EU member). Shawn Bayern et al., *Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators*, 9 HASTINGS SCI. TECHNOL. LAW J. 135–162, 139–53 (2017).

⁶⁴ Thomas Burri, *Free Movement of Algorithms: Artificially Intelligent Persons Conquer the European Union’s Internal Market*, in RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE, 545 (Woodrow Barfield & Ugo Pagallo eds., 2017), <https://papers.ssrn.com/abstract=3010233> (last visited Sep 13, 2018).

⁶⁵ Burri, *supra* note 1 at 95–98. See also Maas, *supra* note 13 at 40.

⁶⁶ See also Crootof and Ard, *supra* note 45 at 31–33.

existing lines between legal categories, so as to explicitly include or exclude the new behaviour, entities or relationships.

5.3.4 Rendering obsolete core assumptions of governance

Fourthly, AI can allow new behaviour that *renders obsolete core assumptions* which underpin a particular law or governance regime. This can occur for one of three subsidiary reasons: because the technology (a) renders once regulated behaviour obsolete (undercutting the continued *need for* the regime); (b) undercuts certain justifying assumptions (challenging the *adequacy* of the manner in which the regime pursues its intended goals); or (c) makes existing laws no longer cost-effective to enforce (challenging the long-term viability or deterrent effects of the regime). In all three cases, AI capabilities can drive functional ‘legal obsolescence,’ where existing law is rendered unfit for its originally intended purposes, consequently requiring a reconsideration of the core aims and purposes of that regime.

5.3.4.1 *Obsolescence or decline of regulated behaviour*

In the first case, obsolescence of a governance regime could result because a formerly common behaviour that was subject to regulation has been superseded or rendered obsolete in practice, or drastically reduced in importance, as a result of new technological capabilities.⁶⁷

One might expect this category of legal obsolescence to be relatively rare. This is because while technological progress can often lower the threshold for certain behaviours, and can occasionally create new affordances or capabilities, it less often results in the categorical disappearance (let alone the impossibility) of older types of behaviour.⁶⁸ Even in cases where it does, there is of course often an implication that certain categories of behaviour could still be revived, which might ensure the continuing relevance of the law. Where it does occur, such obsolescence might be expected to play a role primarily in the context of highly technology-specific laws, such as rules on the management of telegraph infrastructure.⁶⁹

In the context of AI, however, one might consider how a continued extrapolation of recent military trends towards ‘remote warfare’ could result in such changes. If AI-steered combat platforms come to widely replace human soldiers on battlefields, this might indirectly marginalise or render practically uninvoked IHL principles dictating the treatment of prisoners of war.⁷⁰ Indeed, just as submarine warfare involved vessels that were architecturally ill-equipped for shipping back the crews of target vessels to port, which therefore eventually drove new state practice in naval warfare which rendered the older London Naval Treaty ‘jurisprudential space

⁶⁷ Bennett Moses, *supra* note 25 at 48–49. See also Bennett Moses, *supra* note 25 at 595.

⁶⁸ Of course, it is also an open question whether such ‘dead letter’ obsolescence truly requires categorical cessation or obsolescence of certain behaviour, or whether virtual disappearance suffices. Likely there is a spectrum, running from ‘previously regulated behaviour which has not just vanished but rendered practically impossible’, to ‘previously regulated behaviour which still occurs, but at a very low level where it is no longer a nuisance’.

⁶⁹ On the other hand, at least within the common law system, cases dealing with telegraphs or other outmoded forms of communication may nonetheless have continued relevance. In such cases, these laws might be only partially obsolete.

⁷⁰ As set out under the THIRD GENEVA CONVENTION RELATIVE TO THE TREATMENT OF PRISONERS OF WAR, (1949). See also Maas, *supra* note 13 at 42.

junk’,⁷¹ the continuation towards AI-supported ‘remote warfare’ could in extreme cases precipitate a similar shift in customary international law. An alternate example might involve ongoing developments in ‘remote biometric surveillance’, such as gait detection, or laser-measured heartbeat identification, some of which have proven as accurate as facial recognition at identifying individuals.⁷² If such applications were to replace facial recognition in most surveillance systems, this could render any regulation or bans on facial recognition systems relatively moot.⁷³

Of course, one can debate to what extent this type of legal obsolescence is very likely or common—or, if it occurs, whether it truly poses a problem for governance regimes. While in some cases under domestic law, antiquated and virtually uninvoked ‘dead letter’ laws do rear their head,⁷⁴ in most cases the legal disruption imposed on a legal system by having ‘superfluous but un-invoked’ laws might seem problematic, if at all, only from a legal coherentist perspective.⁷⁵ On the other hand, as noted by Bennett Moses, “it can be a problem if the regulated conduct has been replaced by conduct that causes harm of a type the rule sought to avoid, but does not fall within the rule itself.”⁷⁶ Furthermore, this type of legal obsolescence could still pose a problem for international law (even from a functional perspective), if obsolete treaties provide (misleading) judicial metaphors; if they delay awareness of shifts in the behaviour to be regulated; or if they grow into ‘jurisprudential space junk’ that muddles the waters regarding the status of subsequent customary international law.⁷⁷

5.3.4.2 Foundational justifying assumptions are no longer valid

More problematically for international law, legal obsolescence might occur because one or more basic justifying assumptions that underlie the original historical introduction or specific formulation of a particular law or regime are no longer valid in the face of sociotechnical change.⁷⁸

For instance, Rebecca Crootof has discussed how the emergence of increasingly precise weapons technologies is gradually stressing core tenets of the law of armed conflict. Under customary international humanitarian law, it has long been assumed that all state parties that participate in a conflict are subject to the same rights and restrictions, and that these obligations do not depend on their (relative) technological capabilities. However, because certain obligations—such as the nominally tech-neutral obligation to take ‘feasible precautions’ in an attack—are in fact affected by the technological capabilities (e.g. ISR or precision targeting)

⁷¹ Crootof, *supra* note 9.

⁷² Keith J Hayward & Matthijs M Maas, *Artificial intelligence and crime: A primer for criminologists*, CRIME MEDIA CULT. 1741659020917434 (2020).

⁷³ Though under another interpretation, this would involve a situation of inappropriate under-inclusion of any regulation meant to regulate systems of surveillance.

⁷⁴ For instance, see (in the context of laws on blasphemous libel), Jeremy Patrick, *Not Dead, Just Sleeping: Canada’s Prohibition on Blasphemous Libel as a Case Study in Obsolete Legislation*, 41 UBC LAW REV. 193–248 (2008).

⁷⁵ See Graham McBain, *Abolishing Some Obsolete Common Law Crimes*, 20 KINGS LAW J. 89–114 (2009).

⁷⁶ Bennett Moses, *supra* note 25 at 48.

⁷⁷ Crootof, *supra* note 9. On the other hand, it could be countered that treaties that apply solely to extinct technologies or behaviour cannot clash with new customary international law, and therefore cannot produce ‘jurisprudential space junk’.

⁷⁸ Bennett Moses, *supra* note 25 at 49–53. Maas, *supra* note 13 at 41.

available to a party, continued technological progress may functionally erode the equality and symmetry of state obligations under IHL in at least some respects.⁷⁹

In the context of AI, what would be examples of the obsolescence of foundational assumptions? One could consider the (admittedly aspirational) human right to work, as enshrined in the Universal Declaration of Human Rights and the International Covenant on Economic, Social and Cultural Rights,⁸⁰ which is premised on an assumption that societies will continue to need the labour of—or be capable of providing employment for—a large fraction of their population. If at length AI systems continue to reach or approach human performance in more tasks,⁸¹ and a significant population were to be rendered functionally unemployable, this change may render this (admittedly aspirational) right to work functionally obsolete,⁸² with additional repercussions to parts of the global legal regime constructed by the International Labour Organization.⁸³ In the longer term, this could result in situations where many people's status as ('productive') participants in society would increasingly be drawn into question.⁸⁴ This would result in the need for legal development towards regimes that would explore different avenues of securing dignified human activity and inclusion in a post-work era.⁸⁵

5.3.4.3 Reduced cost-effectiveness of enforcement

Finally, legal obsolescence (and the accordant need for corrective legal development) might occur because a law is no longer enforceable, at least not in a cost-effective manner.⁸⁶ Practical difficulties introduced by new technologies can often slow the enforcement of older legal frameworks to new spaces. For instance, the difficulty of attributing attacks in cyberspace has been held as one (if not the only) hurdle to effectively regulating cyberattacks.⁸⁷

While some uses of AI technology could certainly support the monitoring or enforcement of international law,⁸⁸ some developments could also impede enforcement of international law. For instance, much scholarship has discussed the potential of AI-produced 'DeepFakes'

⁷⁹ Crootof, *supra* note 27 at 11–12.

⁸⁰ UN GENERAL ASSEMBLY, *Universal Declaration of Human Rights*, 217A(III) (1948). Article 23. 999 UN GENERAL ASSEMBLY, INTERNATIONAL COVENANT ON CIVIL AND POLITICAL RIGHTS (1966). Article 6.

⁸¹ Katja Grace et al., *When Will AI Exceed Human Performance? Evidence from AI Experts*, 62 J. ARTIF. INTELL. RES. 729–754 (2018). See also the discussion in chapter 2.1.6 (on change from future development).

⁸² See also Mathias Risse, *Human Rights and Artificial Intelligence: An Urgently Needed Agenda*, 41 HUM. RIGHTS Q. 1–16 (2019). However, it is simultaneously possible that such a development could put greater emphasis on workers' rights in regions producing the raw materials or inputs for AI and its infrastructures. I thank Luisa Scarella for this observation.

⁸³ Maas, *supra* note 13 at 43.

⁸⁴ Risse, *supra* note 82.

⁸⁵ See for example JOHN DANAHER, AUTOMATION AND UTOPIA: HUMAN FLOURISHING IN A WORLD WITHOUT WORK (2019).

⁸⁶ Bennett Moses, *supra* note 25 at 53–54; Maas, *supra* note 13 at 42.

⁸⁷ Michael J Glennon, *The Dark Future of International Cybersecurity Regulation*, 6 J. NATL. SECUR. LAW POLICY 563–570 (2013). However, note the counterargument on attribution in Horowitz, *supra* note 10 at 395. (noting that attribution in cyberspace may be more viable than sometimes perceived, not only because of advances in attribution and digital forensics, but also because "the standard of proof for attribution in international politics may differ from that required for international domestic or legal purposes").

⁸⁸ See also the discussion, shortly, under Section 5.4.2. ('Automation of monitoring and enforcement').

techniques to enable the at-scale forging of video documentation.⁸⁹ This could include the faking of documentary evidence of alleged human rights abuses. Various scholars have as such raised the possibility that DeepFakes could adversely affect the probative value of (video) evidence, not only in domestic courts, but also potentially eroding certain epistemological foundation of international journalism, human rights investigations or international criminal judicial proceedings.⁹⁰ Accordingly, such uses could enable malicious (or politically interested) actors to swamp human rights bodies or media observers with fake footage, which could only be revealed as such after laborious analysis. Such strategies would impose heavy costs on these organisations, while generally eroding the credibility of genuine reports—the so-called the ‘liars’ dividend’.⁹¹ This would render the effective monitoring and enforcement of human rights norms more precarious.⁹²

5.3.5 Altering problem portfolio

In addition to these challenges, AI may drive an additional category of ‘governance disruption’, which is not neatly captured under these previous categories.⁹³ Specifically, there may be cases where AI innovation markedly shifts the prevailing sociotechnical ‘distribution of risks’ arising from the technology and its use.⁹⁴ I previously argued—in Paper [II]—that technological innovation can also drive ‘indirect disruption’ of existing governance regimes, by altering the ‘problem portfolio’ of a technology, creating blind spots that the existing regime is not adequately equipped (either in jurisdiction, institutional mandate, in-house expertise, or culture) to address.⁹⁵ This can serve as a fifth situation of legal uncertainty, which creates a need for developments to re-orient or re-balance existing regulatory regimes to track the relative balance of challenges created by a technology as it develops.

After all, as illustrated by our discussion of sociotechnical change, shifts in technological capabilities can unlock new use cases, which may generate new sociotechnical change in new domains. In some cases, this can shift the centre of gravity of the problem portfolio (for instance, from ‘ethics’ to ‘structure’). This can create a peculiar problem for existing governance regimes which accordingly find themselves focused on—and tailored to—the mitigation of only a subset of challenges.

Indeed, the ability of technologies to iteratively shift the ‘salient’ legal challenges confronting a regime, can be found in the changing challenges facing cyberlaw. As Jack Balkin

⁸⁹ Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. LAW REV. 1753–1820 (2019).

⁹⁰ Marie-Helen Maras & Alex Alexandrou, *Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos*, INT. J. EVID. PROOF 1365712718807226, 258 (2018); Livingston and Risso, *supra* note 6 at 144; Hin-Yan Liu & Andrew Mazibrada, *Artificial Intelligence Affordances: Deep-fakes as Exemplars of AI Challenges to Criminal Justice Systems*, UNICRI SPEC. COLLECT. ARTIF. INTELL. (2020), <http://unicri.it/towards-responsible-artificial-intelligence-innovation> (last visited Jul 13, 2020).

⁹¹ Chesney and Citron, *supra* note 89 at 1785.

⁹² Maas, *supra* note 13 at 42.

⁹³ Although on some readings, this could be read as a version of ‘obsolescence of foundational assumptions’ of a regime.

⁹⁴ Maas, *supra* note 14 at 139–141.

⁹⁵ *Id.* at 141.

noted, rather than confronting a single technology with immutable ‘essential features’, the internet proved a changeable regulatory target; whereas in 1991, the internet mostly appeared to raise challenges relating to anonymous communication across jurisdictional lines, by 1999, regulatory concerns focused on filtering and surveillance; and by 2008, the focus was on the internet’s ability to lower the costs of social organisation.⁹⁶ A comprehensive legal system for cyberspace, instituted in 1991, would therefore have faced gradual misalignment—not because of any single explicit situation of legal uncertainty, but simply because it would remain focused on-and tailored to a subset of the technology’s full portfolio of sociotechnical change. The key point is not that cyberlaw could not or did not change and respond to such developments, but rather that—especially in the international legal context, where multilateral treaties are hard to amend—there is a risk of early regimes becoming ‘locked in’.

For instance, in the domain of the regulation of military AI, most current regulatory approaches aim to ensure some forms of ‘Meaningful Human Control’ (MHC) over the exertion of lethal force; and/or to outright ban LAWS as being inherently ‘indiscriminate’ and therefore in violation of IHL.⁹⁷ Either solution might constitute a significant international legal response. Yet in the longer term, it may be that regimes tailored to address these ethical or legal challenges of LAWS are not optimally oriented or equipped to anticipate, let alone address, later risks arising from adjacent AI capabilities, such as destabilizing *structural shifts* produced by AI cyberwarfare systems or military decision support systems.⁹⁸

Of course, one counterargument would be that legal frameworks such as IHL have a clear focus and jurisdiction for a reason, and that they are not, and should not, concern itself much or at all with questions of strategic stability or a lowering threshold to the use of force,⁹⁹ but that these might instead be in the purview of arms control regimes or *jus ad bellum* instead. That may be a key point, though it also highlights the degree to which the ongoing evolution in the AI problem portfolio can erode the efficacy or sufficiency of existing regime distinctions which target only a subset of sociotechnical changes.

5.3.6 Carrying Through Development: Reflections and Risks

These above situations involve situations where AI applications can or could produce various substantive or doctrinal shortfalls in existing governance instruments. Whenever they do, these shortfalls accordingly need to be repaired—either to restore the status *ex quo ante*, or to achieve new regulatory goals in light of the new situation.

Generally speaking, these are situations of ‘governance disruption’ which demonstrate how AI systems might create a need for legal or regulatory change (‘development’), because they create new entities, enable new behaviour, shift incentives of actors to create new situations, or produce markedly new patterns and portfolios of sociotechnical change. Bennett Moses and others

⁹⁶ Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. LAW REV. CIRCUIT 17, 47–48 (2015).

⁹⁷ This is admittedly a rash generalization of a much wider array of possible regulatory approaches that have been coined. Cf. Esther Chavannes, Klaudia Klonowska & Tim Sweijns, *Governing autonomous weapon systems: Expanding the solution space, from scoping to applying*, HCSS SECUR. 39 (2020). However, these two avenues appear to have gotten the most traction at the international level.

⁹⁸ Maas, *supra* note 14 at 139–141. See also Koppelman, *supra* note 16.

⁹⁹ Deeks, Lubell, and Murray, *supra* note 5.

argue that such cases demonstrate how technology can create a ‘recurring dilemma’ for legal systems,¹⁰⁰ and argue that this is a problem against which governance cannot be systematically future-proofed.¹⁰¹ There are a few caveats to the concept of technology-driven *substantive governance disruption*, however.

5.3.6.1 *Caveats to governance disruption*

In the first place, *not all governance disruption is created equal*. Indeed, technological change, even if it produces patterns of sociotechnical change, is often not legally disruptive. As argued by Bennett-Moses;

“[m]ost of the time, a law predating technological change will apply in the new circumstances without any confusion. For example, traffic rules continue to apply to cars with electric windows, and no sane person would seek to challenge these laws as inapplicable or write an article calling for the law to be clarified.”¹⁰²

The same might be the case at the international level, where many new technological innovations come quietly, without producing significant disruption. Moreover, even where there is a degree of disruption, this can come in distinct degrees. There might certainly be times where a new technology is ‘exceptional’ in its disruption, because, as Calo notes, its introduction to the mainstream requires “a systematic change to the law or legal institutes in order to reproduce, or if necessary displace, an existing balance of values”.¹⁰³ Yet there are many other cases where a technology merely “creates one or more small changes in the law, or [...] reveals, at the margins, that an existing interpretation of a particular doctrine is incomplete”,¹⁰⁴ but where a failure of legal or regulatory development to address these minor details would not materially endanger either the coherence or instrumental utility of a legal system.¹⁰⁵ Indeed, some have argued that, at least in certain domestic jurisdictions, the ‘legally disruptive’ impact of AI technologies is relatively rare, with most applications fitting within the remit of most existing laws most of the time.¹⁰⁶ Accordingly, beyond the binary question of whether or not an AI capability creates one of the five types of substantive legal disruption discussed above, we should also examine the

¹⁰⁰ Bennett Moses, *supra* note 25.

¹⁰¹ Friedman, *supra* note 24 at 85.

¹⁰² Bennett Moses, *supra* note 25 at 596.

¹⁰³ Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. LAW REV. 513–564, 552 (2015).

¹⁰⁴ *Id.* at 552.

¹⁰⁵ On the distinction between the regulatory mind-sets of ‘legal coherentism’ and ‘regulatory instrumentalism’ (and how technology may shift the balance amongst them, as well as towards a more ‘technocratic’ regulatory mindset), see also Roger Brownsword, *Law and Technology: Two Modes of Disruption, Three Legal Mind-Sets, and the Big Picture of Regulatory Responsibilities*, 14 INDIAN J. LAW TECHNOL. 1–40 (2018); Roger Brownsword, *Law Disrupted, Law Re-Imagined, Law Re-Invented*, TECHNOL. REGUL. 10–30 (2019).

¹⁰⁶ For instance, Carlos Ignacio Gutierrez Gaviria has recently applied Bennett Moses’s framework to the US regulatory context, to examine which AI systems create genuinely new situations. Reviewing 50 regulatory gaps, he argues that questions of conceptual uncertainty (42%) and targeting or scope (26%) were the most prevalent gaps, but argues that only a limited subset of uses (12%) posed genuinely new legal challenges (i.e. new gaps). Carlos Ignacio Gutierrez Gaviria, *The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the U.S.*, 2020, https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html (last visited Jul 11, 2020).

question of how much stress is put on the existing regulatory regime to adapt—and how adaptive or flexible it generally is.

Secondly, related to this, *substantive governance disruption occurs relative to a particular legal system*. It is not a general feature of a given technology or application (or even certain type of sociotechnical change) to be ‘disruptive’; rather, this effect can only be understood in relation or reference to the conceptual categories or assumptions of one or another specific legal system. As a result, when considered at a national level, a certain AI capability might be disruptive in one national jurisdiction but not in others.¹⁰⁷ Likewise, at the international level, technologies can drive different combinations or sets of governance disruption in different regimes. For instance, Crootof notes how “a weapon that creates one kind of legal disruption in one area of international humanitarian law might cause a very different kind of disruption in others—or in international human rights law, the law of the sea, space law, or another legal regime.”¹⁰⁸

For example, various legal scholars have argued that, with a few exceptions, autonomous maritime vessels create relatively few intractable conceptual problems under the existing law of the sea.¹⁰⁹ Even so, the use of unmanned underwater vehicles by criminal networks to smuggle contraband in a more deniable manner could instead pose an *obsolescence* challenge for operational (cost-effectiveness) assumptions of law enforcement,¹¹⁰ potentially challenging instruments such as the 1988 United Nations Convention Against Illicit Traffic in Narcotic Drugs and Psychotropic Substances,¹¹¹ or the broader mission of the United Nations Office on Drugs and Crime (UNODC). Such cases show how we should not expect AI technology (or any given capability or use case) to be ‘universally disruptive’ to governance. Rather, as a general-purpose technology, it might instead produce varied constellations of governance disruption in different fields, which should be cautiously mapped, as they may require distinct avenues or paths of development.

Thirdly, this also illustrates the potentially *differential ‘resilience’ or absorption capacity of different types of law or governance system* to potential disruption. For instance, much of the above discussion has focused on potential disruptions of hard law instruments (e.g. treaties). However, that should not be very surprising, as they might be the sites where we would expect to find most such disruption. In contrast, broader forms of soft law might be more conceptually ambiguous or flexible, and would be better able to either absorb the technological shocks, or to facilitate quick development. This suggests that hard law and soft law can both be disrupted by

¹⁰⁷ Meg Leta Jones, *Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw*, 2 J. LAW TECHNOL. POLICY (2018), <http://illinoisjlp.com/journal/wp-content/uploads/2018/12/Jones.pdf> (last visited Sep 7, 2020).

¹⁰⁸ Crootof, *supra* note 27 at 13.

¹⁰⁹ See also Klein, *supra* note 38; Robert McLaughlin, *Unmanned Naval Vehicles at Sea: USVs, UUVs, and the Adequacy of the Law*, J. LAW INF. SCI. (2011), <http://www5.austlii.edu.au/au/journals/JLawInfoSci/2012/6.html> (last visited Jun 22, 2020).

¹¹⁰ See Thomas C. King et al., *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, SCI. ENG. ETHICS, 12 (2019), <https://doi.org/10.1007/s11948-018-00081-0> (last visited Feb 21, 2019); Noel Sharkey, Marc Goodman & Nick Ros, *The Coming Robot Crime Wave*, 43 IEEE COMPUT. MAG. 116–115 (2010).

¹¹¹ UNITED NATIONS CONVENTION AGAINST ILLICIT TRAFFIC IN NARCOTIC DRUGS AND PSYCHOTROPIC SUBSTANCES, (1988).

technological change, but that this may be more evident in the case of the former.¹¹² Nonetheless, while this implies that some parts of global governance might be relatively shielded from technology-driven governance disruption, the susceptibility of international hard law instruments to such disruption is nonetheless critical. This is because broader governance tools such as soft law can often be either rooted in- or bracketed by hard law foundations, or are created when it appears that hard law instruments are inviable or ineffective.¹¹³ In the former case, governance disruption of hard law instruments could indirectly cascade through to ‘softer’ instruments which implicitly assume their coherence. In the latter case, governance disruption of hard law instruments might prompt a need for new development that might only be adequately (or quickly) met through soft law tools.¹¹⁴

Fourthly, it is not governance disruption that creates problems, but a governance system’s *inability to adapt through some form of development*. As such, situations of legal uncertainty are not intrinsically problematic in and of themselves. They only become challenging to a given system of law when it, for conceptual or political reasons, is unable to carry out the needed changes or updates in response (or even unable to adequately identify the need for such changes).¹¹⁵ In principle, after all, legal systems are perfectly capable of any revisions or change—bound only, perhaps, by the limits of natural language.¹¹⁶ Many abstract legal notions are flexible enough to be reinterpreted or adapted to many new situations;¹¹⁷ where even they fall short, governance systems can introduce new concepts, rights or legal categories tailored to clarify the situation. In that sense, legal development is almost always possible in principle. For instance, in an international law context, new governance gaps, ambiguities, scope inaccuracies, obsolescence, or problem portfolio shifts can all in principle be addressed through routine ‘legal development’ in a variety of ways: through norm development in customary international law, new treaties, amendments to existing treaties, reinterpretation of existing agreements, or entirely new governance arrangements.¹¹⁸

However, in practice, not all these responses are equally sufficient or equally viable. New treaties or treaty amendments might appear the most valuable solution, as they could in principle be written and tailored directly to the problem at hand. To be sure, there are certainly cases where there is sufficient (state) pragmatism and desire to reach a consensus that such solutions can be

¹¹² I thank Roger Brownsword for this observation.

¹¹³ I thank Laura Emily Christiansen for this observation.

¹¹⁴ Crootof, *supra* note 9 at 124–126.

¹¹⁵ Such situations may result in ‘erosion’ or even ‘destruction’ of governance regimes, as will be discussed shortly, in section 5.5.

¹¹⁶ As well as, in many domestic contexts, by the limits of ‘folk understandings and cultural norms’. Pierre Schlag, *Spam Jurisprudence, Air Law, and the Rank Anxiety of Nothing Happening (A Report on the State of the Art)*, GEORGETOWN LAW J. 803, 812, 821 (2009).

¹¹⁷ Indeed, articulating legal rules in a sufficiently broad and abstract way could in principle resolve any uncertainty, ensuring there would never be a need for development. As noted by George Friedman, “If legal rules are defined in sufficient breadth, legal innovation is never necessary. [...] Indeed, it is arguably possible to resolve all legal issues by a single very broad rule: have whatever legal rules maximize economic efficiency”. Friedman, *supra* note 24 at 85. There would of course be drawbacks to such principles, which would be far too broad to “apply with predictable results at a reasonable cost”. *Id.* at 85. On the general flexibility and limits of legal systems, see also Orin S. Kerr, *A Theory of Law*, 16 GREEN BAG 111 (2012).

¹¹⁸ Picker, *supra* note 12 at 156.

carried through, or treaties can be adapted. For instance, in a recent case study of international law's efforts to come to terms with autonomous vehicles, Bryant Walker Smith notes how states have available various approaches in order to reconcile driverless vehicles with the requirements for a (human) 'driver'—stipulated in both The 1949 Geneva Convention on Road Traffic, as well as the 1968 Vienna Convention on Road Traffic.¹¹⁹ These approaches to carry out 'legal development' in order to reconcile self-driving cars with the existing road traffic conventions show a range of possible actions for states, from soft law (unilateral explanatory memoranda or multilateral Working Party Resolutions under the Global Forum for Road Traffic Safety), to hard law (whether unilateral renunciation or reservations of the treaties, multilateral efforts to amend either convention, or introduce a new convention).¹²⁰ Nonetheless, in general, and particularly in the context of high-stakes technologies, amending a multilateral resolution can often run into political and conceptual challenges.¹²¹

As such, of all of these various approaches to dealing with legal uncertainty at the international level, Crootof argues that treaty 'reinterpretation' is the most common strategy used to fold new (weapons) technologies into an existing governance regime, in part because it is cheaper, quicker, and more reliable than attempting to negotiate or amend a treaty, establish Customary International Law, or seek an Advisory Opinion from the ICJ.¹²² However, while such attempted (re)interpretations may prove a commonly used strategy of first resort in order to achieve the necessary legal development, this is not without its risks.

5.3.6.2 AI at the limits of words: analogical reinterpretation as risky development shortcut

Situations where some needed development in regimes is not carried out create the risk of 'erosion' and the inadequacy of existing regimes. Of course, in the first place, whether a treaty (re)interpretation is even possible depends to a large extend on the authority and legitimacy of the interpreter,¹²³ as well as on conceptual and material features of the technology in question.¹²⁴ However, even if adaptation through reinterpretation appears viable, there may be hidden risks or shortfalls in conducting development in this manner, if it relies on or imports misleading (or unproductive) technological analogies. This matters to AI because AI technology is particularly prone to lending itself to diverse analogies, which each import distinct regulatory programs or priorities.

Of course, appropriate and vivid metaphors or framings of a technology may be valuable from an *agenda-setting* perspective. They can ensure that (global) policymakers or the public engage with a complex emerging technology in the first place—that is, that the international legal

¹¹⁹ Smith, *supra* note 54. Specifically, the 1949 Convention requires that "[e]very vehicle or combination of vehicles proceeding as a unit shall have a driver" and that "[d]rivers shall at all times be able to control their vehicles or guide their animals." 1949 CONVENTION ON ROAD TRAFFIC, 125 UNTS 22 (1949). Art. 8. The 1968 Convention similarly requires that "[e]very moving vehicle or combination of vehicles shall have a driver" and that "[e]very driver shall at all times be able to control his vehicle or to guide his animals." 1968 CONVENTION ON ROAD TRAFFIC, 1042 UNTS 15705 (1968). Art. 8.

¹²⁰ Smith, *supra* note 54 at 154–155.

¹²¹ See generally Crootof, *supra* note 9. See also the discussion in section 5.5.1 (on Erosion).

¹²² Crootof, *supra* note 27 at 13–14.

¹²³ *Id.* at 14.

¹²⁴ See also the discussion of a given interpretation being able to pass the 'laugh test', in section 5.5.1.1.

system recognizes that existing law falls short, and development is needed.¹²⁵ More generally, analogical reasoning is core to many domestic legal and judicial systems, especially in a common-law context. Indeed, especially where it comes to new technologies that may present ‘unprecedented’ (in the most literal and mundane sense of the word) situations, analogical reasoning may well constitute a ‘critical legal tool’, which, as Crootof notes, helps “make new kinds of technology accessible, allow for the application of existing law, and can help identify particular risks or solutions.”¹²⁶

However, there are risks to the unreflexive legal reliance on analogies in either the domestic or the international context. Crootof identifies at least three problems with analogies in regulation. In the first place, the selection and foregrounding of a certain metaphor occludes both that there are always multiple analogies possible for any new technology, and how each of these advances different ‘regulatory narratives’.¹²⁷ Secondly, analogies can be misleading by failing to capture a key trait of the technology, or by alleging certain characteristics that do not actually exist.¹²⁸ Most importantly, “analogies limit our ability to understand the possibilities and limitations of new technology.”¹²⁹ They constrain our imagination and understanding, and close down our conceptions of the other forms a technology might take, the other uses it might lend itself to, and the indirect or longer-term sociotechnical effects these uses would create. For instance, Crootof notes how, when discussing autonomous vehicles, the term “driverless cars” both works to normalize something potentially hazardous, but also narrows our sense of what the technology will look like:

“There is no reason to think autonomous vehicles will look or operate anything like existing cars, just as early cars did not look or operate like horseless carriages. An autonomous vehicle need not have a steering wheel or other means of human interaction with the system. And conceiving of autonomous vehicles as driverless cars locks one into a host of existing assumptions, instead of allowing for more imaginative conceptions of what the technology might permit. For example, rather than being individually owned and operated property, autonomous vehicles could operate as connected nodes on a “smart highway” or as a leasable service.”¹³⁰

This can also be seen in the responses to regulating various technologies in the past. For instance, Crootof discusses how, when talking about the internet, different terms evoke radically distinct problem portfolios, and correspondingly distinct regulatory surfaces or levers:

¹²⁵ For instance, on the importance of framings to the international campaigns to ban blinding lasers and anti-personnel mines (and whether these framings can transfer well to ‘killer robots’), see Elvira Rosert & Frank Sauer, *How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies*, 0 CONTEMP. SECUR. POLICY 1–26 (2020).

¹²⁶ Crootof, *supra* note 27 at 17.

¹²⁷ *Id.* at 17.

¹²⁸ Alternatively, they might impute the technology with certain characteristics that certainly do exist, but which are hardly the most distinguishing or salient feature at stake. Or they might unduly emphasize the technology’s characteristics over its use.

¹²⁹ Crootof, *supra* note 27 at 18.

¹³⁰ Rebecca Crootof, *Autonomous Weapon Systems and the Limits of Analogy*, 9 HARV. NATL. SECUR. J. 51–83, 80 (2018).

“The “World Wide Web” conception suggests an organically created common structure of linked individual nodes, which is presumably beyond regulation. The “Information Superhighway” analogy emphasizes the import of speed and commerce and implies a nationally-funded infrastructure subject to federal regulation. Meanwhile, “cyberspace” could be understood as a completely new and separate frontier, or it could be viewed as yet one more kind of jurisdiction subject to property rules and state control.”¹³¹

Similarly, Jordan Branch has suggested that the use of ‘foundational metaphors’ around the internet has strongly shaped U.S. cybersecurity policies; the U.S. military notably framed the internet and related system as a ‘cyberspace’—just another ‘domain’ of conflict along with land, sea, air, and space—which has had strong consequences both institutionally (expanding the military’s role in cybersecurity, and supporting the creation of US Cyber Command), while also shaping how international law is applied to cyber operations.¹³²

To take another example, one might consider ‘data’. Over the last decade, this curt informational concept, once the preserve of staid ‘data science’, has been cast in a diverse array of roles.¹³³ It has been captured under different metaphors—as ‘oil’, ‘sunlight’, ‘public utility’, or ‘labour’—which each have drastically distinct political and regulatory implications.¹³⁴ For instance, the *oil* metaphor, which has until recently been relatively dominant in the US,¹³⁵ foregrounds how, operationally, data must often be ‘refined’ (‘cleansed’ and ‘tagged’) to be useful; how, economically, data is a traded commodity that is owned by whoever ‘extracts’ it; and how, politically, this resource and its infrastructures might become a source of geopolitical contestation between major state actors.¹³⁶ Within the framing of data as a natural resource, even policies to preserve individual user privacy become implicitly couched as competing ownership claims.¹³⁷

Conversely, the *sunlight* metaphor instead emphasizes data as a ubiquitous public resource that ought to be widely pooled and shared for social good, and underpins the ‘Open Data’ movement, calls to ‘democratize’ AI, and the Chinese approach to pooling types of data such as

¹³¹ Crootof, *supra* note 27 at 17–18. In the specific context of autonomous weapons systems, Crootof has similarly discussed the relative utility (and limits) of analogizing such systems to ‘weapons’, ‘combatants’, ‘child soldiers’, or ‘animal combatants’. Crootof, *supra* note 130. See also Michael C Horowitz, *Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systems*, 30 TEMP INT'L COMP 14 (2016).

¹³² Jordan Branch, *What's in a Name? Metaphors and Cybersecurity*, INT. ORGAN. 1–32 (2020).

¹³³ For a broader, critical discussion, see also Luke Stark & Anna Lauren Hoffmann, *Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture*, 1 J. CULT. ANAL. 11052 (2019).

¹³⁴ See also The Economist, *Are data more like oil or sunlight?*, THE ECONOMIST, 2020, <https://www.economist.com/special-report/2020/02/20/are-data-more-like-oil-or-sunlight> (last visited Feb 26, 2020).

¹³⁵ See for instance Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 UNIV. CALIF. LAW REV. 1149–1210, 1149 (2018). (“[a]lgorithms and AI [artificial intelligence] are the machines; Big Data is the fuel that makes the machines run. Just as oil made machines and factories run in the Industrial Age, Big Data makes the relevant machines run in the Algorithmic Society”).

¹³⁶ For a critique, emphasizing that the ‘oil’ metaphor does not provide good guidance to assessing data competitiveness, see HUSANJOT CHAHAL, RYAN FEDASIUK & CARRICK FLYNN, *Messier than Oil: Assessing Data Advantage in Military AI* (2020), <https://cset.georgetown.edu/research/messier-than-oil-assessing-data-advantage-in-military-ai/>.

¹³⁷ For instance, see Lisa Austin, *We must not treat data like a natural resource*, THE GLOBE AND MAIL, July 9, 2018, <https://www.theglobeandmail.com/opinion/article-we-must-not-treat-data-like-a-natural-resource/> (last visited Sep 14, 2020). (“...ownership language portrays data like a natural resource. [...] Within this framework, privacy becomes a competing claim of control – a kind of “ownership” claim to be carefully balanced against the ownership claims of those doing the extracting and generating economic gains”).

health data; however, it too does not mesh well with those seeking to preserve individual data privacy. In turn, the *public utility* metaphor sees data as ‘infrastructure’, and emphasizes the importance of public investment and new institutions such as data trusts, - cooperatives, or personal data stores to provide ‘data stewardship’.¹³⁸ All these stand in contrast to regulatory programs that treat data as *labour*, and which assert the ownership rights of the individuals generating it against what correspondingly becomes seen as the extractive or exploitative practices of ‘surveillance capitalism’.¹³⁹

It is therefore not surprising that the role of analogies in framing policy debates can be seen in a wide range of AI contexts, also. For instance, Kraftt and colleagues have found that whereas definitions of AI that emphasize ‘technical functionality’ are more widespread amongst AI researchers, definitions that emphasize ‘human-like performance’ are more prevalent amongst policymakers, which in their view might prime policy towards future threats.¹⁴⁰ A recent report by the Centre for Security and Emerging Technology found that the precise AI definition which one takes can paint a very different picture of the relative standing and progress of major states in AI.¹⁴¹

One can also see the importance of framings in the debates, in the last two years, around the desirability of ‘openness’ in AI research, given certain capabilities that might have potential for misuse.¹⁴² In February 2019, the research lab OpenAI sparked debate over this question, after it revealed GPT-2, a capable text-generation AI system, and refused to release the full model out of concerns for misuse in fake news and computational propaganda applications.¹⁴³ While some applauded the move, others critiqued this ‘limited release’ strategy, on the grounds that AI research was an ‘scientific endeavour’ (connoting that all results should be open and public on

¹³⁸ The Economist, *supra* note 134.

¹³⁹ Imanol Arrieta-Ibarra et al., *Should We Treat Data as Labor? Moving beyond ‘Free’*, 108 AEA PAP. PROC. 38–42 (2018).

¹⁴⁰ P. M. Kraftt et al., *Defining AI in Policy versus Practice*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 72–78 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375835> (last visited Feb 12, 2020).

¹⁴¹ DEWEY MURDICK, JAMES DUNHAM & JENNIFER MELOT, *AI Definitions Affect Policymaking* (2020), <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Definitions-Affect-Policymaking.pdf> (last visited Jun 3, 2020). For instance, they note that “the competitive landscape varies significantly in sub-areas such as computer vision (where China leads), robotics (where China has made significant progress), and natural language processing (where the United States maintains its lead).” *Id.* at 2. Their paper is particularly interesting for offering a ‘functional AI definition’ that was itself derived through the use of SciBERT, a neural network-based technique for natural language processing trained on scientific literature. *Id.* at 4.

¹⁴² For surveys of such uses, see Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV180207228 Cs (2018), <http://arxiv.org/abs/1802.07228> (last visited Feb 21, 2018); King et al., *supra* note 110; Hayward and Maas, *supra* note 72. Note, there is also some antecedent work on the larger-scale strategic considerations around whether the development of (advanced) AI systems should proceed in an open or closed manner; Nick Bostrom, *Strategic Implications of Openness in AI Development*, GLOB. POLICY 135–148 (2017).

¹⁴³ Alec Radford et al., *Better Language Models and Their Implications*, OPENAI BLOG (2019), <https://blog.openai.com/better-language-models/> (last visited Feb 17, 2019). For some of the subsequent debates, see also Eric Zelikman, *OpenAI Shouldn’t Release Their Full Language Model*, THE GRADIENT, 2019, <https://thegradient.pub/openai-shouldnt-release-their-full-language-model/> (last visited Mar 4, 2019); Hugh Zhang, *Dear OpenAI: Please Open Source Your Language Model*, THE GRADIENT, 2019, <https://thegradient.pub/openai-please-open-source-your-language-model/> (last visited Mar 4, 2019). See also the later assessment by OpenAI; Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models*, ARXIV190809203 Cs (2019), <http://arxiv.org/abs/1908.09203> (last visited Nov 18, 2019).

principle), or on the grounds that any security problems around AI are just like regular computer (cyber)security problems arising from conventional software development (connoting that a ‘culture of disclosure’ would better enable the identification and patching of vulnerabilities).

Clearly, either of these analogies—*scientific endeavour* or *cybersecurity*—are not entirely baseless in describing aspects of AI.¹⁴⁴ Yet the fact that distinct analogies can be supported does not mean they are all equally productive. In this case, an argument could be made (and, of course, contested) that these analogies are at least somewhat misleading or counterproductive in terms of the policy recommendations they foreground. After all, as Dafoe & Shevlane have argued, an examination of the actual balance of whether published AI research is more useful for attackers or defenders depends on a broad range of factors, including the possibility for adequate defensive measures, or the independent discovery of the knowledge outside of the scientific community. When considering these factors, they argue that “[w]hile [in cybersecurity] disclosure of software vulnerabilities often favours defence, this cannot be assumed for AI research.”¹⁴⁵ To be sure, whether the AI community adopts policies on safe disclosure, or emphasizes full transparency, will also be affected by high-profile incidents; but policy debates are equally shaped by the framings in play. This matters, because effective mitigation of certain risks may require revisions or changes to scientific publication norms.¹⁴⁶

All this does not mean that analogies should not be used in grounding new interpretations, but rather that we should be aware of their ‘scope conditions’ and risks. Accordingly, for AI also, the full and diverse range of analogies should be critically examined. Before grounding governance, it is important to clarify what we understand by ‘AI’—how we could have understood it differently, and what political, regulatory, or strategic effects this choice entails for a given treaty or regime.

This suggests that at the international level, even if a certain analogy or interpretation might be sustained, we should be cautious about too rapidly leaping for the first ‘adaptive interpretation’ available, just for the sake of keeping our regime coherent and nominally applicable by going through the motions of ‘development’. There may be situations where it may prove more sustainable or effective, from a longer-term perspective of ensuring the resilience of a regime or treaty, to forego a ‘patch’ through reinterpretation if it would import forced or loaded technology analogies, and instead to hold out for pursuing new regimes or amendment.

To sum up this section’s exploration of *development*: governance disruption by (AI) technology poses significant challenges only if or when regulatory development is slowed or inhibited by conceptual or political features of the technology. The pertinent question is as such

¹⁴⁴ Indeed, one of them reflects the definitions of AI as science, discussed back in section 2.1.1.1.

¹⁴⁵ Toby Shevlane & Allan Dafoe, *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*, in PROCEEDINGS OF THE 2020 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (AIES ’20) (2020), <http://arxiv.org/abs/2001.00463> (last visited Jan 9, 2020). For broader discussion of how to balance openness and prudence in AI development, in the context of tensions between these core principles within the Montréal Declaration for Responsible AI, see also Jess Whittlestone & Aviv Ovadya, *The tension between openness and prudence in AI research*, ARXIV191001170 Cs (2020), <http://arxiv.org/abs/1910.01170> (last visited Jan 22, 2020).

¹⁴⁶ For a discussion of proposed recommendations, see also Abhishek Gupta, Camille Lanteigne & Victoria Heath, *Report prepared by the Montreal AI Ethics Institute (MAIEI) for Publication Norms for Responsible AI by Partnership on AI*, ARXIV200907262 Cs (2020), <http://arxiv.org/abs/2009.07262> (last visited Sep 21, 2020).

not only when or where new AI capabilities might create situations that are disruptive, but also when and whether governance development to patch the ‘holes’ in regimes are conceptually plausible or politically feasible. In situations where such change is hard, this may well produce regime-specific ‘governance destruction’ in the form of gradual erosion.¹⁴⁷ However, before moving to discuss this topic, we will first explore processes of governance *displacement*.

5.4 Displacement

A second major category of governance disruption concerns the ways in which AI could drive *displacement* in the *processes* or instruments of international governance.¹⁴⁸ That is, more even than many previous technologies, it may drive considerable sociotechnical change in the ‘legal’ domain itself.

Of course, the suggestion that AI technology may not only produce new ‘capabilities’ to be regulated, but can also provide new affordances for *the regulators* who produce and enforce law, is hardly new. Indeed, over the last years, new digital technologies, and particularly AI, have seen particularly rapid and enthusiastic uptake in various branches of domestic public administration, judicial decision-making, policing, and the business of government writ large. At the same time, from the perspective of (private) legal practice, AI systems can be used to automate or contribute to a range of routine legal tasks, such as checking contracts,¹⁴⁹ case law research,¹⁵⁰ or the faster provision of legal services.¹⁵¹ Provided with adequate databases, some machine learning systems have even begun to make headway in predicting the outcomes of legal disputes in court.¹⁵² This has led some scholars to anticipate an increasing automation of legal systems,¹⁵³ with legal rules and standards becoming progressively replaced by algorithmically-tailored ‘micro-directives’ that can predict, *ex ante*, what an *ex post* judicial decision would have held in every specific case, and so can offer perfectly clear yet tailored legal clarity for individuals.¹⁵⁴

¹⁴⁷ Maas, *supra* note 13 at 50–55. We will return to this discussion in section 5.5.1.1. (on governance erosion).

¹⁴⁸ This section loosely tracks, and expands on: *Id.* at 43–49.

¹⁴⁹ LAWGEEX, *Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts* (2018), <https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf> (last visited Aug 27, 2020).

¹⁵⁰ See for instance the CARA AI system. CARA A.I., CASETEXT, <https://casetext.com/cara-ai/> (last visited Sep 8, 2020).

¹⁵¹ See for instance the app ‘DoNotPay’. RJ Vogt, *DoNotPay Founder Opens Up On ‘Robot Lawyers’*, LAW360 (2020), <https://www.law360.com/articles/1241251/donotpay-founder-opens-up-on-robot-lawyers> (last visited Sep 8, 2020).

¹⁵² For instance, a system trained on 584 judicial decisions of the European Court of Human Rights managed to predict the outcome of new cases with 79% accuracy. Nikolaos Aletras et al., *Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective*, 2 PEERJ COMPUT. SCI. e93 (2016). Another study achieved an average accuracy of 75% in predicting the violation of nine articles of the European Convention on Human Rights. Masha Medvedeva, Michel Vols & Martijn Wieling, *Using machine learning to predict decisions of the European Court of Human Rights*, 28 ARTIF. INTELL. LAW 237–266 (2020). Although for a general critique of such projects, see Frank A. Pasquale & Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviorism*, UNIV. MD. FRANCIS KING CAREY SCH. LAW LEG. STUD. RES. PAP. (2017), <https://papers.ssrn.com/abstract=3067737> (last visited Sep 8, 2020).

¹⁵³ Alarie, Niblett, and Yoon, *supra* note 8 at 424.

¹⁵⁴ Casey and Niblett, *supra* note 8 at 430; Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 INDIANA LAW J. 1401–1447, 1410–1411 (2017).

That is not to say that such scenarios—or indeed even more modest cases of legal automation—have been uncritically welcomed. Indeed, recent years have seen extensive scholarship exploring the practical, normative, and legal implications of ‘legal automation’ in domestic practice.¹⁵⁵ There are extensive critiques of the near-term prospects for legal automation, given the limits of current machine learning approaches,¹⁵⁶ as well as in light of the various ‘interface design errors’ that often plague hybrid human and AI ‘cyborg justice’ arrangements, and which can produce inappropriate ‘overtrust’.¹⁵⁷ There have been particularly acute cases involving bias in algorithms used for recidivism prediction;¹⁵⁸ and others have highlighted likely human rights violations that are likely when such algorithms are deployed on or to particularly vulnerable populations, such as migrants.¹⁵⁹

More conceptually, others have cautioned that the automation of law enforcement systems might close the loop on the legal qualities of ‘inefficiency’ and ‘indeterminacy’, arguing that these are both key safeguards against the perfect enforcement of laws that were in fact originally drafted on the implicit assumption of a certain degree of lenience or discretion.¹⁶⁰ Finally, there are concerns that the integration of such systems may produce an increasingly opaque ‘algocracy’.¹⁶¹

However, while these technologies therefore no doubt raise diverse grounds of concern, barring a major public backlash, it may well be the case that they see continued development and deployment of automation in public services and administration, potentially signalling a general shift towards ‘technocratic’ regulatory attitudes in domestic legal systems.¹⁶²

Given its foundational implications, this breadth of work and attention on ‘legal displacement’ in domestic legal systems is surely warranted. However, by comparison, there has been much less attention paid to the prospects of such a shift at the international legal level. In *Paper [III]*, I argued that there are three distinct categories of displacement, which will here be discussed in turn.¹⁶³ Specifically, these are (5.4.1) the automation of processes of rule creation

¹⁵⁵ JOSHUA P. DAVIS, *Law Without Mind: AI, Ethics, and Jurisprudence* (2018), <https://papers.ssrn.com/abstract=3187513> (last visited Jun 23, 2018). For a discussion of the implications of unintelligible legal automation for core theories of law—notably HLA Hart’s account which requires critical officials; Joseph Raz’s concept grounded in reason-based tests of legitimacy, or Ronald Dworkin’s work on deep justifications for coercion—see Sheppard, *supra* note 8.

¹⁵⁶ See for instance: Frank Pasquale & Glyn Cashwell, *Four Futures of Legal Automation*, 26 UCLA LAW REV. DISCOURSE 23 (2015); Frank A. Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEORGE WASH. LAW REV. 1–55 (2019); MIREILLE HILDEBRANDT, *Law As Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics* (2017), <https://papers.ssrn.com/abstract=2983045> (last visited Jul 9, 2018).

¹⁵⁷ Crootof, *supra* note 8.

¹⁵⁸ Lauren Kirchner et al., *Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.*, PROPUBLICA (2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (last visited May 24, 2017).

¹⁵⁹ Petra Molnar, *Technology on the margins: AI and global migration management from a human rights perspective*, 8 CAMB. INT. LAW J. 305–330 (2019).

¹⁶⁰ Hartzog et al., *supra* note 8.

¹⁶¹ Danaher, *supra* note 8.

¹⁶² cf. Brownsword, *supra* note 105.

¹⁶³ Maas, *supra* note 13 at 43–49. It should be noted that in *Paper [III]*, I distinguished between just two categories of displacement: ‘The automation of International Law’, comprising both the use of AI in rule creation or

and adjudication; (5.4.2) the automation of monitoring and enforcement of international regimes; and (5.4.3) potential shifts in the ‘regulatory modality’ of international law, resulting in the gradual replacement of some forms of conflict mediation (e.g. arbitration) with other avenues (e.g. bilateral state diplomacy; or unilateral AI-supported ‘lawfare’).

5.4.1 Automation of rule creation, adjudication or arbitration

In the first place, it has long been recognised that new technologies can change the processes by which international law is created. For instance, technologies can *speed up* international law formation. Already in 1973, Louis B. Sohn noted how new communication and travel infrastructures were making treaty negotiation faster and easier.¹⁶⁴ Of course, while promising, this may not be an unalloyed good, because the increased presence of communications technologies has also, in some readings, rendered strategies of ‘lawfare’ more viable.¹⁶⁵ Others have recently argued that the deluge of big data, and the falling thresholds to collecting it, could facilitate the process of collating and consolidating evidence of state practice, fostering an era of ‘Data-Driven Customary International Law’.¹⁶⁶

Along with speed, new communication technologies have also qualitatively altered the processes of international law, for instance by *expanding participation* to more actors. Indeed, in their discussion of the phenomenon of ‘norm cascades’, Finnemore & Sikkink have discussed how the pace and scope of (general) norm acceptance in international relations can be directly traced to technological change, since;

“changes in communication and transportation technologies and increasing global interdependence have led to increased connectedness and, in a way, are leading to the homogenization of global norms [...] the speed of normative change has accelerated substantially in the later part of the twentieth century.”¹⁶⁷

For instance, in a comparison of the drafting processes behind UNCLOS and the Mine Ban Treaty, Gamble and Ku have argued that modern communications technologies greatly increased the ability of NGOs to get involved in international rule creation, resulting in laws that were less narrowly tailored to state interests.¹⁶⁸

Accordingly, the first type of legal displacement envisions AI systems changing the practices of (international) law *creation* or *adjudication*. Ideas to this effect draw an analogy to growing work on the use of AI systems in domestic legal systems, where algorithms are used in support of the adjudication of cases, the monitoring of crimes and violations, or even the prediction

adjudication as well as in monitoring and enforcement; and ‘The Technological Replacement of International Law’. In this section, I have split out the former category for clarity.

¹⁶⁴ Louis Sohn, *The Impact of Technological Changes on International Law*, 30 WASH. LEE LAW REV. 1, 10 (1973).

¹⁶⁵ Charles Dunlap, *Lawfare Today: A Perspective*, YALE J. INT. AFF. 146–154 (2008). As discussed by Crootof, *supra* note 27 at 6–7.

¹⁶⁶ Megiddo, *supra* note 2.

¹⁶⁷ Martha Finnemore & Kathryn Sikkink, *International Norm Dynamics and Political Change*, 52 INT. ORGAN. 887–917, 909 (1998).

¹⁶⁸ John Gamble & Charlotte Ku, *International Law - New Actors and New Technologies: Center Stage for NGOs*, 31 LAW POLICY INT. BUS. 221, 249–251 (2000); Crootof, *supra* note 27 at 6.

of court rulings. While this rise of ‘AI Justice’¹⁶⁹ has received considerable scrutiny and critique even in the domestic realm, could we expect AI systems to see usage in support of international law? One can consider a (5.4.1.1) ‘strong’ scenario, and a (5.4.1.2) ‘modest’ scenario.¹⁷⁰

5.4.1.1 Strong scenario: a global digital arbiter

The ‘strong’ scenario posits that advances in AI could, in time, posit some sort of ‘global digital arbiter’: a queryable legal model that would facilitate the creation of a ‘completely specified’ international law.¹⁷¹ If possible, such systems could serve as a powerful ‘integrator’ of global governance, as they would be able to identify and resolve conflicts between norms, or goals—from human rights to environmental law—in adjudicating a ‘global constitution’—therefore reducing the challenges of fragmentation and regime complexity.¹⁷² At the limit, such systems would promise the fully automated adjudication of international law.¹⁷³

Of course, while an interesting thought experiment or extrapolation of trends, for the purposes of governance this strong scenario may be less relevant. Technically, such a system remains beyond existing machine learning approaches, and might instead require some forms of future ‘high-level machine intelligence’ capabilities.¹⁷⁴ While it is not clear that such capabilities are categorically out of reach—or even more than a few decades away¹⁷⁵—the fact remains that if they were to be created and implemented, their general societal impact would likely far exceed their direct impact on international law.

More pragmatically, proposing the use of an advanced AI system to create or adjudicate international law in this strong sense would be ‘passing the buck’. The ‘problem’ of global governance (from a coherentist or global constitutionalist perspective) is not that we presently cannot conceive of an authoritative body that could reconcile or decide amongst norm conflicts. It is rather that such hierarchical authorities currently only exist at the state level, whereas the international system notably lacks such bodies. There would certainly be distinct and considerable political (and symbolic) sensitivities involved in handing over the international legal order over to a machine, but these seem modest compared to the extant political difficulty of installing *any* final authority above the international system. In that sense, the ‘strong’ scenario of international legal automation might be somewhat of a red herring, since if one or more AI systems were deployed to integrate and resolve the problem of authority at the international level, by far the bigger achievement would be that this was achieved at all, not that it was done in silicate. As such, a far-reaching ‘automation’ of international law might counteract the trend

¹⁶⁹ Richard M Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STANF. TECHNOL. LAW REV. 48 (2019).

¹⁷⁰ Because the latter is more analytically interesting, we will briefly discuss the former first.

¹⁷¹ Along the lines of Benjamin Alarie’s ‘legal singularity’, Alarie, *supra* note 8.

¹⁷² We will discuss the intersection of (even more modest) AI tools with the lens of ‘regime complexity’ later on, in Chapter 7.3.3.2.

¹⁷³ Although interestingly, while a dramatic change, such a ‘full automation’ of international law would not be the same as the ‘replacement’ of international law, since it would still at its core be reliant on presenting norms to—and adjudicating between (human) parties.

¹⁷⁴ In which case global society might have more pressing problems to contend with.

¹⁷⁵ As per the previous discussion in Chapter 2.1.6 (on change from further AI progress).

towards regime complexity; but its prospects appear dim in the near term, and moot in the long-term.

5.4.1.2 Modest scenario: text-as-data tools and ML support system

By contrast, the ‘modest’ scenario of the automation for rule creation is more analytically fruitful. Of course, a first question to ask is whether even this scenario of ‘displacement’ is anywhere on the horizon. Can international law be even partially automated in this more modest manner? Thomas Burri has been broadly sceptical: he argues that domestic legal areas such as tax law are susceptible to legal automation, because AI systems in those areas can draw on large, dense, structured and homogeneous datasets.¹⁷⁶ However, he argues that the key legal ‘datasets’ at the international law level are either far too small (e.g. ICJ decisions), or far too heterogeneous and ambiguous to allow this.¹⁷⁷

By contrast, Ashley Deeks has been more optimistic about the conditions for gradual but eventually wide-spread take-up of AI technology into ‘high-tech international law’.¹⁷⁸ Contra Burri, she argues that in many areas in international law, there are significant digital sources of texts (covering thousands of documents) which provide extensive text corpora that can allow machine learning ‘text-as-data’ tools to perform various functions. She notes how:

“[o]ne key reason to think that international legal technology has a bright future is that there is a vast range of data to undergird it. [...] there are a variety of digital sources of text that might serve as the basis for the kinds of text-as-data analyses that will be useful to states. This includes U.N. databases of Security Council and General Assembly documents, collections of treaties and their *travaux préparatoires*, European Court of Human Rights caselaw, international arbitral awards, databases of specialized agencies such as the International Civil Aviation Organization, state archives and digests, data collected by a state’s own intelligence agencies and diplomats (memorialized in internal memoranda and cables), states’ notifications to the Security Council about actions taken in self-defense, legal blogs, the U.N. Yearbook, reports by and submission to U.N. human rights bodies, news reports, and databases of foreign statutes.”¹⁷⁹

Given this, text-as-data machine learning systems could be trained to generate new treaty texts (e.g. draft extradition treaties),¹⁸⁰ to predict how an international arbitral panel might rule, to gauge the course and likely outcomes of treaty negotiations, or to identify possible treaty conflicts.¹⁸¹ Moreover, far from being limited only to legal texts, Deeks notes how other uses of AI—from systems that aggregate intelligence information about the preferences or negotiation

¹⁷⁶ This may be too easy a reading of the difficulties involved in tax law automation, especially in civil law systems where precedent does not hold the same value. I thank Luisa Scarella for this point. For further discussion, see Joshua D. Blank & Leigh Osofsky, *Automated Legal Guidance*, 106 CORNELL LAW REV. (2021), <https://papers.ssrn.com/abstract=3546889> (last visited Sep 8, 2020).

¹⁷⁷ Burri, *supra* note 1 at 93–95.

¹⁷⁸ Deeks, *supra* note 1.

¹⁷⁹ *Id.* at 596–597. [citing sources]

¹⁸⁰ *Id.* at 605.

¹⁸¹ *Id.* at 604–606, 616–622 628–630.

strategies of negotiation partners, to emotion-recognition systems or social media sentiment analysis—might also play a role in broader diplomatic processes.¹⁸²

Focusing on the *procedural* contributions that machine learning could make to the creation of new international law, she argues that international lawyers in service of a state's foreign ministry in practice have three roles—negotiating agreements; dispute resolution (whether in judicial or arbitral fora); and advising policymakers about the existence and meaning of international law—and argues that all of these can, in distinct ways, gain from automation.¹⁸³

Contrary to the ‘strong’ scenario discussed above, this ‘modest’ image of legal automation does not envision that these lawmaking processes will be entirely given over to machine decision-making, but rather that many actors will come to find significant benefits in using machine learning tools to support human legal and diplomatic decision-making processes.

To be certain, that does not mean this process of displacement will occur rapidly. As noted previously, there are a number of barriers which can slow procurement and integration of AI tools, and which ensure a lag between the development of state-of-the-art tools in labs, and their use in the field.¹⁸⁴ Deeks does admit that international law can often be ‘conservative’ about taking up new technologies, and that at present, states and their advisors in international legal issues still lag far behind the private sector in contemplating how AI could change their work. As a result, with a few exceptions, “international law generally has been a stranger to a new wave of technological tools – including computational text analysis, machine learning, and predictive algorithms – that use large quantities of data to help make sense of the world.”¹⁸⁵ Furthermore, there remain challenges to the incorporation of AI technologies in international law, including technical challenges around the format of some types of international law data; the fact that international law questions have relatively less precedent to guide predictions; and civil liberties concerns over applications in emotion detection or social media scraping software, amongst others.¹⁸⁶

In spite of this, Deeks does identify several trends that will, in her view, steadily increase the pressure on international lawyers to take up and reckon with these new tools. These are: (1) near-peer ambitions; (2) proofs of concept in private law; (3) client pressure; (4) necessity in the line of work.¹⁸⁷ As such, she argues that while the challenges to adoption are real, they are not insurmountable, and that there are potentially many use cases of AI, whether in the preparation of treaty negotiations, conducting negotiations, or identifying customary international law.¹⁸⁸

In addition to these, there might of course be a more fundamental limit to the significance of governance displacement-through-automation: even if there are many areas where there are significant sources for text-as-data tools to be trained on, these may not be the highest-stakes scenarios that international law faces. After all, international law may be especially engaged by

¹⁸² *Id.* at 613–615.

¹⁸³ *Id.* at 589–593.

¹⁸⁴ See also the discussion in chapter 2.1.4 (on barriers and limits to AI), and 2.1.7.3 (on patterns and rates of global diffusion).

¹⁸⁵ Deeks, *supra* note 1 at 576.

¹⁸⁶ *Id.* at 598–599.

¹⁸⁷ *Id.* at 593–597.

¹⁸⁸ *Id.* at 599.

large historical shocks and changes.¹⁸⁹ That would not prevent such AI systems from seeing wide usage in many rote tasks, but it would imply that international legal history would continue to be written by humans for some time.

Nonetheless, even this modest scenario might have significant effects. While some AI tools might support greater coordination, the fact that they are developed and owned by some parties may, as Deeks has argued, mean that “states with a high level of technological sophistication are likely to treat some of these tools as proprietary and critical to their national security, and so may use them in a way that exacerbates existing power differentials.”¹⁹⁰ This could have erosive effects, and markedly alter dynamics of contestation and the legitimacy of the global governance architecture.

5.4.2 Automation of monitoring & enforcement

The second form of governance displacement envisions the use of AI technologies in strengthening the *enforcement* of existing or future international law and global governance instruments.

To be certain, it has long been recognised that new technologies can play a key role in changing the modes and anticipated effectiveness of compliance monitoring and enforcement. Indeed, the role of various forms of ‘National Technical Means’¹⁹¹ in supporting the monitoring of compliance with international commitments, has been well-chronicled in the high-stakes area of arms control. For many decades, a range of technologies including signals intelligence, satellites, networked arrays of seismic, hydroacoustic, infrasound or radionuclide monitoring stations,¹⁹² and (aided by legal arrangements such as the 1992 Treaty on Open Skies) surveillance or radionuclide ‘sniffer’ aircraft,¹⁹³ have all played roles in enabling states parties to monitor and verify each other’s (non)compliance with treaty commitments or various peremptory norms under international law.¹⁹⁴

Notably, such tools are important, not only because they can improve treaty effectiveness by changing the incentives and calculations of actors regarding the expected results of violating

¹⁸⁹ I thank Laura Emily Christiansen for this point.

¹⁹⁰ Deeks, *supra* note 1 at 582.

¹⁹¹ Eric H. Arnett, *Science, Technology, and Arms Control*, 1 in ENCYCLOPEDIA OF ARMS CONTROL AND DISARMAMENT 477–490 (Richard Dean Burns ed., 1993).

¹⁹² See for instance the International Monitoring System (IMS) sensor network which is currently being developed and operated by the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, and which deploys four complementary verification methods (seismic, hydroacoustic, infrasound and radionuclide sensors) at 321 monitoring stations spread over 89 countries, in order to detect any signs of nuclear testing. CTBTO Preparatory Commission, *Overview of the verification regime*, COMPREHENSIVE NUCLEAR-TEST-BAN TREATY ORGANIZATION, <https://www.ctbto.org/verification-regime/background/overview-of-the-verification-regime/> (last visited Sep 9, 2020).

¹⁹³ TREATY ON OPEN SKIES, CTS No 3 (1992), <https://www.osce.org/files/f/documents/1/5/14127.pdf>. See generally also David A. Koplow, *Back to the Future and up to the Sky: Legal Implications of Open Skies Inspection for Arms Control*, 79 CALIF. LAW REV. 421–496 (1991). However, note that in May 2020, President Donald Trump announced an intention for the US to withdraw from the Open Skies Treaty. Bonnie Jenkins, *A farewell to the Open Skies Treaty, and an era of imaginative thinking*, BROOKINGS (2020), <https://www.brookings.edu/blog/order-from-chaos/2020/06/16/a-farewell-to-the-open-skies-treaty-and-an-era-of-imaginative-thinking/> (last visited Jun 22, 2020).

¹⁹⁴ Maas, *supra* note 13 at 43.

a treaty, but also because in some cases, technological means can provide the guarantees by which parties are willing to bind themselves to agreements in the first place.

For instance, Coe and Vaynmann have argued that one critical barrier to reaching arms control agreements is a so-called ‘security-transparency’ tradeoff.¹⁹⁵ This refers to the dilemma whereby an arms control agreement between A and B to cap B’s capabilities must on the one hand have sufficient provisions in place to ensure transparency of B’s actions (to ensure A that any cheating can and will be detected), but on the other hand, must not offer A so much transparency of B’s military actions, that B holds the arrangement to unacceptably erode its security. In some cases, it may be impossible to satisfy both requirements at once, for, as Coe and Vaynman note:

“[a]ny deal that is transparent enough to assure that one side complies with the deal may also shift the balance of power so much that the other side reneges to exploit this shift. Any deal that preserves the balance of power well enough to be safe for the arming side may not be transparent enough to assure the monitoring side of its compliance. When this is true, no arms control deal will be viable.”¹⁹⁶

While the ‘security-transparency’ trade-off might appear to provide grounds for pessimism regarding the prospects of arms control, however, this model also allows that whenever this trade-off is less steep, or where one side can assure itself of the ability to unilaterally monitor the other’s compliance, agreements can be struck. Importantly, this highlights structural ‘levers’ that can be affected, not just through institutional arrangements or confidence-building measures, but also through new technologies, in order to mitigate or sidestep this trade-off. As such, technologies can structurally shift the conditions for an arms control treaty if they increase unilateral compliance monitoring, or if they can reduce the ‘transparency-security’ trade-off.

In the first case, technological advances that *increase either party’s ability to unilaterally monitor* certain capabilities can render a ‘closed deal’ (whereby parties agree on an arms control deal, but refuse to allow access to facilities for the purposes of mutual verification) viable nonetheless, because either party can be assured that they are able to unilaterally monitor compliance. For instance, in the domain of nuclear arms control, Coe and Vaynmann suggest that the Strategic Arms Limitations Talks (SALT I) negotiations succeeded where the earlier 1964 Freeze negotiations had failed, in large part because the development of satellite surveillance in the intervening years had improved unilateral monitoring capabilities sufficiently to make a ‘closed deal’ viable.¹⁹⁷

In the second place, there may be ways to take measures that create mutual perceptions that the *trade-off between transparency and security is mild*. For instance, during the first round of negotiations over the Intermediate Nuclear Forces (INF) Treaty (1980-1983), US negotiators initially proposed ‘anytime anywhere inspections’, which were rejected not only by the USSR but also, before long, by many on their own side.¹⁹⁸ Instead, during the second round of negotiations

¹⁹⁵ Andrew J. Coe & Jane Vaynman, *Why Arms Control Is So Rare*, 114 AM. POLIT. SCI. REV. 342–355 (2020).

¹⁹⁶ *Id.* at 343. (arguing that this trade-off has driven Iraq’s nuclear weapons programs after the Gulf War, great power competition in arms in the interwar period, and superpower military rivalry during the Cold War, noting that this accounts for almost 40% of all global arming in the past 2 centuries).

¹⁹⁷ *Id.* at 352.

¹⁹⁸ *Id.* at 352.

(1985-1987), the US put forward proposals for more limited inspections with access to only one missile production facility. Moreover, the US devised a way to outfit Soviet facilities with a sensor that would reveal whether an exiting missile was of the banned type, without revealing the technical characteristics of non-banned missiles.¹⁹⁹ Since the Cold War, there has been continued scientific research into new avenues to, for instance, validate nuclear warheads that are slated for retirement, in ways that do not reveal key engineering features.²⁰⁰

This highlights the potentially important role that new technologies can play in improving the ability of states and other parties to make verifiable claims about the properties, capabilities, or inputs of new technologies.²⁰¹ Technology can therefore not just incrementally contribute to the deterrent or compliant pull of treaties; it can also, in fact, shift the possibility frontier for which arms control or non-proliferation treaties are politically possible in the first place.

Moreover, it is important to note how since the days of the Cold War, the rise of the digital economy has increasingly complemented ‘national technical means’ with a host of distributed sensors. As a result, over the past decades, the International Atomic Energy Agency (IAEA) has begun to draw on open-source information and commercial satellite imagery in order to provide early warning in nuclear safeguards and to counter nuclear proliferation;²⁰² commercial satellites have also been proposed to monitor the implementation of the Kyoto Protocol’s Global Emissions Trading Program.²⁰³ Moreover, as noted by Livingston & Risse, the emergence and ubiquity of twenty-first century digital technologies, including mobile phones, commercial high-resolution imaging satellites and social media has already begun to enable near-constant surveillance not just by states or corporations, but also by non-state actors such as human rights observers, journalists and open-source (citizen) investigation networks such as Bellingcat or the Syrian Archive.²⁰⁴

Such open-source technologies have begun to play a key role in monitoring various war crimes, as well as human rights violations.²⁰⁵ In at least one case, this has led to an indictment

¹⁹⁹ *Id.* at 352.

²⁰⁰ For recent developments, see for instance Areg Danagoulian, *Verification of Arms Control Treaties with Resonance Phenomena*, 30 NUCL. PHYS. NEWS 25–30 (2020).

²⁰¹ For a discussion of the security-transparency trade-off in arms control regimes for military AI systems themselves—and the possible role of (AI) technologies in alleviating it, see also the later discussion in Chapter 7.1.2.2.

²⁰² Giacomo G. M. Cojazzi et al., *Collection and Analysis of Open Source News for Information Awareness and Early Warning in Nuclear Safeguards*, 50 ESARDA BULL. 94–105 (2013); David Albright, Sarah Burkhard & Allison Lach, *Commercial Satellite Imagery Analysis for Countering Nuclear Proliferation*, 46 ANNU. REV. EARTH PLANET. SCI. 99–121 (2018).

²⁰³ Allison F. Gardner, *Environmental Monitoring’s Undiscovered Country: Developing a Satellite Remote Monitoring System to Implement the Kyoto Protocol’s Global Emissions-Trading Program*, 9 N. Y. UNIV. ENVIRON. LAW J. 152 (2000).

²⁰⁴ Livingston and Risse, *supra* note 6 at 143. STEVEN LIVINGSTON & SUSHMA RAMAN, *Human Rights Documentation in Limited Access Areas: The Use of Technology in War Crimes and Human Rights Abuse Investigations* (2018), https://carrcenter.hks.harvard.edu/files/cchr/files/ccdp_2018_003_humanrightsdocumentation.pdf (last visited Aug 26, 2020). On the ‘multi-use’ of remote sensing satellites, see also Nathan Edward Clark, *Blurred Lines: Multi-Use Dynamics for Satellite Remote Sensing*, 10 J. INT. HUMANIT. LEG. STUD. 171–183 (2019).

²⁰⁵ Ella McPherson, Isabel Guenette Thornton & Matt Mahmoudi, *Open Source Investigations and the Technology-Driven Knowledge Controversy in Human Rights Fact-Finding*, in DIGITAL WITNESS: USING OPEN SOURCE

under international criminal law: in 2017, digital sleuths identified Libyan execution sites, by triangulating geographical features found in propaganda videos that had been posted to social media, leading to an International Criminal Court warrant for the arrest of Mahmoud Mustafa Busayf Al-Werfalli, a Libyan warlord.²⁰⁶ While such analysis is predominantly done by humans at present, it involves distinct tasks that could be automated or enhanced by AI tools. Indeed, one recent experiment involved the development and testing of a machine learning algorithm, trained on synthetic data, to go through footage of airstrikes in order to help investigators identify UK-manufactured cluster munitions that are illegally used in conflict; in tests, this system sped up analysis 100-fold, while reducing risk of trauma for investigators.²⁰⁷ In another project, run by Amnesty International and the Citizen Evidence Lab, researchers deployed machine learning to support the large-scale analysis of satellite data, in order to detect the destruction of human settlements in Sudan's Darfur region.²⁰⁸ There are various other ways machine learning tools could contribute to the detection or investigation of various violations under international law.²⁰⁹

As such, from a more modest perspective, there may be various roles for machine learning approaches—if appropriately and rigorously vetted—in enforcing international law, and in facilitating the negotiation of agreements that are sensitively dependent on the ability to monitor compliance.²¹⁰

5.4.3 The replacement of international ‘law’? AI and shifts in regulatory modality

Notably, even if AI were to be used to change the ‘*input*’ of international law (e.g., the process of treaty negotiation or adjudication), or strengthen its enforcement, this would not

INFORMATION FOR HUMAN RIGHTS INVESTIGATION, DOCUMENTATION AND ACCOUNTABILITY 21 (A Koenig, S Duberley, & D Murray eds., 2019).

²⁰⁶ Boutin, *supra* note 1.

²⁰⁷ Nesta, *Documenting mass human rights violations through collective intelligence*, NESTA (2020), <https://www.nesta.org.uk/feature/collective-intelligence-grants/documenting-mass-human-rights-violations-through-collective-intelligence/> (last visited Jun 29, 2020); Karen Hao, *Human rights activists want to use AI to help prove war crimes in court*, MIT TECHNOLOGY REVIEW, 2020, <https://www.technologyreview.com/2020/06/25/1004466/ai-could-help-human-rights-activists-prove-war-crimes/> (last visited Jun 29, 2020).

²⁰⁸ Milena Marin, Freddie Kalaitzis & Buffy Price, *Using artificial intelligence to scale up human rights research: a case study on Darfur*, AMNESTY INTERNATIONAL & CITIZEN EVIDENCE LAB (2020), <https://citizenevidence.org/2020/07/06/using-artificial-intelligence-to-scale-up-human-rights-research-a-case-study-on-darfur/> (last visited Aug 26, 2020); Julien Cornebise et al., *Witnessing atrocities: quantifying villages destruction in Darfur with crowdsourcing and transfer learning* (2018), https://aiforsocialgood.github.io/2018/pdfs/track1/80_aisg_neurips2018.pdf (last visited Aug 26, 2020). Notably, they did not publicly share the source code of this project, to avoid it being used to target villages.

²⁰⁹ Maas, *supra* note 13 at 44–45. (reviewing several recent use cases of AI tools in predicting the activities of poachers, the ‘Sentry’ system providing civilians in the Syrian conflict through advance warning of incoming airstrikes, and in DNA sequencing for use in forensic investigations of war crimes).

²¹⁰ MARY L. CUMMINGS ET AL., ARTIFICIAL INTELLIGENCE AND INTERNATIONAL AFFAIRS: DISRUPTION ANTICIPATED 3 (2018), <https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>. (“Monitoring the outputs of sensors set up to verify compliance with, for instance, a nuclear, chemical or biological arms control treaty might well be a deadening job for human analysts – albeit one that would require significant specialist training and experience. By contrast, a machine learning system set up to do the same would never tire or grow bored of its task. And while it might (especially in the process of learning) flag a large number of ‘false positives’, by assigning human specialists to oversee and correct it the system’s accuracy would quickly increase.”).

change the nature of the ‘*output*’ of those legal processes (normative rules). ‘Legal automation’ would no doubt alter the texture of international law irrevocably—but it would presumably remain *law*. This would therefore not change the core, normative ‘regulatory modality’ of international law—the instruments through which this system seeks to regulate and change state behaviour.

However, even in a more modest usage, the deployment of AI systems in the service of international law (or the measuring of global governance) could shift the preferences or values of regulators in key ways. As such, AI could drive or facilitate a shift towards novel ‘modes’ of achieving societally desired regulatory outcomes, which no longer rely to the same extent on normative laws. This discussion draws on Lawrence Lessig’s theory of the various ‘regulatory modalities’ of law, social norms, markets, and architecture.²¹¹ For instance, in the domestic law context, Lessig famously chronicled how the architecture of computer ‘code’ enabled US regulators to emphasize different ‘modalities’ in the production and enforcement of law on cyberspace. Specifically, it allowed them to reduce their reliance on normative ‘law’, and instead regulate tech companies in order to embed architectural constraints on online behaviour, in areas such as zoned speech or privacy protection.²¹²

Likewise, Roger Brownsword has argued that new technologies introduce a ‘double disruption’ to law.²¹³ In the first place, they affect the substance of legal rules. Secondly, they drive a regulatory shift, away from seeking to shape behaviour by normative codes or laws and towards the use of non-normative ‘technological management’.²¹⁴ Accordingly, Brownsword has chronicled how emerging digital technologies of social control may even shift the core ‘regulatory attitude’ of authorities, and effect a shift away from traditional ‘legal coherentism’, towards ‘regulatory-instrumentalism’ or even ‘technocracy’.²¹⁵ If a similar shift occurred in the context of global governance, this could also drive value shifts. Speculatively, one can imagine the availability of more powerful AI monitoring capabilities would reduce the need for consensus (so long as there was consensus about trusting the monitoring capabilities).

More directly, Deeks has suggested how the proliferation of certain AI technologies in state departments may also shift how states seek to resolve international disputes. It could lead to more unilateral or strategic uses of this technology. For instance, she notes how “[i]f a state could predict in advance which way a tribunal is likely to resolve a particular dispute, it would allow the state to decide whether to pursue the case or to choose an alternative, such as settling it through diplomatic negotiations or dropping the matter entirely.”²¹⁶

However, there are also more collaborative cases of such ‘replacement’. For instance, there may be situations where algorithms can be used to speed up the negotiation process between states, either as ‘negotiation support systems’ that can propose potential divisions of interests, or

²¹¹ Lawrence Lessig, *The New Chicago School*, 27 J. LEG. STUD. 661–691 (1998); Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. LAW REV. 501 (1999).

²¹² Lessig, *supra* note 211 at 503–505.

²¹³ Brownsword, *supra* note 105 at 6–15.

²¹⁴ See also Roger Brownsword, *In the year 2061: from law to technological management*, 7 LAW INNOV. TECHNOL. 1–51 (2015); Brownsword, *supra* note 8.

²¹⁵ Brownsword, *supra* note 105.

²¹⁶ Deeks, *supra* note 1 at 628.

by helping locate third-party proposals.²¹⁷ However, in such cases, rather than support traditional channels of international law (e.g. arbitration bodies or international courts) through the ‘automation’ of their operations, AI tools would contribute to the ‘replacement’ of such tools as dominant instruments of international conflict resolution or collaboration.

5.5 Destruction

The third and final version of AI-driven governance disruption is found in cases where AI facilitates or contributes to ‘destruction’ of governance architectures either in part or in full, and either locally or globally.

Such governance *destruction* can happen for two reasons.²¹⁸ In the first place, a certain AI issue may prove a particularly *intractable problem* for existing governance regimes to address, because the required ‘developments’ are hard for either conceptual or political reasons, with the result that these regimes experience gradual ‘erosion’ through their sustained inability to correct an increasingly clear mismatch between technological realities and governance instruments. In the second place, AI capabilities may (be perceived to) offer certain capabilities that shift parties’ interests against continued compliance with certain arms control regimes, or even provide them with the tools to challenge multilateralism more directly, contributing to ‘decline’ in the underlying normative or political scaffolding. The former case I refer to as (5.5.1) ‘erosion’; the latter as (5.5.2) ‘decline’.

5.5.1 Erosion: AI as conceptually or politically intractable puzzle for development

As noted in the previous discussion of development, new AI capabilities can, on occasion, produce new affordances (new entities, relationships, or behaviours) in different domains, which can substantively ‘disrupt’ governance instruments for various reasons—they can reveal clear ‘gaps’ or uncertainties and conceptual ambiguities in existing concepts; they can blur the scope of application of existing regimes, or render regimes obsolete (either because the regulated behaviour is rendered very rare, or because it challenges a foundational assumption, or because it makes cost-effective enforcement very difficult); or they can shift the ‘problem portfolio’ of a technology beyond the remit of existing regimes. When such situations occur, they create a need for legal and regulatory systems to adapt to these challenges (development).

However, while domestic legal systems are usually able to carry out regulatory or legislative changes in response to sufficiently urgent situations of disruption, the same cannot be said of international law, where the updating of treaties can be a slow process.²¹⁹ This yields the first version of governance destruction: an ‘erosion’ of certain legal frameworks which seek to regulate certain AI capabilities, but find the requisite development difficult for either conceptual (5.5.1.1) or political (5.5.1.2) reasons. In such cases, just like certain forms of cyberwarfare,²²⁰ AI

²¹⁷ *Id.* at 633–637.

²¹⁸ The discussion in these sections mostly tracks, and expands on, Maas, *supra* note 13 at 50–56. *Paper [III]*.

²¹⁹ Crootof, *supra* note 9.

²²⁰ Glennon, *supra* note 87.

capabilities might create ‘unsolvable puzzles’ (or at least, ‘*unsolved* puzzles’) for global governance, creating situations where the need for legal adaptation is clear to many if not all, but where governance regimes remain largely unable to carry through (meaningfully effective) changes in response.²²¹

5.5.1.1 *Conceptual friction*

As previously noted, the mere creation of a situation that requires ‘development’ of law is not in and of itself beyond the scope of a legal system.²²² Indeed, in principle, legal systems of any stripe would seem perfectly capable of reconfiguration to accommodate nearly any new challenges that can be understood and meaningfully described within natural language. Hypothetically speaking, if one were given complete freedom to alter or redraw disrupted doctrinal categories, or to introduce wholly novel legal categories, standards, or ‘legal fictions’, there would be few or no technology-driven challenges that could not be captured in this way.²²³

In practice, however, matters are obviously more constrained and path dependent. This is because governance development is subject to various limits of language, intelligibility, analogical reasoning, available technological metaphors, and the need to remain in accordance with ‘common sense’.²²⁴ These all limit which avenues of governance development are considered or even viable.

For example, while Crootof has argued that treaty ‘reinterpretation’ is the most common strategy used in order to fold new weapons technologies into an existing governance regime,²²⁵ she also grants that there are situations where an attempted reinterpretation of an existing technology-specific treaty in order to force a fit with a new technology simply can no longer pass the ‘laugh test’.²²⁶ In such cases, attempts to capture patently (or at least apparently) revolutionary technologies under age-old governance instruments or doctrines may exceed all bounds of credibility, threatening the credibility of those governance initiatives, as well as, by extension, the more general legitimacy of the governance system that produced or allowed it (as any critics can add an anecdote to a list of ‘absurdities’ produced by that extant system). In such cases, conceptual or analogical friction means that ‘the law runs out’, and new treaties or law might be needed to carry out development.²²⁷

²²¹ Maas, *supra* note 13 at 55. (discussing conceptual overlap between situations of legal disruption through obsolescence, and legal erosion, but arguing that “the two [...] can be seen as reverse faces of the same coin: for instance, it is a (perhaps the) basic assumption of any prospective legal regime that regulation is in fact politically and operationally possible. If that assumption is no longer justified for a new technology, then the resulting situation can be understood as legal obsolescence with respect to existing legal frameworks, and as legal erosion (or soft legal destruction) with respect to any future or extended legal frameworks which might have been considered to fill the place: the old is no longer fit to serve, but the new is out of reach.”).

²²² See also the caveats in section 5.3.6 (on ‘carrying through development’).

²²³ Which is not the same as saying that they would consequently be efficiently addressed or mitigated; but simply that they would no longer blur the existing law.

²²⁴ Indeed, the need to ultimately ground all legal analyses in understandable ‘folk understandings’ of certain phenomena or criteria (such as criminal intent), may be a categorical challenge for legal scholarship. See for instance: Schlag, *supra* note 116.

²²⁵ Crootof, *supra* note 27 at 13–14.

²²⁶ *Id.* at 15.

²²⁷ *Id.* at 18. On the downside risks of relying on ‘reinterpretation’, even when it appears conceptually possible, given the risks of unreflexive technology analogies, see also the previous discussion in 5.3.6.2 (‘AI at the limits of words’).

However, while domestic legal systems may generally be able to develop new legislation in response to such systems, at the international level, this process can become particularly thorny.

5.5.1.2 Political knots

In addition to these conceptual limits, certain technologies can prove recalcitrant to governance development for practical or political reasons. For instance, they might create a range of capabilities that empower some stakeholders much more than others, creating unequal stakes.²²⁸ Or alternatively, their future trajectory and implications might be very unclear, inhibiting effective regulatory action at an early stage, until the technology has already become firmly embedded at a later stage, and many parties have established stakes, and it is hard to dislodge.²²⁹ The technology might map onto open political fault lines or pervasive ethical disagreements, or have architectural features that mean it simply is too easily disseminated to be constrained by conventional export control mechanisms.²³⁰

For instance, take cyber-weapons: in spite of calls for global regulation of cyberwar,²³¹ it has been argued that traditional arms control regimes are unlikely to be successfully transferred to the realm of cyberspace, because of the differences in how the ‘weapons’ in question are used (specifically the difficulty of attribution), and the relative difficulty of monitoring compliance.²³² In such cases, the reason that there are no global cyberwarfare conventions is not that no problem is perceived to exist,²³³ nor that the existing governance regimes are perceived to be entirely adequate and up to the task. Instead, it combines various conceptual difficulties of classification, with significant underlying political gridlock,²³⁴ which prevent the global resolution of this visible, technology-precipitated problem.²³⁵

²²⁸ Picker, *supra* note 12 at 191–194. (on ‘unevenly shared technology’).

²²⁹ The so-called ‘Collingridge Dilemma’. DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY (1981).

²³⁰ NELSON, *supra* note 7.

²³¹ Mette Eilstrup-Sangiovanni, *Why the World Needs an International Cyberwar Convention*, 31 PHILOS. TECHNOL. 379–407 (2018).

²³² Erica D. Borghard & Shawn W. Lonergan, *Why Are There No Cyber Arms Control Agreements?*, COUNCIL ON FOREIGN RELATIONS (2018), <https://www.cfr.org/blog/why-are-there-no-cyber-arms-control-agreements> (last visited Jan 22, 2018); Glennon, *supra* note 87. Although for a proposed approach see also; Erin D. Dumbacher, *Limiting cyberwarfare: applying arms-control models to an emerging technology*, 25 NONPROLIFERATION REV. 203–222 (2018).

²³³ Although some have argued that the risk of cyberwar has been exaggerated. THOMAS RID, CYBER WAR WILL NOT TAKE PLACE (1 edition ed. 2013). See also the discussion in chapter 4.2.4.

²³⁴ On the topic of global governance ‘gridlock’ generally, see also THOMAS HALE & DAVID HELD, BEYOND GRIDLOCK (2017). On how scientific projects might bypass or dissolve such situations, see Mark Robinson, *The CERN Community; A Mechanism for Effective Global Collaboration?*, GLOB. POLICY (2018), <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12608> (last visited Nov 30, 2018).

²³⁵ For instance, cyberwarfare has only very limited governance at the international level. The Tallinn Manual on the International Law Applicable to Cyberwarfare, while a significant development, only provides non-binding advice on the application of international law in cyberspace, and has only been endorsed by NATO member states. TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS, (2 ed. 2017), <https://www.cambridge.org/core/books/tallinn-manual-20-on-the-international-law-applicable-to-cyber-operations/E4FFD83EA790D7C4C3C28FC9CA2FB6C9> (last visited Sep 9, 2020). Conversely, the Shanghai Cooperation Organisation’s ‘Information Security Agreement’ only has six member states, and failed to get approval at the UN General Assembly.

More generally, it has been argued that along with diverse historical, political, and social factors, a technology's architectural characteristics can render some (weapons) technologies more 'regulation-resistant' than others.²³⁶ For instance, Sean Watts and Rebecca Crootof have articulated a set of criteria which influence whether an outright global ban on a technology is likely to be successful.²³⁷

As a result, attempts to carry out needed 'development' of existing multilateral treaties in order to address the disruptive challenge of a new technology regularly faces the fraught political challenge of securing the agreement of all existing treaty partners, or else risks a fragmentation of the treaty that would endanger the authority, coherence, clarity or appeal of the entire regime.²³⁸ While in some cases, states have shown remarkable pragmatism in reconciling treaties,²³⁹ in general there may be many political challenges to carrying out legal development in response to disruptive technological change, whether the attempted modification is sought through treaty supersession, amendment, additional protocols, or adaptive interpretations.²⁴⁰

This is the case even when the existing regime has mechanisms to enable the modernisation in the face of technological change. For instance, technological progress has often challenged arms exports control regimes, even as those have provisions to enable modernisation. As argued by Nelson, there are certain patterns in regime modernisation efforts.²⁴¹ For instance, she notes that complementing agreements such as export guidelines or the interpretation and application of international law can help modernize provisions or control lists. To be sure, there are examples of non-proliferation and export control regime modernisation, amongst others in the nuclear domain,²⁴² or in the conventional weapons domain in the Wassenaar Arrangement,²⁴³ and the Missile Technology Control Regime.²⁴⁴ At the same time, there also been attempts to increase

²³⁶ Sean Watts, *Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War*, 91 INT. LAW STUD. 83 (2015).

²³⁷ Crootof, *supra* note 9 at 114–115. A similar discussion of criteria that facilitate weapons bans, examined in the context of a prospective ban on military AI, is offered by PAUL SCHARRE, ARMY OF NONE: AUTONOMOUS WEAPONS AND THE FUTURE OF WAR 13 (1 edition ed. 2018). See also the discussion of these criteria in chapter 7.1.1.3.

²³⁸ Brian Israel, *Treaty Stasis*, 108 AJIL UNBOUND 63–69 (2014); As discussed in Crootof, *supra* note 9 at 110–111.

²³⁹ Smith, *supra* note 54.

²⁴⁰ Crootof, *supra* note 9 at 110–111.

²⁴¹ Amy J Nelson, *The Impact of Emerging Technologies on Arms Control Regimes* (2018), <http://www.isodarco.it/courses/andalo18/paper/iso18-AmyNelson.pdf>; NELSON, *supra* note 7.

²⁴² For instance, the 1968 Nuclear Non-Proliferation Treaty (NPT) was revised or supplemented with various additional regimes or institutions, including (1971) the Zanger Committee, formed to implement the NPT requirement, in Article III, that member states adopt export controls; (1975) the Nuclear Suppliers Group, established in the wake of India's first nuclear explosion in order to prevent dual-use technology from being used in a nuclear weapons program; (1991) a Nuclear Suppliers Group (NSG) Control List Update, in response to Iraq's nuclear weapons program; (1997) establishment of broader safeguards agreement under IAEA Additional Protocol. NELSON, *supra* note 7 at 10–11.

²⁴³ Most prominently, the regime saw a 2013 attempt to broaden the scope of controlled items to 'intrusion software'; however, overly broad language caught dual-use software used by industry to improve cybersecurity, leading the US to call for a do-over in 2015. *Id.* at 12.

²⁴⁴ Including a 1992 modernization effort to add unmanned aerial vehicles (UAVs) with a minimum payload and range threshold that would qualify them as delivery systems for nuclear weapons; and a 2016 effort to include missile mission software. *Id.* at 13–14. However, it should be noted that in spring 2020, the US sought to move armed drones off the MTCA control lists. PAUL K KERR, *U.S.-Proposed Missile Technology Control Regime Changes* 3 (2020), <https://fas.org/sgp/crs/nuke/IF11069.pdf>.

cross-regime harmonisation.²⁴⁵ However, often such updates are only carried out successfully in response to a vivid crisis, and updating ‘control lists’ to keep pace with emerging technologies otherwise remains a cumbersome process. In particular, such processes are frequently held back by industry objections over the stifling of innovation, as well as the general problems around understanding or predicting the indirect, second- or third-order effects of certain innovations.²⁴⁶ Finally, more general trends in digitisation have begun to challenge export control regimes across the board, intersecting at certain key operating assumptions with regards to the materiality of the technologies under question.²⁴⁷

In the scenario of legal *erosion*, the regulatory texture of certain AI applications—including their architectural, operational, strategic and political features, along with their ‘problem logic’—can render them a particularly thorny issue for parties to reach meaningful agreement on. As such, even though international law would in principle be capable of the developments necessary to fix the legal gaps or ambiguities created by some AI capabilities, in practice, many of these cases may prove resistant to the sources and tools available to it. As discussed previously,²⁴⁸ customary international law is generally slow and requires unambiguous evidence of state practice (which is not present with ‘hidden’ capabilities such as cyberwarfare AIs, or which would be hindered by definitional problems around ‘AI’).

For their part, new treaties often require that all parties have roughly even stakes in the technology, clear expectations of benefit for abiding by the treaty, the ability to jointly agree on clear definitions and the ability to effectively verify compliance. All of these are difficult in the context of AI development. The highly unequal distribution of leading technological capabilities amongst a narrow set of powerful states, or the need to rely on (self-interested) private-sector for scarce technical expertise, as well as uncertainty over the technology’s long-term impacts, are all features that have historically inhibited swift international consensus or regulation.²⁴⁹ They are likely to do so again for AI, especially given the relative prominence of private actors in this space. Finally, international courts are often slow,²⁵⁰ react to specific cases in front of them, are non-expert on technologies, often have fractured jurisdiction, and at any rate have experienced at least some measure of political backlash in recent years.²⁵¹

That is not to say that international regimes will not be able to carry out some developments to resolve many situations of uncertainty,²⁵² but it may slow or inhibit progress in

²⁴⁵ For instance, some UAVs are on the MTCD control lists, and UAV-related goods can be found on both the Wassenaar Arrangement’s Dual-Use List and Munitions List. Nonetheless, this has not been very effective at halting the diffusion of lethal UAVs. NELSON, *supra* note 7 at 14.

²⁴⁶ Nelson, *supra* note 241.

²⁴⁷ *Id.*; NELSON, *supra* note 7. But see also International Panel on the Regulation of Autonomous Weapons (iPRAW), *LAWS and Export Control Regimes: Fit for Purpose?* (2020), <https://www.ipraw.org/working-paper-diffusion-export-control/> (last visited Sep 25, 2020).

²⁴⁸ In Chapter 2.3.1.

²⁴⁹ Picker, *supra* note 12 at 191–194.

²⁵⁰ For example, the ICJ only provided a (non-binding) Advisory Opinion on The Legality of the Threat or Use of Nuclear Weapons after the end of the Cold War, in 1996—over half a century after the technology’s creation.

²⁵¹ See Mikael Rask Madsen, Pola Cebulak & Micha Wiebusch, *Backlash against international courts: explaining the forms and patterns of resistance to international courts*, 14 INT. J. LAW CONTEXT 197–220 (2018).

²⁵² For reviews of recent activities, see Kunz and Ó hÉigearthaigh, *supra* note 39; Eugenio V Garcia, *Multilateralism and Artificial Intelligence: What Role for the United Nations?*, in THE GLOBAL POLITICS OF

key and high-profile cases. While such legal erosion may therefore not be a categorical threat to international law, it may leave semi-permanent ‘wounds’ in the fabric of global governance. This would have broader implications, however. As noted by Crootof, “[j]urisprudential space junk is not just outdated law that makes it difficult to identify the most relevant regulations. Seemingly ineffectual law on the books creates the perception that international rules have little power, thereby weakening the entire international legal system.”²⁵³ Likewise, the dysfunction, contestation, or inefficacy of global governance instruments in the face of certain clear and present global challenges of AI could well challenge the perceived efficacy and legitimacy of much broader swathes of global governance in the eyes of states, non-state actors, or the global public.

5.5.2 Decline: AI as persistent threat to governance architectures

Finally, and more speculatively, there is a ‘hard’ version of the ‘governance destruction’ argument which, while rarer, would be more foundationally damaging to international law. Such governance ‘decline’ involves new AI capabilities that could threaten the political scaffolding of either specific AI regimes, or more broadly. It could do so by (5.5.2.1) ‘increasing the spoils’ of noncompliance, by (5.5.2.2) creating active avenues for resistance, evasion or contestation, or (5.2.2.3) by shifting values.

5.5.2.1 *Increasing the spoils of noncompliance: decay in political regime foundations*

Regime scholars have long recognised the ways in which the prevailing state of technology acts as one structural factor that shapes state interests, and determines the incentives they have to pursue or participate in the establishment of international regimes.²⁵⁴ However, that argument may cut both ways, if shifts in technologies could in time also erode the structural conditions or grounds for particular regimes. That is not to say that these would immediately trigger the decline of a regime. However, it could result in significant new pressure on specific regimes, as technological innovation in AI shift the incentives of states in various ways that would put sudden or considerable pressure on extant agreements.²⁵⁵ For instance, AI capability advancement could increase the perceived ‘spoils’ of noncompliance with certain regimes (most prominently restrictive bans), in various ways.

In the first place, innovation might radically increase the absolute or relative *gains* which a party perceives it foregoes by continuing full compliance with a ban on restrictive regime. This might occur in many contexts, but pressures might be particularly steep in certain competitive or military domains, where, as Payne has argued, “marginal quality might prove totally decisive” because ‘other things being equal, we can expect higher-quality AI to comprehensively defeat inferior rivals.’²⁵⁶ Such performance gains are never easy to anticipate, but once conceivable or

ARTIFICIAL INTELLIGENCE 18 (Maurizio Tinirello ed., 2020). PHILIPPE LORENZ, *AI Governance through Political Fora and Standards Developing Organizations* 41 (2020), <https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations>. As well as the survey in Chapter 2.3. ²⁵³ Crootof, *supra* note 9 at 128.

²⁵⁴ Arthur A. Stein, *Coordination and Collaboration: Regimes in an Anarchic World*, 36 INT. ORGAN. 299–324, 319–320 (1982).

²⁵⁵ Maas, *supra* note 14 at 141–142.

²⁵⁶ Kenneth Payne, *Artificial Intelligence: A Revolution in Strategic Affairs?*, 60 SURVIVAL 7–32, 24 (2018).

apparently in reach, would steeply increase the perceived (counterfactual) strategic costs states consider they incur by continued compliance with bans or regimes which they acceded to when the (plausible) advantages appeared more modest.²⁵⁷

In the second place, technological change could reduce the barriers to access and proliferation (creating an opportunity for noncompliance). As noted by Nelson, advances in digitisation and additive manufacturing ('3D printing') are already leading to 'regime perforation' of various export control regimes.²⁵⁸ In the case of AI, there are a wide range of developments in synthetic data as well as learning approaches, which could also steadily reduce the barriers to accessing AI capabilities.²⁵⁹ In some cases, this by itself can lead to doctrinal problems. For instance, while the 1967 UN Outer Space Treaty is widely considered as successful,²⁶⁰ its drafters had not anticipated that the costs of the technology would fall so much, nor that non-state actors such as SpaceX and BlueOrigin would ever accrue the resources to undertake space exploration.²⁶¹

Thirdly, innovation in AI could also reduce the anticipated public or political *costs* of noncompliance, either by making violations less likely to be detected or proven, or by making the public reputational repercussions of deploying certain (e.g. 'non-kinetic') AI capabilities relatively low even if they are nominally in violation of treaties.²⁶²

Finally, AI's potential as 'replacement' tool could offer specific capabilities that result in broader legal *decline*. This could be because AI systems might offer strategic capabilities which shift interests; they chip away at the rationales for certain powerful states to engage fully in, or comply with, international law regimes. While they would be unlikely to completely erode interest, AI applications in areas such as surveillance could arguably reduce states' (perceived) dependence on multilateral security regimes to ensure their security from terrorist threats. Such capabilities could increase actor's willingness to 'defect', and erode their willingness to support multilateralism.

5.5.2.2 *AI as weapon enabling general contestation*

In extreme cases, AI systems could enable new strategies by which various actors could directly challenge the legitimacy of international law or its component regimes. For instance, Deeks notes how, amongst various unilateral uses of machine learning, there might be the use of AI tools to measure or even sway local public perceptions against or in favour of certain solutions

²⁵⁷ An extreme version of this scenario, exploring the strategic incentives offered by hypothetical advanced AI capabilities (although concluding that this is likely some distance away) is discussed in John-Clark Levin & Matthijs M. Maas, *Roadmap to a Roadmap: How Could We Tell When AGI is a 'Manhattan Project' Away?* (2020), http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf.

²⁵⁸ NELSON, *supra* note 7.

²⁵⁹ As discussed in Section 2.1.7.1, on lowering thresholds for use.

²⁶⁰ TREATY ON PRINCIPLES GOVERNING THE ACTIVITIES OF STATES IN THE EXPLORATION AND USE OF OUTER SPACE, INCLUDING THE MOON AND OTHER CELESTIAL BODIES, *supra* note 56. See the discussion of the lessons for AI governance in TURNER, *supra* note 48 at 244–247.

²⁶¹ See Verity Harding, *Lessons from history: what can past technological breakthroughs teach the AI community today* (2020), <https://www.bennettinstitute.cam.ac.uk/blog/lessons-history-what-can-past-technological-breakt/> (last visited Aug 17, 2020).

²⁶² Maas, *supra* note 14 at 144–146.

to a conflict.²⁶³ Significantly, such tools can damage the standing of international agreements merely by existing, simply if there is a suspicion about their involvement in certain negotiations, to sway certain diplomatic processes. After all, as Deeks note, the mere proliferation (if not the use) of simple algorithms might “confound dispute resolution and the conclusion of treaties, perhaps because non-technologically advanced states may reject any process that appears to involve the use of text-as-data tools.²⁶⁴

That is not to say that AI tools are the only ‘weapon’ allowing the unilateral manipulation or contestation of multilateral processes—nor that they even are an unusually powerful one. However, they can contribute to such trends.

5.5.2.3 Shift of values

Finally, beyond shifting the calculations of states with regards to continuing to comply with specific regimes ('increasing spoils'), or providing capabilities that allow certain states to better achieve pre-existing desires to contest global legal processes, certain AI capabilities could also *shift* the goals, values or norms of key actors.

That is not to say that AI would be the sole source of these challenges, or that technology alone would be sufficient to shift norms. Rather, in recent years, international legal scholars have begun to note a period of tension for the global legal order. This includes cases of state backlash against international courts,²⁶⁵ dismissive rhetoric from various world leaders, and (threatened or actual) withdrawals from various treaties and multilateral arrangements.²⁶⁶ Most recently, the US withdrawal from the World Health Organization amidst a global pandemic,²⁶⁷ along with apparent US political momentum to withdraw from the World Trade Organization, following continuing spats over that organisation's Appellate Body.²⁶⁸ This has all raised concern over the sustained health of a multilateral rules-based order in the long run. Of course, this trend should not be over-exaggerated, yet its possible intersection with the proliferation of new AI capabilities is salient.²⁶⁹

In the first place, at a *positional* level, AI innovation could simply change the balance of power between actors, asymmetrically empowering those states that are already unsympathetic

²⁶³ Deeks, *supra* note 1 at 637–640. see also Katarina Kertysova, *Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Counteracted*, 29 SECUR. HUM. RIGHTS 55–81 (2018); John Akers et al., *Technology-Enabled Disinformation: Summary, Lessons, and Recommendations*, ARXIV181209383 Cs (2019), <http://arxiv.org/abs/1812.09383> (last visited May 12, 2020).

²⁶⁴ Deeks, *supra* note 1 at 648.

²⁶⁵ Karen J. Alter, James Thuo Gathii & Laurence R. Helfer, *Backlash Against International Courts in West, East and Southern Africa: Causes and Consequences*, 27 EUR. J. INT. LAW (2016); See also broadly Madsen, Cebulak, and Wiebusch, *supra* note 251.

²⁶⁶ James Crawford, *The Current Political Discourse Concerning International Law*, 81 MOD. LAW REV. 1–22 (2018).

²⁶⁷ Michelle Nichols, *U.S. withdrawal from WHO over claims of China influence to take effect July 2021: U.N.*, REUTERS, July 8, 2020, <https://www.reuters.com/article/us-health-coronavirus-trump-who-idUSKBN2482YZ> (last visited Sep 9, 2020).

²⁶⁸ Philip Blenkinsop, *U.S. trade offensive takes out WTO as global arbiter*, REUTERS, December 10, 2019, <https://www.reuters.com/article/us-trade-wto-idUSKBN1YE0YE> (last visited Jul 13, 2020); Keith Johnson, *U.S. Effort to Depart WTO Gathers Momentum*, FOREIGN POLICY (2020), <https://foreignpolicy.com/2020/05/27/world-trade-organization-united-states-departure-china/> (last visited Jul 13, 2020).

²⁶⁹ Maas, *supra* note 14 at 147.

to international norms, enabling them to either outperform, flout or directly challenge the international legal order.²⁷⁰ For instance, some have predicted that AI surveillance capabilities and military tools could be particularly attractive for authoritarian regimes, strengthening their ability to monitor their citizens, and maintaining centralised control over military forces.²⁷¹ There remains considerable contestation over this prospect, however.²⁷² Much of these shifts may depend on how widely available such tools will be.²⁷³ Others have suggested that AI systems can serve as a valuable tool for asymmetric information warfare against democracies,²⁷⁴ deploying attacks that are disproportionately effective against the distinct informational ‘attack surfaces’ and threat models that democracies suffer from more than autocracies.²⁷⁵

In the second place, AI capabilities, along with general trends in digitalisation and changes in media consumption, might simply enable an *increase in the ability of many smaller actors to disseminate norms*, and thereby generate a much greater fragmentation or plurality of views, under which the maintenance of (any) overarching global order becomes increasingly untenable.²⁷⁶

In the third place, and more fundamentally, AI might have a role in *shifting the goals or values of certain parties* that were previously supportive of the global governance order. That is, in certain domains, AI tools might be(come) perceived (whether or not correctly) as a tempting technological substitute for achieving the interests or assurances once secured through give-and-take compromise. For instance, states may have various concrete national interests or goals (domestic security, global stability, prosperity, domestic or global legitimacy or soft power, migration management) which they have long perceived they might only or best achieve through reciprocal compliance with international norms. However, there may be certain new AI capabilities—from enhanced surveillance to new military force projection capabilities, commercial

²⁷⁰ Richard Danzig, *An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order*, 5 LAWFARE RES. PAP. SER. (2017), <https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf> (last visited Sep 1, 2017). See also the discussion in Maas, *supra* note 14 at 147–149.

²⁷¹ Michael C. Horowitz, *Who'll want artificially intelligent weapons? ISIS, democracies, or autocracies?*, BULLETIN OF THE ATOMIC SCIENTISTS, 2016, <http://thebulletin.org/who%20want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692> (last visited May 13, 2017).

²⁷² In a 2019 expert survey, a majority of polled scholars ‘agreed’ or ‘strongly agreed’ with the premise that ‘Technological change today is strengthening authoritarianism relative to democracy’. Foreign Affairs, *Does Technology Favor Tyranny?*, FOREIGN AFFAIRS (2019), <https://www.foreignaffairs.com/ask-the-experts/2019-02-12/does-technology-favor-tyranny> (last visited Mar 11, 2019). However, the idea of an authoritarian ‘AI advantage’ has been contested. Cf. Jaron Lanier & Glen Weyl, *AI is An Ideology, Not A Technology*, WIRED, 2020, <https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/> (last visited Mar 16, 2020).

²⁷³ However, it is likely that many such tools will proliferate relatively fast, as discussed in section 2.1.7. One can also compare the spread of AI-enabled surveillance tools by China: Paul Mozur, Jonah M. Kessel & Melissa Chan, *Made in China, Exported to the World: The Surveillance State*, THE NEW YORK TIMES, April 24, 2019, <https://www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-police-government.html> (last visited Jun 24, 2019); Ross Andersen, *The Panopticon Is Already Here*, THE ATLANTIC, 2020, <https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/> (last visited Sep 9, 2020).

²⁷⁴ Alina Polyakova, *Weapons of the weak: Russia and AI-driven asymmetric warfare*, BROOKINGS (2018), <https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/> (last visited Jan 13, 2019).

²⁷⁵ HENRY FARRELL & BRUCE SCHNEIER, *Common-Knowledge Attacks on Democracy* (2018), <https://papers.ssrn.com/abstract=3273111> (last visited Jan 13, 2019).

²⁷⁶ Which might certainly, to some, be a good thing. Amitav Acharya, *The Future of Global Governance: Fragmentation May Be Inevitable and Creative Global Forum*, GLOB. Gov. 453–460 (2016).

uses, computational propaganda, or the use of AI to model and ‘predict’ other states’ military operations or negotiation strategies²⁷⁷—which such actors can perceive as shortcuts to achieving these goals unilaterally and without costly compromise. This could well shift the instrument choice of leading nations in AI development, which include powerful states on whose buy-in or at least acquiescence multilateral efforts have often been reliant.

Of course, such a gloomy scenario should be tempered. States are not the only actors on the global stage, and even amongst them, many states are not solely driven by interest-based considerations, but also are informed by various norms to support a range of values enshrined in the global legal order. Even if one supposes that certain AI capabilities could strengthen the hand of domestic political coalitions sceptical, hostile or indifferent to international law, this will not necessarily trigger terminal contestation or outright decline. The key consideration might be less whether such value shifts away from multilateralism might occur in a few powerful states, and rather whether such shifts might occur broadly. For instance, Karen Alter has in recent years theorised the future of the international liberal order in the face of a sudden break in US support.²⁷⁸ Drawing on John Ruggie’s argument that international regimes combine fuse power and social purpose,²⁷⁹ she theorizes that even when a previous hegemon retrenches from the international order, momentum can keep that order in place so long as the broader social purpose remains and is shared by many others.²⁸⁰ This is because international law can maintain its appeal and survive the retrenchment of individual states, so long as the broader social purpose of the international liberal order does not lose support.²⁸¹ As such, even if technocratic AI capabilities marginally erode the multilateral commitment of a few powerful states, they may ultimately do little to shift the overall trajectory of the global governance system if other states (especially those with less access to the technology) remain invested in the rule of law.

However, this presumes that other states that do not lead in AI development will retain confidence and trust in international law. Yet as noted previously in the discussion of displacement, the (unilateral) use of AI capabilities in the processes of international law may also erode the legitimacy of the entire system in the eyes of these smaller states and parties also. After all, while the use of AI tools in the processes of global governance can have beneficial effects in speeding up the creation of international law, or strengthening its enforcement, these tools may also have considerable distributional consequences.²⁸² Specifically, if only a few countries have (or retain) access to a set of high-tech tools by which they might predict the rulings of arbitration bodies, identify opportunities to engineer treaty conflicts, identify patterns of state practice that favour their positions, stack treaty negotiations in their favour, or manipulate another state’s media during or after negotiations, then even the mere existence of such capabilities is likely to

²⁷⁷ Deeks, Lubell, and Murray, *supra* note 5.

²⁷⁸ Karen J. Alter, *The Future of International Law*, 101 ICOURTS WORK. PAP. SER. (2017), <https://papers.ssrn.com/abstract=3015177> (last visited Jun 11, 2020).

²⁷⁹ John Ruggie, *International Regimes, Transactions and Change: Embedded Liberalism in the Postwar Economic Order*, in INTERNATIONAL REGIMES 195–232, 198 (Stephen Krasner ed., 1983); Referenced in Alter, *supra* note 278 at 9.

²⁸⁰ Alter, *supra* note 278 at 9.

²⁸¹ *Id.* at 16–17.

²⁸² Deeks, *supra* note 1 at 643–650.

foster a profound sense of unfairness.²⁸³ This might both undermine the negotiation of specific treaties or agreements,²⁸⁴ but also more generally challenge the fiction of sovereign equality amongst states.

However, while all the above might give reason for pause and consideration, it is not meant to be a tale of doom. Ultimately, in spite of the dramatic name, ‘destruction’ is not meant to imply that AI will irrevocably usher in the end of international law, let alone global governance. That might be giving the technology too much credit, certainly for the foreseeable future. Indeed, for one, many of AI’s ‘destructive’ effects discussed here may well be tempered, countered, or offset.²⁸⁵ Moreover, when compared to other ongoing global developments that are currently putting pressure on global governance, the additional contribution of AI systems to processes of destruction might be relatively modest. AI’s societal disruption may be pervasive—but it is not (at least in the near term) as visibly and widely disruptive as is the current global COVID-19 pandemic.²⁸⁶ AI’s role as an ‘erosive’ intractable puzzle may come to threaten the legitimacy of global governance—but perhaps not remotely as much as that architecture’s legitimacy may already be threatened by its inability to effectively address, say, climate change. Finally, AI may offer a variety of actively destructive tools by which to contest international law—but these tools may derive much of their salience from the broader trend towards increasing unilateralism and populist backlash against international law.²⁸⁷ The point of exploring how AI capabilities can drive processes of *destruction* is therefore not to call for international lawyers to forfeit, but instead should urge an examination of responses that can ensure resilience in this changing landscape.

5.6 Conclusion: AI and Governance Disruption

This chapter has covered the second perspective on ‘change’ in AI governance, by examining various changes *in* governance as a result of AI. On this basis, I set out a taxonomy of types of governance disruption created by AI systems.²⁸⁸ Accordingly, this model discussed a variety of ways in which AI can drive changes that require *development* in the substance or norms of governance, the ways it could produce *displacement* in the processes of international law, from the way international law is created or enforced, to the underlying choice of normative law as

²⁸³ *Id.* at 648.

²⁸⁴ *Id.* at 648. (“Use of these tools by one or a few states, but not all states, may foster a sense of unfairness. In extreme cases, states may shy away from acceding to agreements that they believe were not reached by a fair process. Thus, a state’s perception of the other side’s use of these tools may affect its willingness to reach agreement in the first place, or its willingness to ratify the agreement after the fact. It is also possible that the use of machine learning tools to manipulate a state’s media during negotiations might lead that state to allege that the treaty was void because the state was fraudulently induced to conclude.”)

²⁸⁵ For instance, Deeks suggests redistributing various ‘text-as-data’ algorithms to many states, in order to mitigate the distributional effects of their spread. *Id.* at 650–651.

²⁸⁶ Of course, the lack of visibility in an erosive trend may itself warrant a need for caution.

²⁸⁷ Alter, *supra* note 278.

²⁸⁸ See the overview in Table 5.1, in section 5.2. These types of changes could in principle apply to any new technology, though they can be seen particularly well in the context of AI, given its wide range of applications. See also Liu et al., *supra* note 22.

regulatory tool. Finally, it enables an exploration of the way AI capabilities might drive gradual *destruction* of specific regimes or even political pillars of the broader governance architecture.

To be certain, one should not overstate the ubiquity of these dynamics. No technology—not even AI—is so inimical to existing laws that it challenges, blurs, or confuses it at every point of contact. The ecology of global governance is broad and diverse, and consists of many varied institutions, instruments, and norms. Many—and likely most—applications of AI capabilities will likely be absorbed into these instruments, if not entirely smoothly, at least relatively quietly or without much ‘disruptive’ confusion or upset. Yet for better or for worse, the global governance landscape is one that can in some ways be more defined by its failures or public breakdowns, than by its quietly functional backwaters. Moreover, more than many other technologies, AI has a breadth of application that suggests it may serve as the trigger of relatively many simultaneous and intersecting sites of governance disruption. As such, it matters to engage these phenomena, even if they are ‘unusual’ in the daily state of affairs.

Accordingly, ‘governance disruption’ is a valuable lens to keep on hand in exploring strategies and regimes for AI governance, for several reasons: (1) from an instrument choice- or design perspective, it informs more resilient debates over the comparative merits of distinct global governance instruments for AI, keeping in mind that these will have to be able to scale or adapt along with further change; (2) it helps put focus on the ways in which AI tools may themselves change the processes as well as broader political scaffolding of global governance in coming years, in ways that can both afford opportunities for governance regimes (whether for AI or in other areas), but which might also challenge or upend previous assumptions.

Having discussed two conceptual lenses, we now take a brief step away from AI technology, in order to engage with ‘change’ in the broader sense, of ongoing and broader shifts in governance architectures.

Chapter 6. Changes in Governance Architecture: Regime Complexity and AI

The regime complexity lens focuses on patterns and drivers of institutional and normative change in the broader institutional ecologies and governance architectures around AI governance regimes, as well as inter-institutional interactions. I first (6.1.) provide a brief background on the development of regime theory, and (6.2) the articulation of the concept of a ‘regime complex’. I then (6.3) examine the debates around the consequences and desirability of fragmentation or centralisation in a regime complex. Finally, I (6.4) explore how the regime complexity lens frames a research agenda to explore an AI governance regime at five levels—origin, topology, evolution, consequences, and strategies.

We have now two facets of ‘change’ in AI governance. In Chapter 4, the lens of sociotechnical change highlighted how, when, and why technical change in AI capability translates into cross-sector sociotechnical changes, and how these can be approached as rationale or target for governance. In Chapter 5, governance disruption allowed an examination of how AI capabilities change the instruments of global governance.

This sets the stage for the third and final exploration of ‘change’ in AI governance (See Figure 6.1), which emphasizes underlying processes of change in the broader global governance architecture—and dynamics between distinct AI regimes and institutions—which have implications for the viability or organisation of AI governance.¹ This discussion draws on regime theory, and in particular the analytical lens of a ‘regime complex’.²

¹ It should be noted that, to prevent (further) duplication, most of this chapter will articulate the conceptual categories and framework of regime complex theory, with most of the application to the context of AI governance being carried out in Chapter 7 (particularly in sections 7.2 through 7.5).

² For additional theoretical background on the regime complex lens, see also Chapter 3.2.2. This discussion draws in particular on the accounts in Amandine Orsini, Jean-Frédéric Morin & Oran Young, *Regime Complexes: A Buzz, a Boom, or a Boost for Global Governance?*, 19 GLOB. GOV. REV. MULTILATERALISM INT. ORGAN. 27–39 (2013); Benjamin Faude & Thomas Gehring, *Regime Complexes as Governance Systems*, in RESEARCH HANDBOOK ON THE POLITICS OF INTERNATIONAL LAW 176–203 (Wayne Sandholtz & Christopher Whytock eds., 2017), <https://www.elgaronline.com/view/9781783473977.xml> (last visited Oct 15, 2019); Karen J. Alter & Kal Raustiala, *The Rise of International Regime Complexity*, 14 ANNU. REV. LAW SOC. SCI. 329–349 (2018); Laura Gómez-Mera, Jean-Frédéric Morin & Thijs Van De Graaf, *Regime Complexes*, in ARCHITECTURES OF EARTH SYSTEM GOVERNANCE: INSTITUTIONAL COMPLEXITY AND STRUCTURAL TRANSFORMATION 137–157 (Frank Biermann & Rakhyun E. Kim eds., 2020).

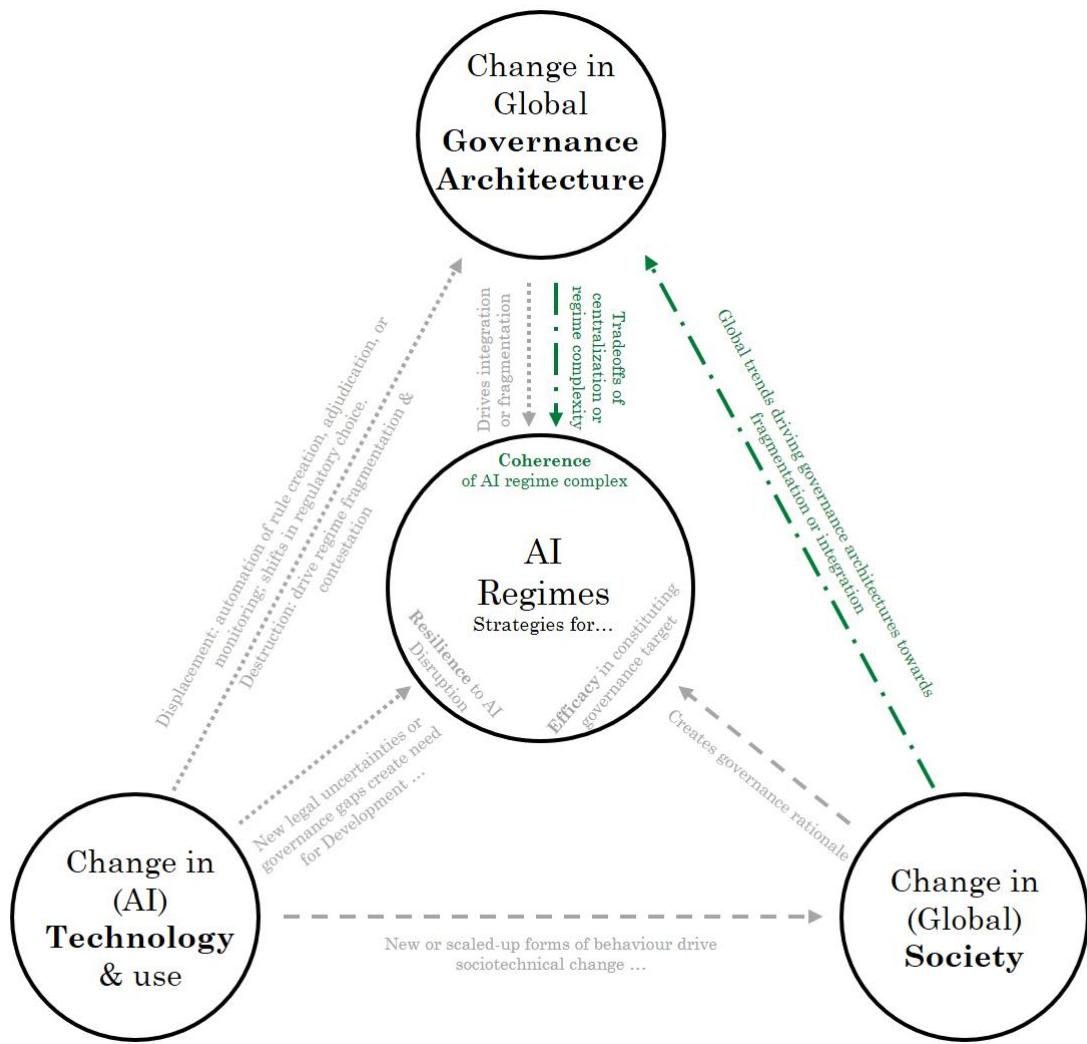


Figure 6.1. Conceptual sketch: Regime Complexity

6.1 Regime Theory and the Fragmentation of Governance

Regime theory originally emerged in the 1970s and 1980s, in response to the unanticipated stability of international economic institutions in the aftermath of World War II.³ Krasner defined such regimes as “sets of implicit or explicit principles, norms, rules and decision-making procedures around which actors’ expectations converge in a given area of international relations”.⁴ Through a combination of norms as well as coordination and enforcement

³ THOMAS GEHRING, DYNAMIC INTERNATIONAL REGIMES: INSTITUTIONS FOR INTERNATIONAL ENVIRONMENTAL GOVERNANCE (1994).

⁴ Stephen D. Krasner, *Structural Causes and Regime Consequences: Regimes as Intervening Variables*, 36 INT. ORGAN. 185–205, 186 (1982).

mechanisms, such regimes can help shape state behaviour within an anarchic international system, enabling cooperation or coordination on specific issues of international concern.⁵

Accordingly, a strand of rationalist research began to examine the conditions under which states seek to create regimes to address discrete problems⁶ as well as what factors affected the resulting design of international institutions.⁷ This research agenda was further catalysed by a pervasive and extended process of global institutional proliferation.⁸ Whereas a century ago there were only a handful of international institutions, today the global governance landscape is populated by thousands of intergovernmental organisations, tens of thousands of nongovernmental organisations, and hundreds of thousands of international agreements.⁹ While some argue that this rate of growth in new treaty and institutional creation has ‘stagnated’ in more recent years,¹⁰ it has nonetheless been extensive enough to transform the tapestry of the global governance architecture.

In particular, the trend of institutional proliferation, as well as the development of ever more specialised regimes, has had far-reaching effects. Rather than being subject to a global code that is coherent across the entire macro-scale ‘global governance complex’, many issue areas¹¹ find themselves covered by a distinct *global governance architecture*, which Biermann, Pattberg, van Asselt and Zelli have defined as:

⁵ See also the (rationalist) account in Arthur A. Stein, *Coordination and Collaboration: Regimes in an Anarchic World*, 36 INT. ORGAN. 299–324 (1982). But also the more recent comparative analysis of both rationalist and constructivist theories of international regimes in Caroline Fehl, *Explaining the International Criminal Court: A Practice Test for Rationalist and Constructivist Approaches*; 10 EUR. J. INT. RELAT. 357–394 (2004).

⁶ For other work on this, see STEPHEN D. KRASNER, INTERNATIONAL REGIMES (1983); KENNETH A. OYE, COOPERATION UNDER ANARCHY (1986), <https://press.princeton.edu/books/paperback/9780691022406/cooperation-under-anarchy> (last visited Jan 21, 2020).

⁷ Barbara Koremenos, Charles Lipson & Duncan Snidal, *The Rational Design of International Institutions*, 55 INT. ORGAN. 761–799 (2001); Barbara Koremenos, Charles Lipson & Duncan Snidal, *Rational Design: Looking Back to Move Forward*, 55 INT. ORGAN. 1051–1082 (2001).

⁸ Kal Raustiala, *Institutional Proliferation and the International Legal Order*, in INTERDISCIPLINARY PERSPECTIVES ON INTERNATIONAL LAW AND INTERNATIONAL RELATIONS: THE STATE OF THE ART 293–320 (Jeffrey L. Dunoff & Mark A. Editors Pollack eds., 2012).

⁹ The true numbers are unclear, and measurements are beset by definitional difficulties. Alter & Raustiala (2018:330) reference “more than 2,400 intergovernmental organizations, 37,000 organizations involved in international politics, and hundreds of thousands of international agreements”, Alter and Raustiala, *supra* note 2 at 330; Referring to Daniel W. Drezner, *The Tragedy of the Global Institutional Commons*, in BACK TO BASICS: STATE POWER IN A CONTEMPORARY WORLD, 284 (Martha Finnemore & Judith Goldstein eds., 2013), <https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199970087.001.0001/acprof-9780199970087-chapter-13> (last visited Jan 30, 2020); OONA A. HATHAWAY & SCOTT J. SHAPIRO, THE INTERNATIONALISTS: HOW A RADICAL PLAN TO OUTLAW WAR REMADE THE WORLD xviii (Advance Reading Copy edition ed. 2017). However, these figures appear based on data from 2003. More recent estimates for 2015, by the Union of International Associations, list 7,657 intergovernmental organizations and 60,919 nongovernmental organizations. UNION OF INTERNATIONAL ASSOCIATIONS, *Yearbook of International Organizations: Volume 5: Statistics*, 2016 2.1. (2016), <http://brill.uia.org/content/4384> (last visited Jan 18, 2020).

¹⁰ J. Pauwelyn, R. A. Wessel & J. Wouters, *When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking*, 25 EUR. J. INT. LAW 733–763 (2014).

¹¹ In this articulation, an ‘issue area’ should be understood as that thematic object (e.g. climate), activity (e.g. trade, security), or subject (e.g. refugees) of legal and political relevance, on which different governance instruments, rules, and jurisdictions converge and overlap. BEATRIZ MARTINEZ ROMERA, REGIME INTERACTION AND CLIMATE CHANGE: THE CASE OF INTERNATIONAL AVIATION AND MARITIME TRANSPORT 37 (2017), <https://www-taylorfrancis-com.ep.fjernadgang.kb.dk/books/9781315451817> (last visited Jan 11, 2020).

“the overarching system of public and private institutions that are valid or active in a given issue area of world politics. This system comprises organizations, regimes, and other forms of principles, norms, regulations, and decision-making procedures.”¹²

Critically, as a result of institutional proliferation, along with the specialisation of regimes, the past few decades have seen an extensive trend of *fragmentation*, which has since then been the subject of active and fierce debate, particularly in international legal scholarship.¹³ This fragmentation of governance architectures has been characterised by Biermann and others as:

“a patchwork of international institutions that are different in their character (organizations, regimes, and implicit norms), their constituencies (public and private), their spatial scope (from bilateral to global), and their subject matter (from specific policy fields to universal concerns).”¹⁴

As such, one thing which the processes of global institutional proliferation and the accordant rise in regime fragmentation have made abundantly clear, is that governance regimes or institutions can no longer be approached in isolation. Instead, for a comprehensive picture on governance, one must reckon with how a patchwork quilt of overlapping and sometimes contradictory regimes converge on the same topics, affecting both law and politics. Accordingly, it should be little surprise that there has been growing attention to the interactions between this panoply of organisations and actors that make up the global governance architecture. In the mid-1990s, Oran Young first developed a preliminary taxonomy of ‘institutional interplay’.¹⁵ Much other work since has examined potentially problematic interactions or conflicts between regimes, in terms of incompatible norms, operations, or impacts that functionally work at cross purposes. Most prominently, scholars working at the intersection of international law and international relations articulated the concept of a ‘regime complex’ to characterize and study these interactions.

6.2 The Concept of a Regime Complex

In an influential 2004 article, Kal Raustiala and David Victor originally defined a regime complex as “an array of partially overlapping and nonhierarchical institutions governing a particular issue-area”.¹⁶ While key, this definition has been critiqued on the basis that it is unclear whether ‘institution’ refers only to formal international organisations, or all elements of

¹² Frank Biermann et al., *The Fragmentation of Global Governance Architectures: A Framework for Analysis*, 9 GLOB. ENVIRON. POLIT. 14–40, 15 (2009).

¹³ MARTTI KOSKENNIEMI & STUDY GROUP OF THE INTERNATIONAL LAW COMMISSION, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law* (2006), http://legal.un.org/ilc/documentation/english/a_cn4_1682.pdf; Gerhard Hafner, *Pros and Cons Ensuing from Fragmentation of International Law*, 25 MICH. J. INT. LAW 849–863 (2004); Raustiala, *supra* note 8.

¹⁴ Biermann et al., *supra* note 12 at 16.

¹⁵ Oran R. Young, *Institutional Linkages in International Society: Polar Perspectives*, 2 GLOB. GOV. 1–23 (1996).

¹⁶ Kal Raustiala & David G. Victor, *The Regime Complex for Plant Genetic Resources*, 58 INT. ORGAN. 277–309, 279 (2004).

a ‘regime’, such as norms.¹⁷ As such, Alter & Meunier have reasserted the broader, Krasnerian conception of ‘regimes’, and extended the term ‘international regime complexity’ as “the presence of nested, partially overlapping, and parallel international regimes that are not hierarchically ordered.”¹⁸ As can be seen, however, both of these definitions emphasize the lack of hierarchical relations amongst elemental institutions as a distinct feature of international regime complexity. By contrast, Orsini, Morin and Young have instead defined a regime complex as: “[a] network of three or more international regimes on a common issue area. These should have overlapping membership and cause potentially problematic interactions”.¹⁹ In so doing, they explicitly omit the ‘non-hierarchical relationship’ criterion, arguing that it is an ‘ambiguous and unnecessary feature’.²⁰ This definition is more focused on potential overlaps amongst regimes or institutions on a particular issue area, and the resulting problematic interactions that can arise in terms of norms, rules, procedures, or impact.

In addition to the ‘horizontal’ dimension of potential thematic overlap, partially overlapping membership of regimes add a ‘vertical’ dimension to regime complexity; this means that, as Gómez-Mera, Morin and Van De Graaf put it, “a regime complex is composed neither of parallel regimes with a clear division of labour, nor of nested regimes embedded within each other like Russian dolls [but] is messier than these neatly organized ideal types.”²¹ This conceptual fuzziness relates to the observation that the ‘fragmentation’ of a global governance system, like ‘regime complexity’, should be considered not a matter of kind, but of degree. Importantly, in the absence of a world government, all governance—and all international law—is more or less fragmented. Indeed, fragmentation, in this reading is “an ubiquitous structural characteristic” of the international system.²² Likewise a ‘regime complex’ is always somewhere on the continuum between fully integrated regime, and a completely fragmented collection of institutions.²³

Of course, it should be clarified that regime complexity is one lens on the structure of global governance, and that competing or complementary perspectives highlight distinct features, and to some extend have distinct policy implications.²⁴ Nonetheless, it can offer a particularly

¹⁷ David J. Galbreath & Sascha Sauerteig, *Regime complexity and security governance*, in HANDBOOK OF GOVERNANCE AND SECURITY 82–97, 84 (James Sperling ed., 2014), <https://www.elgaronline.com/view/edcoll/9781781953167/9781781953167.00014.xml> (last visited Oct 15, 2019).

¹⁸ Karen J. Alter & Sophie Meunier, *The Politics of International Regime Complexity*, 7 PERSPECT. POLIT. 13–24, 13 (2009).

¹⁹ Orsini, Morin, and Young, *supra* note 2 at 29.

²⁰ *Id.* at 31. However, note that there is still a recurring debate over the importance of ‘hierarchy’ to the definition of a regime complex; for instance, Alter and Raustiala reject the definition by Orsini et al., countering that ‘interaction’ need not be part of the definition since it is already ‘intrinsic’ to complex systems, and arguing that “the lack of hierarchy is central and that regime complexes do not necessarily require interaction—what matters is that there is overlapping authority”; Alter and Raustiala, *supra* note 2 at 7.

²¹ Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 140.

²² Biermann et al., *supra* note 12 at 31.

²³ Robert O. Keohane & David G. Victor, *The Regime Complex for Climate Change*, 9 PERSPECT. POLIT. 7–23 (2011). This relates to the fact that In practice, as noted by Gomez-Mera et al., there is no objectively ‘right’ level of analysis that merits the label ‘regime complex’: instead, the concept is a ‘heuristic construct’, to be used to facilitate analysis at the level of analysis most appropriate to a specific research question. Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 139.

²⁴ Rakhyun E. Kim, *Is Global Governance Fragmented, Polycentric, or Complex? The State of the Art of the Network Approach*, INT. STUD. REV., 4 (2019), <https://academic.oup.com/isr/advance-article/doi/10.1093/isr/viz052/5571549>

fertile lens, as it grounds analysis at the ‘meso-level’, between the macro-level of the fragmentation of the entire system of international law, and the micro level of individual inter-institutional ‘interface conflicts’.²⁵ Moreover, this scholarship has explored at length questions of the effects and implications of such regime complexity.

6.3 The effects and importance of regime complexity

What are the outcomes of regime complexity? There is a wide range of theorised effects for global governance and international cooperation.²⁶ For instance, Alter & Meunier anticipate that higher institutional density can lead to diverse effects, including: (1) A greater role for implementation in determining outcomes; (2) greater reliance on bounded rationality in actor decision-making; (3) more social interaction; (4) more forum-shopping; (5) more institutional competition; and (6) more feedback among institutions.²⁷ These effects are clearly diverse, and there is accordingly live disagreement amongst scholars in regime complex theory over the functional consequences and normative desirability of regime complexity. Critics and proponents argue from a range of positions.²⁸

6.3.1 Critiques of regime complexity: dysfunction, inequality, strategic vulnerability

Some legal scholars, especially those writing from a *global constitutionalist* perspective, have taken exception to the increasing ‘fragmentation’ of international law.²⁹ For instance, Fischer-Lescano and Teubner identify four resulting problems: contradictions between individual decisions of international courts and tribunals; rule collisions; inconsistencies between legal doctrines, and conflicts between legal principles.³⁰ Taking a holistic view of the overall governance system, such scholars argue that increasing regime complexity not only introduces adverse effects for governance on the local level, but that this fracturing also undermines the overarching unity and authority of international law.³¹

Other critiques focus on *operational dysfunction*. From a governance perspective, scholars have suggested that regime complexity, and accordant phenomena such as ‘treaty congestion’,³²

(last visited Feb 16, 2020). (“[f]ragmentation points to flat and nonhierarchical structures, polycentricity points to uneven and rugged structures, while complexity points to modular and hierarchical structures”).

²⁵ Christian Kreuder-Sonnen & Michael Zürn, *After fragmentation: Norm collisions, interface conflicts, and conflict management*, 9 GLOB. CONST. 241–267 (2020).

²⁶ This section draws a lot on the review by Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 144–147. As well as those by Faude and Gehring, *supra* note 2 at 188–193; Alter and Raustiala, *supra* note 2 at 338–341.

²⁷ Alter and Meunier, *supra* note 18 at 15–21.

²⁸ The application or implications of these debates for the organization and efficacy of the AI governance regime will be worked out at greater length in the next Chapter, in section 7.4 (‘Consequences’).

²⁹ Faude and Gehring, *supra* note 2 at 182.

³⁰ Andreas Fischer-Lescano & Gunther Teubner, *Regime-Collisions: The Vain Search for Legal Unity in the Fragmentation of Global Law*, 25 MICH. J. INT. LAW 999–1046 (2004); As discussed in Faude and Gehring, *supra* note 2 at 180–181.

³¹ Hafner, *supra* note 13 at 854.

³² Don Anton, “Treaty Congestion” in *International Environmental Law*, in ROUTLEDGE HANDBOOK OF INTERNATIONAL ENVIRONMENTAL LAW (Shawkat Alam et al. eds., 2012), <https://www.taylorfrancis.com/books/9780203093474> (last visited Oct 7, 2019).

can drive a host of operational problems. They can introduce confusion over authority, unclear boundaries, and rule uncertainty, which in turn may lead to reduced accountability for actors and lower compliance with international commitments.³³ Within the regime complex, institutional interaction may be hallmark by inefficiencies,³⁴ duplication of efforts,³⁵ conflicts in objectives, obligations or procedures, or even antagonistic competition or rhetorical bickering.³⁶

In addition, some scholars critique the perceived *power inequality* of a fragmented regime complex. Such work suggests that, rather than promote inclusion and access, the presence of multiple overlapping institutions may strengthen already-powerful actors, who are better able to navigate and support representation at all these fora, while disadvantaging weaker actors who are less able to carry the participation costs of monitoring all governance developments or attending all these simultaneous meetings. This could perversely exacerbate power imbalances,³⁷ and undermine democratic participation and the legitimacy of the resulting governance arrangements.³⁸

Most importantly, there are concerns over the *strategic vulnerabilities* which a fragmented regime complex creates, and the way these can or are exploited by states. As Gehring & Faude argue, at their core, “[r]egime complexes create new opportunities for strategic action through forum shopping, regime shifting, and the creation of a competing institution intended to induce strategic inconsistency with an already existing institution’.³⁹ Others have noted that states have on occasion deliberately created treaty conflicts in order to facilitate certain changes in multilateral regimes.⁴⁰

6.3.2 Defences of regime complexity: flexible problem-solving, democratic dynamics

In contrast, other scholarship has countered that ‘fragmented’ global governance may, for all its warts, suffice as ‘good enough’ global governance.⁴¹ Orsini, Morin & Young suggest that rather than a sign of dysfunction, “problem solving is enhanced in a context of regime complexes, even if the complex is fragmented, because the existence of a complex means that potential problems are likely to be sorted out.”⁴²

³³ Raustiala, *supra* note 8.

³⁴ Alter and Meunier, *supra* note 18; Biermann et al., *supra* note 12.

³⁵ Amandine Orsini, *The negotiation burden of institutional interactions: non-state organizations and the international negotiations on forests*, 29 CAMB. REV. INT. AFF. 1421–1440 (2016).

³⁶ Galbreath and Sauerteig, *supra* note 17 at 88–89.

³⁷ Eyal Benvenisti & George W Downs, *The Empire’s New Clothes: Political Economy and the Fragmentation of International Law*, 60 STANFORD LAW REV. 595–632 (2007); Daniel W. Drezner, *The Power and Peril of International Regime Complexity*, 7 PERSPECT. POLIT. 65–70 (2009).

³⁸ Amitav Acharya, *The Future of Global Governance: Fragmentation May Be Inevitable and Creative* Global Forum, GLOB. GOV. 453–460, 457 (2016).

³⁹ Thomas Gehring & Benjamin Faude, *The Dynamics of Regime Complexes: Microfoundations and Systemic Effects*, 19 GLOB. GOV. 119–130, 126 (2013).

⁴⁰ SURABHI RANGANATHAN, STRATEGICALLY CREATED TREATY CONFLICTS AND THE POLITICS OF INTERNATIONAL LAW (2014), <https://www.cambridge.org/core/books/strategically-created-treaty-conflicts-and-the-politics-of-international-law/55EACC81A929FC19216E3380D2E9DF69> (last visited Jun 18, 2020).

⁴¹ Stewart Patrick, *The Unruled World: The Case for Good Enough Global Governance*, 93 FOREIGN AFF. 58–73 (2014). But see also the critique by Dan Plesch & Thomas G. Weiss, *1945’s Lesson: “Good Enough” Global Governance Ain’t Good Enough*, 21 GLOB. GOV. 203 (2015).

⁴² Orsini, Morin, and Young, *supra* note 2 at 35.

Indeed, some even go further, and assert that fragmented governance is not just ‘good enough’. Instead, they argue that what they call ‘polycentral’ governance may in fact be a positively superior model of governance that is preferable to ‘monocentric’ regimes.⁴³ Like its critics, the proponents of regime complexity reason from a variety of angles.

Some argue that regime complexity is *problem-solving*. These scholars argue that the redundant overlap in competences makes it less likely that issues will get overlooked because of blame avoidance.⁴⁴ Regime complexes are also thought to have “greater flexibility (across issues) and adaptability (across time) over single, legally-integrated regimes”.⁴⁵ Specifically, organisational ecology theory suggests institutional competition could foster beneficial adaptation, as competing institutions specialize in niches—corresponding to certain issue area themes—or to specific governance tasks (e.g. generating knowledge, strengthening norms, building capacity, or monitoring and enforcing compliance with rules).⁴⁶ These scholars therefore expect that “the invisible hand of a market of institutions leads to a better distribution of functions and effects”.⁴⁷ Moreover, global ‘legal pluralists’ see overlapping jurisdictions as desirable because they can be ‘a source of innovation’⁴⁸ in legal (substantive or procedural) norms. Analogously, it is suggested that polycentric governance provides more opportunities for policy experimentation.⁴⁹

Along with those arguments, others argue that regime complexity creates better *dynamics* amongst governance actors. Rather than putting all ‘eggs in the one basket’ of an authoritative institution, which might ossify or fail to adequately adapt, polycentric governance is held to create improved opportunities for inter-institutional learning.⁵⁰ Others suggest it might increase communication and interaction across parties and institutions, helping foment mutual trust.⁵¹

Moreover, and in direct contrast to the concerns over power asymmetries and democratic deficit, above, some of these proponents also expect that polycentric governance can allow for *greater democratisation*, because of the decentralisation of authority and the (potential) inclusion of more actors in diverse sets of regimes. Indeed, it is pointed out that the strategic flexibility afforded by fragmented regime complexes can also strengthen democratic representation: the same strategies of forum-shopping or regime shifting which can be used in ‘non-cooperative’ ways by powerful (state) actors to bypass or hollow out governance, are also available to less powerful

⁴³ Elinor Ostrom, *Polycentric systems for coping with collective action and global environmental change*, 20 GLOB. ENVIRON. CHANGE 550–557 (2010).

⁴⁴ Aynsley Kellow, *Multi-level and multi-arena governance: the limits of integration and the possibilities of forum shopping*, 12 INT. ENVIRON. AGREEM. POLIT. LAW ECON. 327–342 (2012); Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 145.

⁴⁵ Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 145; See also Keohane and Victor, *supra* note 23 at 19.

⁴⁶ Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 145.

⁴⁷ Fariborz Zelli & Harro van Asselt, *Introduction: The Institutional Fragmentation of Global Environmental Governance: Causes, Consequences, and Responses*, 13 GLOB. ENVIRON. POLIT. 1–13, 7 (2013).

⁴⁸ Paul Schiff Berman, *A Pluralist Approach to International Law*, 32 YALE J. INT. LAW 301–329, 321 (2007). (“normative conflict among multiple, overlapping legal systems is often unavoidable and might even sometimes be desirable both as a source of innovation and as a site for discourse among multiple community affiliations”).

⁴⁹ Alter and Meunier, *supra* note 18 at 19; Ostrom, *supra* note 43.

⁵⁰ Sebastian Oberthür, *Interplay management: enhancing environmental policy integration among international institutions*, 9 INT. ENVIRON. AGREEM. POLIT. LAW ECON. 371 (2009).

⁵¹ Ostrom, *supra* note 43.

actors, and as such can allow civil society or even institutions to route around obstructions thrown up in one forum,⁵² or highlight issue linkages, ultimately affording them greater voice.⁵³ Finally, some argue that fragmented global governance may simply be appropriate, as it reflects the reality of plural values amidst an “inevitable transition from the unipolar moment to a multiplex world.”⁵⁴ As such, it has been suggested that AI governance regimes should not pursue harmonisation, but become more reflective of the plurality of value systems and value judgments.⁵⁵

6.4 Analysing a Regime Complex

The value of the regime complex lens is that it allows analysis of a number of key features or facets of a governance architecture for any specific issue area. Drawing on analytical frameworks developed in this literature,⁵⁶ I argue that this lens highlights five key aspects of any governance system: *origins, state, evolution, consequences, and strategies*. These are five aspects that will be valuable to explore for scholars or proponents of AI governance.

In the first place, examining the *origins* of a prospective AI regime can help understand whether, for a given AI issue that has not yet been subject to a regime, states or other parties may see sufficient cause to create (sets of) new regimes. Is a meaningful regime even viable? Should we expect the emergence of a regimes or regime complexes for a specific AI issue, or may those issues remain in a ‘non-regime’ state? Such an assessment can draw on ‘outside-view’ comparisons with the governance of other technologies or issue areas. But from an ‘inside-view’, it can also draw on established scholarship on the causes and origins of specific regimes or institutions. Such an analysis can draw on more traditional rationalist accounts of the potential *functions* that a regime can fulfil, in terms of the (state) interests such a regime could meet. But from a constructivist perspective, it can also consider possible avenues of *norm* creation and development, by exploring the ways in which ‘norm entrepreneurs’ such as epistemic communities or transnational issue networks can shift state perceptions and preferences. This can yield insights, not just in the ‘viability’ of successful governance for a certain issue, but also in comparing the desirability and sufficiency of distinct strategies.

⁵² On the related concept of ‘institutional bypasses’, see also Mariana Mota Prado & Steven J. Hoffman, *The promises and perils of international institutional bypasses: defining a new concept and its policy implications for global governance*, 10 TRANSNATL. LEG. THEORY 275–294 (2019).

⁵³ Laura Gómez-Mera, *Regime complexity and global governance: The case of trafficking in persons*; 22 EUR. J. INT. RELAT. 566–595 (2016). (discussing the ‘cooperative’ use of regime shifting, forum linking, and other cross-for-a strategies by IGOs, civil society organizations, and other principled actors, to promote their normative mandates).

⁵⁴ Acharya, *supra* note 38 at 460.

⁵⁵ Han-Wei Liu & Ching-Fu Lin, *Artificial Intelligence and Global Trade Governance: A Pluralist Agenda*, 61 HARV. INT. LAW J. (2020), <https://papers.ssrn.com/abstract=3675505> (last visited Sep 26, 2020).

⁵⁶ For instance, Laura Gómez-Mera, Jean-Frédéric Morin, and Thijs van De Graaf have argued that work on regime complexity has, variably, aimed at the investigation of the (1) causes and origins of regime complexes; (2) their evolution over time; and (3) the consequences and effects (and consequently the normative implications) of high regime complexity, and ways to manage these. Gómez-Mera, Morin, and Van De Graaf, *supra* note 2 at 141–147. While departing from this typology, I make a few additional distinctions, as identified.

In the second place, this lens allows one to examine the extant *topology* (state) of a given AI governance architecture at a given moment. This allows a snapshot description of the size, density and organisation (that is, level of centralisation) of networks of treaties, institutions and norms that pertain to a given AI issue.⁵⁷ In turn, this allows an initial exploration of gaps, tensions, or synergies between its elements, whether at the level of norms, goals, institutions or impacts.

Thirdly, this framework allows exploration of the *evolution* of regime complexity over time, either to shed light on past developments and trends,⁵⁸ or, more speculatively, to understand both regime-specific and general factors affecting regime complexity, in order to extrapolate various near-future trajectories of fragmentation or integration for an AI regime. Specifically, we can explore such potential trajectories by considering various general factors that have been theorised to drive further fragmentation; by complementing this analysis with the impact, on regime complexity, of AI governance disruption, and by considering eventual patterns of regime complex consolidation over time.

Fourthly, the regime complex lens allows a structured study of (4) the *consequences* of fragmentation and high regime complexity. That is, drawing on lessons from other international regimes, one is able to anticipate, whether in analytical or normative terms, some of the implications of distinct trajectories of AI regime complexity and various forms of inter-institutional interaction. Whether the regime remains fragmented or will become centralised may, as discussed, have significant implications for the overall legal coherence, efficacy, or legitimacy of the resulting AI governance architecture. This in turn can inform advance debates over what may be the trade-offs or risks involved in various trajectories (whether decentralised or centralised), and accordingly which of these scenarios would be optimal or acceptably functional for governing distinct AI issues.

Fifthly, taking the above steps into consideration can enable a tentative exploration of *strategies* that can be used to manage the resulting AI governance architecture in more productive ways. In a fragmented regime complex, this can involve strategies aimed at managing the interactions of these diverse institutions better, in ways that avert conflict and increase abilities for accommodation or synergy. For specific AI issues or types of change which require a centralised regime, such strategies can involve various considerations of institutional design, in order to avert or mitigate some of the risks of centralisation. By coupling various strategies developed within regime complex scholarship to the insights drawn from the ‘sociotechnical change’ and ‘governance disruption’, this yields a set of tentative recommendations to enable AI governance regimes—of either form—to be more effective, resilient or coherent.

⁵⁷ Kim, *supra* note 24.

⁵⁸ For instance, see Caroline Fehl’s exploration of the evolution of institutional inequalities in the arms control regime. Caroline Fehl, *Unequal power and the institutional design of global governance: the case of arms control*, 40 REV. INT. STUD. 505–531 (2014).

6.5 Conclusion: AI and regime complexity

This chapter has briefly introduced and sketched the third and final perspective on ‘change’ in AI governance: changes *for* AI governance that derive *from* broader trends and dynamics in governance architectures. To do so, it drew from the scholarship on ‘regime complexity’, to lay out a rough research programme of questions that can and should be productively considered in mapping the course, state, and development of AI governance regimes, including potential obstacles and challenges.

The regime complexity perspective is useful for offering a lens on governance systems that considers the ‘messy’ political interactions and strategic moves of various actors within an institutional landscape that is not frozen but quite dynamic. In this way, it contributes a *political* lens to the theoretical sociotechnical change perspective, and the legal perspective on governance disruption. In doing so, it will structure much of the analysis and examination of AI governance in the next and final chapter.

The three Chapters in Part II have presented three perspectives on ‘AI governance under change’. These lenses—on sociotechnical change, governance disruption, and regime complexity—can each individually be applied to problems and questions in AI governance. However, they can also be used into a complementary fashion. As such, the final step of this investigation will apply insights from all three theories to an exploration of various facets of the emerging AI governance regime. It will in turn explore the five analytical perspectives on a governance system (origin, topology, evolution, consequences and strategies), developed in this chapter, in order to organize this analysis. Such an exploration cannot be comprehensive, and does not seek to be definitive, but rather is aimed at drawing out the strengths and limits of these conceptual lenses in coming to terms with the complexity of global governance for AI.

Part III. FRAMEWORKS FOR CHANGE

Chapter 7. AI Governance in Five Parts

The sociotechnical change, governance disruption, and regime complexity lenses can each be used to reframe key conceptual, strategic, and political questions or choices in AI governance. I show how these lenses can offer insight into five questions that can be asked of diverse existing, emerging, or proposed AI governance systems in distinct domains. In terms of (7.1) *origins*, these lenses allow consideration of an AI regime's purpose, viability, and design considerations. In terms of (7.2) *topology*, these lenses allow an examination of the state and degree of normative or institutional fragmentation or integration of an AI governance architecture at a certain moment. In terms of (7.3) *evolution*, these lenses enable the examination of potential trajectories of the AI regime complex, by considering how external governance trends as well as governance disruption can serve as drivers of regime complex integration or fragmentation. In terms of (7.4) *consequences*, these lenses enable us to consider the political, institutional and normative challenges and trade-offs that the AI governance architecture may face, conditional on whether it remains fragmented, or whether it is integrated. Finally, in terms of (7.5) *strategies*, these three lenses highlight potential shifts in conceptual approach, instrument choice, or instrument design, which might enable the AI regime complex to ensure efficacy, resilience, and coherence in the face of these patterns of change.

Throughout the three chapters of Part II, I have drawn on various fields of scholarship in order to introduce and explore three perspectives on change that are relevant to AI governance.

Chapter 4 introduced the lens of *sociotechnical change*, as a thinking tool to enable scholars and policymakers to consider when and where changes in AI technology or capabilities enable sociotechnical changes in (international) society; when and why these changes create a *rationale* for governance interventions; and what material features and problem logics define its texture as a governance *target*.

Chapter 5 introduced the lens of *governance disruption*, a taxonomy that helps examine when and where changes in AI capabilities might affect changes in the substance, processes, or instruments of international law.

Chapter 6 briefly introduced the lens of *regime complexity*, which helps structure analysis in considering when, how, and why changes in the broader global governance architectures—and interactions amongst individual AI regimes or institutions—could shift conditions for AI governance.

Together, these three lenses provide complementary perspectives on how AI regimes are affected and shaped—directly and indirectly—by these three drivers of (socio)technical, legal, normative, and political change (see Figure 7.1).

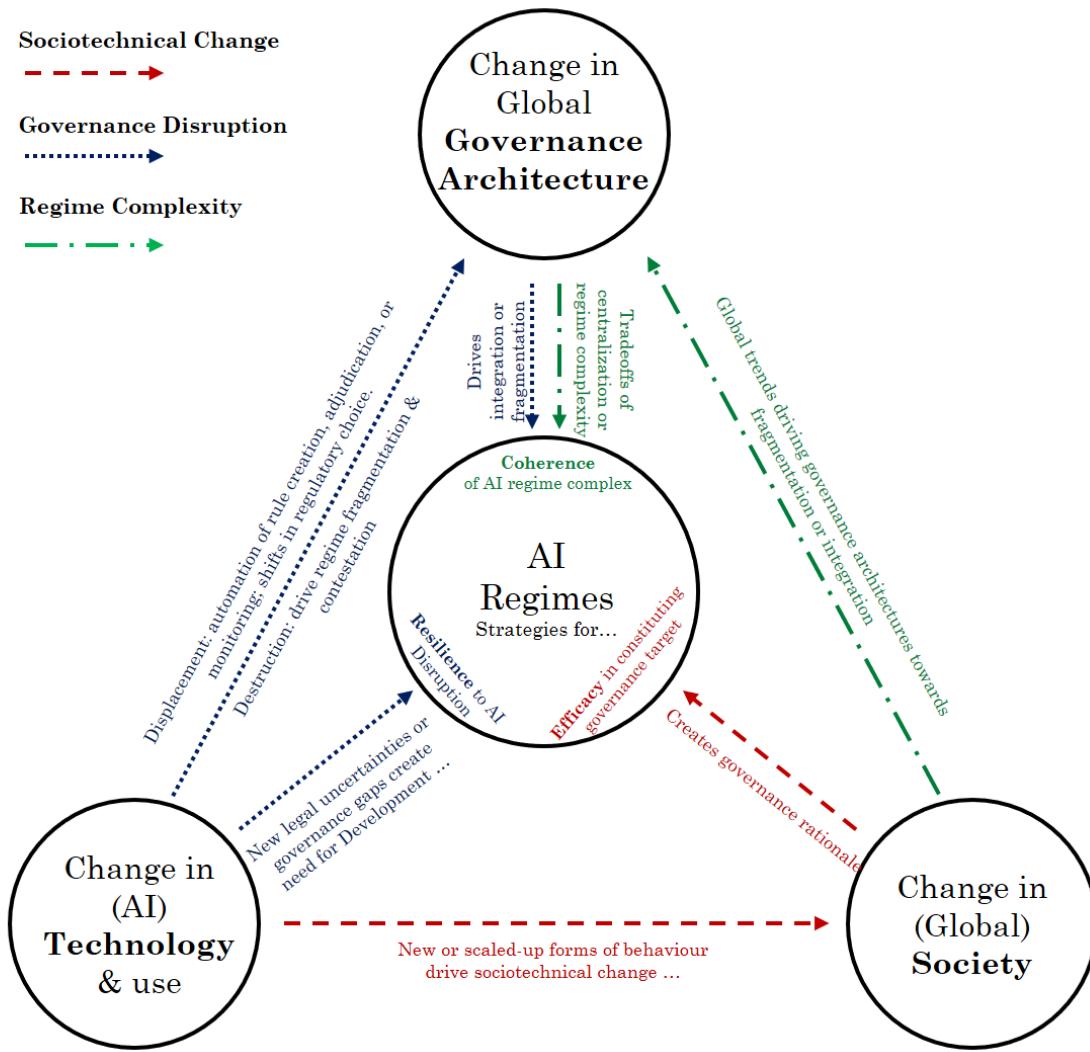


Figure 7.1. AI Governance and Three Facets of Change (revisited again)

Having sketched these three perspectives, we can explore their implications for existing, emerging or proposed AI regimes. This chapter accordingly charts a series of considerations, opportunities, challenges and trade-offs for AI governance to take stock of.

The aim here is not to provide exhaustive analysis or conclusive answers, but to highlight sets of questions that can be asked about initiatives in the AI regime complex, and indicative implications of distinct forms of organisation. As such, this chapter will explore how an examination of AI governance systems or architectures—whether existing, emerging, or proposed—can and should engage with patterns, trends or shocks of ‘change’ at the five levels of *origins, topology, evolution, consequences, and strategies*. At each of these levels, the three facets of change explored in this research can enrich analysis, and improve choices of policy and strategy options (see Table 7.1).

	Theme	Questions
Origins Of individual regimes	Purpose: Is a regime needed?	<ul style="list-style-type: none"> • What are the underlying technological developments? • What (anticipated) sociotechnical change do these enable? • What governance rationales are raised? (e.g. market failures; risks to human health; moral interests; social solidarity; democratic process, or international law itself) • What material features and problem logics characterize this governance target?
	Viability: (why) is any regime viable?	<ul style="list-style-type: none"> • From a comparative historical perspective, were past regimes for similar (technological) challenges viable? • Which functions would this regime serve? What (state) interests would it meet? • How might various actors shift norms to render it (more) viable?
	Design: what regimes would be optimal or adequate?	<ul style="list-style-type: none"> • What strategy? (e.g. reliance on (1) deterrence or (2) gradual norm development; (3) extension of existing regimes; (4) new regime) <ul style="list-style-type: none"> ◦ If new regime, which type? (full ban or regulatory treaty?) ◦ Differential resilience to governance disruption?
	Demographics	<ul style="list-style-type: none"> • Size and composition of network (what are the applicable norms or treaties, active institutions or governance initiatives?)
	Organisation of network	<ul style="list-style-type: none"> • Density of institutional network (number of membership overlaps; institutional contact points on AI issue area) • Links: relating to norms, goals, impacts or institutional relations.
	Interactions and outcomes of linkages	<ul style="list-style-type: none"> • Gaps: functional non-regime; issue unaddressed • Conflictive: active norm conflicts, operational externalities, turf wars • Cooperative: loose integration, but norm relationships unclear • Synergistic: mutually reinforcing norms or institutional labour divisions
Topology of regime complex at a given time	Scope of analysis	<ul style="list-style-type: none"> • Macro: e.g. interactions of AI regime complex with other regimes (trade; data privacy; transport); or with general international law. • Meso: e.g. interactions of AI security regime with other AI regimes • Micro: e.g. internal institutional dynamics in AI security regime complex
	General trends in regime complexity?	<ul style="list-style-type: none"> • Density; accretion; power shifts over time; preference changes; modernity; representation and voice goals; local governance
	Effects of AI governance disruption?	<ul style="list-style-type: none"> • Development: AI as generator or trigger of latent regime fault lines • Displacement: AI as shield, patch, cure or accelerator of fragmentation. • Destruction: AI as driver of governance contestation
Consequences of trajectories...	If regime complex remains fragmented	<ul style="list-style-type: none"> • Drawbacks: undercuts coherence of international law; operational dysfunction; barriers to access and power inequalities; strategic vulnerability to forum shopping • Benefits: problem-solving; more democratic, inclusive; greater trust
	If regime complex is integrated	<ul style="list-style-type: none"> • Drawbacks: slowness, brittleness, 'breadth vs. depth' dilemma • Benefits: greater political power, efficiency and participation, can avert forum shopping
Strategies for managing AI regimes to ensure...	Efficacy (sociotechnical change)	<ul style="list-style-type: none"> • Conceptual approach (x3), instrument choice (x3), instrument design (x1)
	Resilience (governance disruption)	<ul style="list-style-type: none"> • Conceptual approach (x4), instrument choice (x6), instrument design (x3)
	Coherence (regime complexity)	<ul style="list-style-type: none"> • Conceptual approach (x2), instrument choice (x2), instrument design (x2)

Table 7.1. Overview of analysis of AI Governance Regime

7.1 Origins: the Purpose, Viability and Design of an AI regime

In the first place, we must consider the origins or foundations of any anticipated or desired regime for AI issues. This involves an explicit analysis of three aspects of a proposed or envisioned AI regime: its *purpose* (7.1.1), its *viability* (7.1.2), and its *design* (7.1.3).

To work through these questions, we will, in these next sections, focus in on an example of the incipient security governance regime complex for military AI applications.¹ While perhaps not a flawless comparator to governance for other AI issues, this can nonetheless be a potentially valuable case. This is because public debates around military developments and uses of AI—particularly LAWS—have been very high on current agendas of AI governance. Moreover, arms control regimes for emerging technologies can more generally serve as high-stakes ‘hard case’ tests for global governance systems under the most high-stakes circumstances.

7.1.1 Regime purpose: Sociotechnical changes and governance rationales

In the first place, we should ask, from the perspective of sociotechnical change, what concerns does AI technology raise? What new behaviours, entities, or relations does a certain AI capabilities create? Will these be widely pursued and deployed? If so, what governance rationales would this create? And how should we think about governance, taking into consideration both the technology’s material features, as well as cross-application problem logics? We shall briefly review these in turn, in the context of developments in- and risks from military AI.

7.1.1.1 Technological developments and applications in military AI

As has been discussed, over the past decade, various AI technologies have begun to see introduction in diverse sectors. In these, there may be many industries where companies or users may honestly claim that the introduction of AI is an entirely new development in their line of work. The military is not amongst these. Indeed, as with many other technologies, militaries may be amongst the ‘oldest’ users of AI. Indeed, the use of computing and automation technologies in military operations has a history that dates back to the very dawn of these disciplines, and many early developments in computing technologies and AI found their direct genesis in military projects, or at least were intimately tied up with strategic needs.² In spite of occasional disillusionment during the ‘AI winters’, developments in AI have already influenced the character of conflict. For instance, during the first Gulf War, the US military’s use of the ‘DART’ tool for automated logistics planning and scheduling was supposedly so successful that DARPA claimed this single application had promptly paid back thirty years of investment in AI.³

¹ The selection of security regimes also follows the direct focus of *Papers [II] and [III]*. See also the discussion of the use of examples in Chapter 3.3.1.

² See for instance SHARON WEINBERGER, THE IMAGINEERS OF WAR: THE UNTOLD STORY OF DARPA, THE PENTAGON AGENCY THAT CHANGED THE WORLD (2017); ALEX ROLAND & PHILIP SHIMAN, STRATEGIC COMPUTING: DARPA AND THE QUEST FOR MACHINE INTELLIGENCE, 1983-1993 (2002).

³ Stephen E. Cross & Edward Walker, *DART: Applying Knowledge Based Planning and Scheduling to CRISIS Action Planning*, in INTELLIGENT SCHEDULING 711–29 (Monte Zweben & Mark Fox eds., 1994); Sara Reese Hedberg, *DART: Revolutionizing Logistics Planning*, 17 IEEE INTELL. SYST. 81–83 (2002).

Moreover, beyond use in logistics, the incorporation of a degree of automatic operation or ‘autonomy’ in lethal weapons systems is not historically unprecedented.⁴ In certain contexts, militaries have been operating ‘fire and forget’ weapons (such as ‘homing’ torpedoes) for over 70 years.⁵ In defensive roles, autonomous systems have for some time already been entrusted with operating fully autonomous weapons systems, such as the ‘Close-In Weapon Systems’ which are operated by dozens of countries to defend naval vessels from incoming missiles.⁶

Nonetheless, recent years have seen a speeding up in the militarisation of AI.⁷ A 2017 review identified “49 deployed weapon systems with autonomous targeting capabilities sufficient to engage targets without the involvement of a human operator”.⁸ Indeed, the first fully autonomous weapons systems have begun to see some limited deployments. The South Korean military, for instance, has for some years deployed Samsung SGR-A1 sentry gun turrets to the Korean Demilitarized Zone, which are reported to have a fully autonomous capability.⁹ In recent years, Israel has begun to deploy the ‘Harpy’ loitering anti-radar drone,¹⁰ and news reports have alleged that various countries have begun to develop and market weaponised drones capable of varying degrees of autonomy.¹¹ Unsurprisingly, this area also sees active research. For instance, in August 2020, DARPA ran AlphaDogFight, a simulated dogfight between a human F-16 pilot and a reinforcement-learning based AI system, which saw the AI defeating

⁴ For a distinction between ‘automatic’, ‘automated’, ‘autonomous’, and ‘intelligent’, see PAUL SCHARRÉ, *Autonomous Weapons and Operational Risk* 12 (2016), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf.

⁵ For instance, acoustic homing torpedoes already saw some use during the Second World War, such as the ‘Wren’ torpedo deployed by German submarines. Paul Scharre, *Autonomous Weapons and Stability*, March, 2020.

⁶ *Id.* at 302–303.

⁷ Justin Haner & Denise Garcia, *The Artificial Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development*, 10 GLOB. POLICY 331–337 (2019).

⁸ VINCENT BOULANIN & MAAIKE VERBRUGGEN, *Mapping the development of autonomy in weapon systems* 147 26 (2017), https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

⁹ Loz Blain, *South Korea’s autonomous robot gun turrets: deadly from kilometers away* (2010), <http://newatlas.com/korea-dodamm-super-aegis-autonomos-robot-gun-turret/17198/> (last visited Sep 20, 2016); Alexander Velez-Green, *The Foreign Policy Essay: The South Korean Sentry—A “Killer Robot” to Prevent War*, LAWFARE (2015), <https://www.lawfareblog.com/foreign-policy-essay-south-korean-sentry%E2%80%94killer-robot-prevent-war> (last visited Sep 11, 2020).

¹⁰ Israel Aerospace Industries, *Harpy Loitering Weapon*, <https://www.iai.co.il/p/harpy> (last visited Sep 11, 2020); Scharre, *supra* note 5 at 21.

¹¹ For instance, in 2019, Chinese company Ziyan began to market its ‘Blowfish A3’, a machine-gun carrying assault drone that was allegedly marketed as sporting ‘full autonomy’. Patrick Tucker, *SecDef: China Is Exporting Killer Robots to the Mideast*, DEFENSE ONE (2019), <https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/> (last visited Sep 11, 2020). 2020 has seen claims that Turkey developed (semi)autonomous versions of a ‘kamikaze’ drone. Joseph Trevithick, *Turkey Now Has Swarming Suicide Drones It Could Export*, THE DRIVE, 2020, <https://www.thedrive.com/the-war-zone/34204/turkey-now-has-a-swarming-quadcopter-suicide-drone-that-it-could-export> (last visited Jun 22, 2020).

the human pilot in all of five matches.¹² Moreover, several parties are making advances in domains such as swarming.¹³

The broader trend in the militarisation of robotic and AI technologies is projected to continue and perhaps accelerate further. Global military spending on autonomous weapons systems and AI is projected to reach \$16 and \$18 billion respectively, by 2025.¹⁴ A range of states, most notably the US, China, Russia, South Korea, and the EU, are now investing in exploring military applications of this technology.¹⁵

The appeal of AI technologies to military organisations should not be surprising, given the technology's potential breadth.¹⁶ Indeed, AI systems could see diverse applications across the battlefield—fulfilling diverse roles across what the US Army has called the 'Internet of Battle Things',¹⁷ or in pursuit of what the Chinese PLA has referred to as a 'Battlefield Singularity'.¹⁸ These applications might include systems deployed in a direct offensive and kinetic capability. However, beyond the 'killer robots' that feature so prominently in public debates, versions of 'lethal autonomy' can come in much more 'mundane' forms. For instance, Sweden has developed a 'smart' artillery shell that can compare detected vehicles in a target area, and only engage those contained on a programmable target list.¹⁹ The U.S. Air Force has been developing its 'Golden Horde' initiative, a project to create munition swarms which, once fired, work together to gather reconnaissance, carry out aerial refuelling, or attack targets matching certain criteria.²⁰

More generally, there is a broad diversity of military uses of AI, many of which extend far beyond individual 'killer robots'.²¹ AI systems could see application in coordinating units to fire

¹² Will Knight, *A Dogfight Renews Concerns About AI's Lethal Potential*, WIRED, 2020, <https://www.wired.com/story/dogfight-renews-concerns-ai-lethal-potential/> (last visited Aug 26, 2020). However, it should be noted that the matchup might not have been entirely even or representative, given that the AI had perfect information within the simulation, something it would not have in the real world.

¹³ For instance, in September 2020, defence company Anduril began tests on the 'Ghost 4', a swarm of which can be controlled by a single operator to perform reconnaissance. Will Knight, *Anduril's New Drone Offers to Inject More AI Into Warfare*, WIRED, 2020, <https://www.wired.com/story/anduril-new-drone-inject-ai-warfare/> (last visited Sep 11, 2020).

¹⁴ Haner and Garcia, *supra* note 7 at 331.

¹⁵ *Id.* at 332. However, other active parties include India, Israel and Japan. Justin K. Haner, *Dark Horses in the Lethal AI Arms Race* (2019), <https://justinkhaner.com/aiarmsrace> (last visited Sep 10, 2020).

¹⁶ See also the earlier discussion, in Chapter 2.1.7.2, of the broad usability of AI capabilities (and particularly the marginal added value of autonomy in various military contexts).

¹⁷ Alexander Kott, *Challenges and Characteristics of Intelligent Autonomy for Internet of Battle Things in Highly Adversarial Environments* (2018), <https://arxiv.org/ftp/arxiv/papers/1803/1803.11256.pdf>.

¹⁸ ELSA B. KANIA, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power* (2017), <https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235804> (last visited Mar 28, 2018).

¹⁹ John Cherry & Christopher Korpela, *Enhanced distinction: The need for a more focused autonomous weapons targeting discussion at the LAWS GGE*, HUMANITARIAN LAW & POLICY BLOG (2019), <https://blogs.icrc.org/law-and-policy/2019/03/28/enhanced-distinction-need-focused-autonomous-weapons-targeting/> (last visited Jul 14, 2020); As cited in Ashley Deeks, *Coding the Law of Armed Conflict: First Steps*, in THE LAW OF ARMED CONFLICT IN 2040, 18 (Matthew C. Waxman ed., 2020), <https://papers.ssrn.com/abstract=3612329> (last visited Jun 25, 2020).

²⁰ Rachel S. Cohen, *Air Force to Test Weapons Swarming Software in October*, AIR FORCE MAGAZINE (2020), <https://www.airforcemag.com/air-force-to-test-weapons-swarming-software-in-october/> (last visited Sep 22, 2020).

²¹ Cf. Hin-Yan Liu, Léonard Van Rompaey & Matthijs M. Maas, *Editorial: Beyond Killer Robots: Networked Artificial Intelligence Systems Disrupting the Battlefield?*, 10 J. INT. HUMANIT. LEG. STUD. 77–88 (2019); Léonard Van Rompaey, *Shifting from Autonomous Weapons to Military Networks*, 10 J. INT. HUMANIT. LEG. STUD. 111–128

on certain targets at the tactical level—creating what the US Army has called an ‘artificial intelligence enabled kill web’.²² AI could also see uses that are adversarial but non-kinetic, such as autonomous cyberwarfare systems or adaptive radar-jamming or electronic warfare capabilities.²³ Moreover, there are also many applications of AI in ‘non-offensive’ military roles,²⁴ including pervasive tactical surveillance,²⁵ command and control capabilities, improved (satellite) sensing capabilities, general logistics,²⁶ training,²⁷ or medevac or medical triage.²⁸

At a strategic level, there have even been concerns that aspects of automation might make their way into the command and control architectures for nuclear deterrents.²⁹ Indeed, while senior US officials maintain that there will always be a human in the loop for nuclear weapons,³⁰

(2019); Hin-Yan Liu, *From the Autonomy Framework towards Networks and Systems Approaches for ‘Autonomous’ Weapons Systems*, 10 J. INT. HUMANIT. LEG. STUD. 89–110 (2019).

²² Ashley Roque, *Project Convergence 2020: US Army hosting kill chain demonstration*, JANES.COM (2020), <https://www.janes.com/defence-news/news-detail/project-convergence-2020-us-army-hosting-kill-chain-demonstration> (last visited Sep 20, 2020). See also Theresa Hitchens, *Kill Chain In The Sky With Data: Army’s Project Convergence*, BREAKING DEFENSE (2020), <https://breakingdefense.com/2020/09/kill-chain-in-the-sky-with-data-armys-project-convergence/> (last visited Sep 20, 2020). Sydney J. Freedberg, *A Slew To A Kill: Project Convergence*, BREAKING DEFENSE (2020), <https://breakingdefense.com/2020/09/a-slew-to-a-kill-project-convergence/> (last visited Sep 20, 2020). In another test, the US Army used target identification and recommendation algorithms to shorten an artillery ‘kill chain’ from 10 minutes to 20 seconds. Sydney J. Freedberg, *Target Gone In 20 Seconds: Army Sensor-Shooter Test*, BREAKING DEFENSE (2020), <https://breakingdefense.com/2020/09/target-gone-in-20-seconds-army-sensor-shooter-test/> (last visited Sep 20, 2020).

²³ See for instance the DARPA ARC and BLADE programs; Paul Tilghman, *Adaptive Radar Countermeasures (ARC)*, DARPA, <https://www.darpa.mil/program/adaptive-radar-countermeasures> (last visited Mar 12, 2018); Paul Tilghman, *Behavioral Learning for Adaptive Electronic Warfare (BLADE)*, DARPA, <https://www.darpa.mil/program/behavioral-learning-for-adaptive-electronic-warfare> (last visited Mar 12, 2018).

²⁴ Although these might still be considered ‘lethality-enabling’; Arthur Holland Michel, *The Killer Algorithms Nobody’s Talking About*, FOREIGN POLICY (2020), <https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/> (last visited Jan 21, 2020).

²⁵ Compare project ‘MAVEN’; Deputy Secretary of Defense, *Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)* (2017), <https://www.scribd.com/document/346681336/Establishment-of-the-AWCFT-Project-Maven> (last visited Apr 30, 2017), or also the ‘Gorgon Stare’ program. ARTHUR HOLLAND MICHEL, EYES IN THE SKY: THE SECRET RISE OF GORGON STARE AND HOW IT WILL WATCH US ALL (2019), <https://www.hmhbooks.com/shop/books/Eyes-in-the-Sky/9780544972001> (last visited Jun 24, 2019); Sharon Weinberger, *Hollywood and hyper-surveillance: the incredible story of Gorgon Stare*, 570 NATURE 162–163 (2019).

²⁶ On various non-kinetic uses of military AI, see also STEPHAN DE SPIEGELEIRE, MATTHIJS M. MAAS & TIM SWEIJS, ARTIFICIAL INTELLIGENCE AND THE FUTURE OF DEFENSE: STRATEGIC IMPLICATIONS FOR SMALL- AND MEDIUM-SIZED FORCE PROVIDERS (2017), <http://hess.nl/report/artificial-intelligence-and-future-defense> (last visited May 19, 2017).

²⁷ Krystal K. Hachey, Tamir Libel & Zack Partington, *The Impact of Artificial Intelligence on the Military Profession*, in RETHINKING MILITARY PROFESSIONALISM FOR THE CHANGING ARMED FORCES 201–211 (Krystal K. Hachey, Tamir Libel, & Waylon H. Dean eds., 2020), https://doi.org/10.1007/978-3-030-45570-5_13 (last visited Jun 22, 2020).

²⁸ Kenneth H. Wong, *Framework for guiding artificial intelligence research in combat casualty care*, 10954 in MEDICAL IMAGING 2019: IMAGING INFORMATICS FOR HEALTHCARE, RESEARCH, AND APPLICATIONS 109540Q (2019), <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10954/109540Q/Framework-for-guiding-artificial-intelligence-research-in-combat-casualty-care/10.1117/12.2512686.short> (last visited Mar 22, 2019).

²⁹ Mark Fitzpatrick, *Artificial Intelligence and Nuclear Command and Control*, 61 SURVIVAL 81–92 (2019); Matt Field, *Strangelove redux: US experts propose having AI control nuclear weapons*, BULLETIN OF THE ATOMIC SCIENTISTS (2019), <https://thebulletin.org/2019/08/strangelove-redux-us-experts-propose-having-ai-control-nuclear-weapons/> (last visited Sep 2, 2019); James Johnson, *Delegating strategic decision-making to machines: Dr. Strangelove Redux?*, 0 J. STRATEG. STUD. 1–39 (2020).

³⁰ Michael T. Klare, “*Skynet*” Revisited: The Dangerous Allure of Nuclear Command Automation, 50 ARMS CONTROL TODAY WASH. 10–15, 12–13 (2020).

recent years have seen informal calls in the US defence establishment for the creation of an AI-supported nuclear ‘Dead Hand’.³¹ Such proposals may remain extreme, but nonetheless, recent years have seen modest steps towards integrating more automation in nuclear forces.³²

Finally, beyond the context of direct military conflict, AI systems could support a broad range of strategic functions that could shape the decision to go to war, and the conditions of conflict once initiated.³³ AI capabilities could be used to support intelligence operations,³⁴ accelerate fundamental and applied defence research R&D programs,³⁵ provide early warnings of conflict, or help “war game” rival state’s moves and decisions.³⁶ As such, it is clear recent years are seeing technological progress—and high expectations or concerns. However, it is not clear that will all necessarily translate into widespread (global) societal disruption.

³¹ Adam Lowther & Curtis McGiffin, *America Needs a “Dead Hand”*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/08/america-needs-a-dead-hand/> (last visited Sep 2, 2019). The reference here is to the ‘Perimitr’/‘Dead Hand’, a Soviet system developed in the final decades of the Cold War, which was configured to (semi)automatically launch the USSR’s nuclear arsenal if its sensors detected signs of a nuclear attack and it lost touch with the Kremlin. Nicholas Thompson, *Inside the Apocalyptic Soviet Doomsday Machine*, WIRED, 2009, <https://www.wired.com/2009/09/mf-deadhand/> (last visited Jun 17, 2020); DAVID HOFFMAN, *THE DEAD HAND: THE UNTOLD STORY OF THE COLD WAR ARMS RACE AND ITS DANGEROUS LEGACY* (1 edition ed. 2010).

³² Klare, *supra* note 30.

³³ GREG ALLEN & TANIEL CHAN, *Artificial Intelligence and National Security* 2 (2017), <http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf> (last visited Jul 19, 2017) (“[a]dvances in AI will affect national security by driving change in three areas: military superiority, information superiority, and economic superiority.”).

³⁴ ALEXANDER BABUTA, MARION OSWALD & ARDI JANJAVA, *Artificial Intelligence and UK National Security: Policy Considerations* 57 (2020), https://rusi.org/sites/default/files/ai_national_security_final_web_version.pdf; Anthony Vinci, *The Coming Revolution in Intelligence Affairs*, FOREIGN AFFAIRS, 2020, <https://www.foreignaffairs.com/articles/north-america/2020-08-31/coming-revolution-intelligence-affairs> (last visited Sep 1, 2020).

³⁵ For instance, when designing the T-7 Red Hawk trainer aircraft, Boeing and Saab were able to use a wide array of digital models to test numerous possible designs before creating a prototype, considerably speeding up the development process. Aaron Gregg & Paul Sonne, *Air Force seeks a radical shift in how jets, missiles and satellites are designed*, WASHINGTON POST, September 15, 2020, <https://www.washingtonpost.com/business/2020/09/15/air-force-digital-design/> (last visited Sep 16, 2020). Similar digital tools reportedly enabled the U.S. Air Force to design and fly a prototype sixth-generation fighter jet for its Next Generation Air Dominance (NGAD) project within less than a year, a radical shortening of the usual decade-long development timelines which have previously characterised the development of modern fighter aircraft such as the F-35. Valerie Insinna, *The US Air Force has built and flown a mysterious full-scale prototype of its future fighter jet*, DEFENSE NEWS (2020), <https://www.defensenews.com/breaking-news/2020/09/15/the-us-air-force-has-built-and-flown-a-mysterious-full-scale-prototype-of-its-future-fighter-jet/> (last visited Sep 18, 2020).

³⁶ Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NATL. SECUR. LAW POLICY 1–25 (2019); Benjamin Jensen, Scott Cuomo & Chris Whyte, *Wargaming with Athena: How to Make Militaries Smarter, Faster, and More Efficient with Artificial Intelligence*, WAR ON THE ROCKS, 2018, <https://warontherocks.com/2018/06/wargaming-with-athena-how-to-make-militaries-smarter-faster-and-more-efficient-with-artificial-intelligence/> (last visited Jun 3, 2020). See also the DARPA ‘Gamebreaker’ AI contest, which encourages participants to generate new AI war-game strategies in various strategy videogames. Kelsey Atherton, *DARPA Wants Wargame AI To Never Fight Fair*, BREAKING DEFENSE (2020), <https://breakingdefense.com/2020/08/darpa-wants-wargame-ai-to-never-fight-fair/> (last visited Aug 26, 2020).

7.1.1.2 Military AI, Sociotechnical Change, and Governance Rationales

How big of a change would the military use of AI be? There remains widespread uncertainty over exactly how disruptive AI will prove in military contexts. In the longer term, many anticipate far-reaching or even foundational changes. AI has been described by some as ‘revolutionary’,³⁷ on the argument that virtually any aspect of military operations might be improved, made faster, or more accurate through the integration of this technology.³⁸ Autonomy has been described as “the third revolution in warfare, after gunpowder and nuclear arms”.³⁹ This has led some to anticipate widespread and unconstrained proliferation of such technology, on the assumption that “[t]he applications of AI to warfare and espionage are likely to be as irresistible as aircraft.”⁴⁰ Kenneth Payne has suggested that, if the use of AI systems will shift and displace human judgment, they might have a larger impact on the psychological essence of strategic affairs than did even nuclear weapons in their time.⁴¹ Even those that take a more muted perspective on the technology still agree that AI systems can serve as a potent ‘evolving’ and ‘enabling’ technology that will have diverse impacts across a range of military fields.⁴²

That is not to say all of these changes would be imminent or will occur in the near-term.⁴³ Indeed, as discussed previously,⁴⁴ AI technology often does face operational, organisational, and cultural barriers to adoption and deployment, such that the reality of military AI may appear relatively mundane at least for the immediate future. For instance, militaries may face particular and unexpected hurdles in procuring such systems from private sector tech companies, because of mismatches in organisational processes, development approach, and system requirements.⁴⁵

³⁷ For instance, US accounts frequently describe it as a ‘revolution’ in military affairs. ALLEN AND CHAN, *supra* note 33 at 70. Similar framings can be seen in Chinese accounts of AI enabling a ‘revolution’ in warfare, and the shift from ‘informatized’ to ‘intelligentized’ warfare. Elsa Kania, *数字化 – 网络化 – 智能化: China’s Quest for an AI Revolution in Warfare*, THE STRATEGY BRIDGE (2017), <https://thestrategybridge.org/the-bridge/2017/6/8/-chinas-quest-for-an-ai-revolution-in-warfare> (last visited Jun 20, 2017).

³⁸ However, for a more nuanced study of the conditions under which autonomy adds value, see DEFENSE SCIENCE BOARD, *Defense Science Board Summer Study on Autonomy* (2016), <https://www.hsdl.org/?abstract&did=794641>. See also Section 2.1.7.2.

³⁹ Future of Life Institute, *Open Letter on Autonomous Weapons*, FUTURE OF LIFE INSTITUTE (2015), <https://futureoflife.org/open-letter-autonomous-weapons/> (last visited Nov 22, 2018).

⁴⁰ ALLEN AND CHAN, *supra* note 33 at 50.

⁴¹ Kenneth Payne, *Artificial Intelligence: A Revolution in Strategic Affairs?*, 60 SURVIVAL 7–32, 7, 30 (2018).

⁴² Amy J Nelson, *The Impact of Emerging Technologies on Arms Control Regimes* (2018), <http://www.isodarco.it/courses/andalo18/paper/iso18-AmyNelson.pdf>. In an interesting taxonomy, she distinguishes between *emerging technologies* (novel technologies currently under development or which have come into existence during the last 10–15 years, and which are therefore beset by unknowns); *evolving technologies* (technologies that are already employed for military purposes, but which are undergoing significant refinements or enhancements); *disruptive technologies* (technologies that are poised to alter prevalent approaches to war and security, or which can challenge the laws of war); *enabling technologies* (those capable of augmenting or otherwise contributing to the creation or enhancement of military capabilities); and *convergence* (where one emerging technology augments another). *Id.* at 24.

⁴³ As argued by Michael Horowitz, “[m]ost applications of AI to militaries are still in their infancy, and most applications of algorithms for militaries will be in areas such as logistics and training rather than close to or on the battlefield”. Michael C. Horowitz, *Do Emerging Military Technologies Matter for International Politics?*, 23 ANNU. REV. POLIT. SCI. 385–400, 396 (2020).

⁴⁴ See also the discussion in Chapter 2.1.4.1.

⁴⁵ Maaike Verbruggen, *The Role of Civilian Innovation in the Development of Lethal Autonomous Weapon Systems*, 10 GLOB. POLICY 338–342 (2019).

That may not stop such systems, but it can certainly slow them, and high-profile failures might temper some enthusiasm. In other cases, highly-anticipated applications may remain beyond current AI capabilities. For instance, it has been argued that current machine learning systems do not lend themselves well to integration in nuclear targeting, given the difficulty of collating sufficient (and sufficiently reliable) training datasets of imagery of nuclear targets (e.g. mobile launch vehicles).⁴⁶ Nonetheless, if even a small subset of these diverse capabilities would be achieved and deployed, their impact on practices of war could become considerable.

As emphasised earlier, the point is not to attempt to predict specific technological trends or individual use cases. The argument is rather about understanding when, under what conditions, and for what reasons, certain AI uses could enable or drive sociotechnical changes that would correspond to overarching *governance rationales*. To be sure, certainly not all envisaged military uses of AI would necessarily create problems that require governance. Indeed, certain applications, such as in optimizing medical evacuation logistics, could appear broadly beneficial.⁴⁷ In other cases, such as discussions over the impact of AI weapons on conventional deterrence or conflict stability, there still is less clarity over what these effects would actually be—that is, whether or not they would be stabilizing or destabilizing, or how they would affect relative and absolute military power generally.⁴⁸

Nonetheless, a number of military applications of AI may produce sociotechnical changes that would constitute plausible or even pressing rationales for governance. Considered in terms of general rationales for regulation, these can involve new harms to humans, or potential threats to rights. Considered from the tenets of international law, the new behaviours enabled by some of these systems may create distinct problems for international peace and stability, International Humanitarian Law, or human rights.

7.1.1.3 Military AI as Governance Target: Materiality and Problem Logics

Once one or more governance rationales have been established, the next step is to consider the texture of the AI issue as a governance *target*. As discussed, this involves a simultaneous consideration of relevant material factors, along with the problem logics that are engaged.

⁴⁶ See Rafael Loss & Joseph Johnson, *Will Artificial Intelligence Imperil Nuclear Deterrence?*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/> (last visited Dec 4, 2019). At the same time, it has been countered that even if AI is not useable for direct nuclear targeting, they could still disrupt nuclear stability through their integration in novel autonomous platforms or in (anti-submarine) sensing capabilities. Zachary Kallenborn, *AI Risks to Nuclear Deterrence Are Real*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/> (last visited Oct 16, 2019).

⁴⁷ In general, the controversy of robotic and autonomous technology's application to military tasks depends more on certain categories of tasks, and the specific level of autonomy that is being proposed. See HUGO KLIJN & MAAIKE OKANO-HEIJMANS, *Managing Robotic and Autonomous Systems (RAS)* 8 (2020), <https://www.clingendael.org/publication/managing-robotic-and-autonomous-systems-ras> (last visited Mar 17, 2020). On the other hand, it would be clear that even systems that do not raise *ethical challenges* (such as the use of AI in military medical diagnosis) might still pose new *security risks* through their susceptibility to adversarial input. See for instance Samuel G. Finlayson et al., *Adversarial attacks on medical machine learning*, 363 SCIENCE 1287–1289 (2019).

⁴⁸ For one early exploration of how AI developments could affect military organizations and powers, see Benjamin M. Jensen, Christopher Whyte & Scott Cuomo, *Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence*, 22 INT. STUD. REV. 526–550 (2020).

7.1.1.3.1 Material & artefactual features

In terms of *material factors*, what are the governance-relevant material or architectural features of the technology? It should be emphasised that in most cases these characteristics are not ‘essential features’, but they can involve temporary but relatively stable aspects which affect or foreground certain patterns of use over others;⁴⁹ or which shape public perceptions or first impressions of the technology.

In the case of military AI applications, what are such relevant characteristics? Watts and Crootof have previously theorised a set of technological features that affect how ‘regulation-tolerant’ or ‘regulation-resistant’ new weapons technologies can be.⁵⁰ Drawing these together, Crootof has produced a list of conditions under which a weapons ban is most likely to be successful.⁵¹ To provide an indicative overview, these are sketched and assessed in the context of various military AI applications (see Table 7.2).

Condition	AI?	Notes
<i>The weapon is [considered] ineffective,*</i>	-	(-) AI seen by number of parties as cornerstone of next military revolution. (+) few countries appear interested in ‘counter-value’ (e.g. ‘slaughterbots’), interest mostly for ‘counter-force’ (e.g. anti-ship; anti-aircraft) systems. (+) some military concerns over effectiveness or applicability of machine learning to all military contexts, given data scarcity, risks of adversarial spoofing.
<i>Other means exist for accomplishing a similar military objective;</i>	+	(+) Is potentially the case for applications where AI is ‘merely’ labour-saving (e.g. automated image processing); (+) Arguably the case for anti-personnel LAWS (‘slaughterbots’). (-) May not be the case for applications where AI provides key competitive edge or marginal advantage (e.g. aerial superiority; cyberwarfare).
<i>The weapon is not novel: it is easily analogized to other weapons, and its usages and effects are well understood*</i>	--	(-) Technology remains fairly novel (?) Analogies to anti-personnel landmines or blinding lasers, but may be limited (-) Has already led to conceptual ambiguity over various terms. (-) Usage, effects and limits of present AI capabilities not well understood
<i>The weapon or similar weapons have been previously regulated</i>	+/-	(-) AI generally considered a novel unprecedented technology—but... (+) previous attempts to regulate UAVs and/or militarily useful ‘software’... but (-) with limited success (cryptography; 2016 missile software under MTCR). (+) integrations of AI into existing weapons could ensure the resulting assemblage is more easily compared to ‘already-regulated’ weapons (+) comparison to WMD could set precedent for regulation
<i>The weapon is unlikely to cause social or military disruption*</i>	--	(-) Many uses of AI might cause large-scale military disruption (-) Kinetic uses of AI (‘Killer Robots’), especially by non-state actors in terrorism, might cause large-scale social disruption
<i>The weapon has not already been integrated into a state’s armed forces</i>	-	(+) Some advanced applications (e.g. nuclear NC2) not yet deployed (-) But: many other applications or in trial stage, significant investments.
<i>The weapon causes superfluous injury or suffering in relation to prevailing standards of medical care</i>	-	(-) LAWS-mounted weapons do not intrinsically cause more suffering; injuries are less visceral or distinctive than for blinding lasers or anti-personnel mines. (+) ‘Machine killing’ carries its own peculiar taboo; drone swarms especially so. (-) Many ‘lethality-enabling’ AI systems may allow non-kinetic interventions.

⁴⁹ For a discussion of the impact of material features on processes of new weapon design and procurement, see also Maaike Verbruggen, *In Defense of Technological Determinism* (2020).

⁵⁰ Sean Watts, *Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War*, 91 INT. LAW STUD. 83 (2015); Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO LAW REV. 1837–1915, 1884 (2015).

⁵¹ Rebecca Crootof, *Jurisprudential Space Junk: Treaties and New Technologies*, in RESOLVING CONFLICTS IN THE LAW 106–129, 115–116 (Chiara Giorgetti & Natalie Klein eds., 2019), <https://brill.com/view/book/edcoll/9789004316539/BP000015.xml> (last visited Mar 15, 2019).

<i>The weapon is inherently indiscriminate</i>	+	(+) Potentially true for some AI capabilities today, but... (-) many AI weapons might be(come) more ‘accurate’ over time (+) malicious use could involve perversely ‘discriminate’ targeting (e.g. by race, gender; political affiliation), which would be politically visceral. (+) implicit risk of hacking or spoofing may result in ‘indiscriminate’ firing
<i>The weapon is or is perceived to be sufficiently notorious to galvanize public concern and spur civil society activism*</i>	++	(+) Significant public campaign against ‘Killer Robots’ (+) Surveys show rising global opposition to LAWS (but conditional on framing) (+) Early deployment incidents might galvanize further acute concern (-) Public opprobrium may have narrow scope (‘Killer Robots’); with comparatively less outcry about broader ‘lethality-enabling’ use
<i>There is sufficient state commitment to enacting regulations*</i>	+/-	(+) 30 states support outright ban, many others see human control as critical (+) Some (though mixed) private sector support in ban or restrictions (-) But: currently not much interest amongst leading developer states.
<i>The scope of the ban is clear and narrowly tailored (... states understand precisely what technology and associated capabilities they are voluntarily relinquishing);*</i>	--	(-) ‘Meaningful Human Control’ is promising line, but often conceptually and operationally blurred (-) Definitional problems around AI, and uncertainties over exact scope, breadth, and reach of future capability development, suggest states cannot clearly (perceive they) understand what capabilities they are relinquishing.
<i>Violations can be identified*</i>	-	(-) In many cases may be hard to monitor compliance (in non-intrusive way); (-) In context of cyberwarfare AI, doubly so (attribution also hard), but... (+) AI capabilities could help in monitoring and detecting violations

Table 7.2. Military AI as governance target: considering material and political features⁵²

To be clear, these assessments are not conclusive. One caveat is that it is not particularly useful to list the prospects for a ban on ‘military AI’ as a general class, since AI’s diverse applications in diverse military contexts may have very different material, strategic and political features. As such, while at a glance the above might appear to paint a somewhat pessimistic picture regarding the viability of an outright or blanket ban on military AI, it is important to disaggregate these also.

To be sure, there appear to be few military AI use cases that would score very ‘well’ across all 12 criteria.⁵³ Nonetheless, this set of criteria does help in examining whether, or to what extent, certain distinctive military applications of AI—such as variably in LAWS; AI-enabled cyberwarfare systems; predictive war-gaming algorithms; or the use of AI in nuclear command and control systems—would score differently on these above criteria, and therefore might be more or less ‘regulation-tolerant’ governance *targets* from a purely material or artefactual point of view.

7.1.1.3.2 Problem logics

In addition to material features, we can consider military AI technologies as governance targets by examining the problem logics of the challenges they raise. While military uses of AI technology have fed high-profile concerns across diverse dimensions,⁵⁴ in terms of the Chapter 5

⁵² Criteria by Rebecca Crootof (*Id.* at 115–116.); assessments here my own. But see also the earlier evaluation in Crootof, *supra* note 50 at 1891–1893. Note: starred (*) criteria depend on state perceptions or expectations of AI technology, which might be shifted through shocks, incidents, or epistemic community action.

⁵³ One possible exception might be the (hypothetical) use of AI systems or robots for the purposes of torture or interrogation. On which, see Amanda McAllister, *Stranger than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture*, MINN. LAW REV. 47 (2018).

⁵⁴ For useful overviews, see also FORREST E MORGAN ET AL., *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World* 224 (2020), https://www.rand.org/content/dam/rand/pubs/research_reports/RR3100/RR3139-1/RAND_RR3139-1.pdf; Denise Garcia, *Lethal Artificial Intelligence and*

taxonomy of problem logics, most of these problems can be characterised by some combination of *ethical challenges, security threats, safety risks, and structural shifts*.

Ethical challenges include, for instance, concerns that AI weapons would challenge human dignity;⁵⁵ or that practically they create problems in terms of accountability and responsibility.⁵⁶ Others concerns have centred on the potential for these technologies to see gradual osmosis towards domestic theatres, resulting in the ‘militarisation’ of policing and the blurring of the critical distinction between ‘war’ and ‘peace’.⁵⁷ As suggested previously,⁵⁸ ethical challenges can be difficult to govern, because they typically face one of two barriers to governance. One is actor ignorance or relative apathy (to the threatened values or rights); a deeper barrier is underlying societal disagreement over the values, interests or rights at stake. For these issues, governance logics tend to foreground governance proposals such as bans, oversight and accountability mechanisms, or ‘machine ethics’ and behavioural limits installed into the systems in questions.⁵⁹

In other cases, however, the governance problem posed by military AI capabilities can be clustered as new *security threats*. As such, there are concerns around possible hacking or adversarial attacks. Governance logics tailored to such challenges have to reckon not just with the ‘malice’ of actors, but should also consider the broader ‘offense-defense balance’ of a given line of AI research or knowledge.⁶⁰ In general, the problem logic of security challenges foregrounds perpetrator-focused solutions (e.g. the prevention of access, the improving of detection and forensic capabilities to ensure attribution), as well as target-focused measures (such as reducing exposure of systems). Because militaries have a clear interest in averting spoofing attacks on their AI capabilities, this might suggest policies would not need to be set at the international level. Nonetheless, even here, states could gain from sharing best practices or security systems.⁶¹

In terms of *safety risks*, there is concern that the use of military AI capabilities in complex and confused theatres of war might be highly susceptible to new accidents, in the form of

Change: The Future of International Peace and Security, 20 INT. STUD. REV. 334–341 (2018); Michael C. Horowitz, *Artificial Intelligence, International Competition, and the Balance of Power*, TEXAS NATIONAL SECURITY REVIEW, 2018, <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/> (last visited May 17, 2018).

⁵⁵ Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOB. POLICY 370–375 (2019). Though for a critical examination, see Daniel Lim, *Killer Robots and Human Dignity* 6 (2019).

⁵⁶ Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 UNIV. PA. LAW REV. 1347–1402 (2016).

⁵⁷ Denise Garcia, *Future arms, technologies, and international law: Preventive security governance*, 1 EUR. J. INT. SECUR. 94–111 (2016); On this trend, see also more generally Rosa Brooks, *War Everywhere: Rights, National Security Law, and the Law of Armed Conflict in the Age of Terror*, 153 UNIV. PA. LAW REV. 675–761 (2004).

⁵⁸ See chapter 4.4.1.

⁵⁹ See also the survey of approaches in Esther Chavannes, Klaudia Klonowska & Tim Sweijns, *Governing autonomous weapon systems: Expanding the solution space, from scoping to applying*, HCSS SECUR. 39 (2020).

⁶⁰ Toby Shevlane & Allan Dafoe, *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*, in PROCEEDINGS OF THE 2020 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (AIES ’20) (2020), <http://arxiv.org/abs/2001.00463> (last visited Jan 9, 2020).

⁶¹ RICHARD DANZIG, *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority* 40 (2018), <https://www.cnas.org/publications/reports/technology-roulette>. For instance, in the area of nuclear weapon security, the US has previously shared—or ‘leaked’—the technologies around the ‘permissive action links’ locks on nuclear weapons to both allies and rivals (respectively) in order to promote security and stability. Joseph S. Nye, *Nuclear Learning and U.S.-Soviet Security Regimes*, 41 INT. ORGAN. 371–402 (1987); See generally PETER STEIN & PETER FEAVER, *ASSURING CONTROL OF NUCLEAR WEAPONS: THE EVOLUTION OF PERMISSIVE ACTION LINKS* (1989).

unexpected behaviour, inability to discriminate sufficiently amongst targets, or because of emergent accident risks.⁶² This leads to a range of concerns over how or why false positives, spoofing, or catastrophic interaction between systems could exacerbate international crises.⁶³ In terms of their problem logics, it is important to understand how safety challenges in various military AI applications arise from (or interact with) actor negligence, overtrust and automation bias at the *operator* level; the complexity and ‘tight coupling’ of AI systems at the *organisational* level; and the ‘many hands’ problem at the level of *design and production*. Accordingly, safety issues around military AI systems in diverse domains and roles foreground their own set of governance logics. These include (1) framework solutions that pursue ‘Meaningful Human Control’ not merely at the operator level, but throughout the system design and deployment lifecycle;⁶⁴ (2) the funding and global dissemination of broad-spectrum AI *safety engineering* techniques to ensure AI system reliability, corrigibility, or interpretability; (3) institutional and legal mechanisms that ensure appropriate liability and state responsibility;⁶⁵ (4) international agreements to provide for the (temporary or permanent) relinquishment of AI capabilities from contexts (e.g. nuclear command and control) were safety failures would be utterly unacceptable.

Finally, many anticipated military AI capabilities would give rise to impacts that might be approached as *structural shifts*. Examples of these can be found in the concerns that AI weapons might erode global stability,⁶⁶ undermine the ‘democratic peace’;⁶⁷ or might proliferate easily to terrorist groups.⁶⁸ Others have suggested AI systems could fundamentally change

⁶² SCHARRE, *supra* note 4; JOHN BORRIE, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies* (2016), <http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf> (last visited Mar 7, 2018); STEPHANIE CARVIN, *Normal Autonomous Accidents: What Happens When Killer Robots Fail?* (2017), <https://papers.ssrn.com/abstract=3161446> (last visited Dec 17, 2019). For an argument that many competitive AI systems could be vulnerable to such failure modes, see Matthijs M. Maas, *Regulating for “Normal AI Accidents”: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment*, in PROCEEDINGS OF THE 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 223–228 (2018), <https://doi.org/10.1145/3278721.3278766> (last visited Sep 8, 2020).

⁶³ Scharre, *supra* note 5; See also Nathan Leys, *Autonomous Weapon Systems, International Crises, and Anticipatory Self-Defense*, 45 YALE J. INT. LAW 377–411 (2020).

⁶⁴ See Chavannes, Klonowska, and Sweijs, *supra* note 59 at 17–19; Elke Schwarz, *The (im)possibility of meaningful human control for lethal autonomous weapon systems*, ICRC HUMANITARIAN LAW & POLICY BLOG (2018), <https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/> (last visited Jun 3, 2020). Chavannes et al. also refer to Australia’s suggested ‘system of control’ (submitted to the CCW GGE meeting in March 2019) as one example of such a more holistic lifecycle approach. Australia, *Australia’s System of Control and applications for Autonomous Weapon Systems* (2019), [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/39A4B669B8AC2111C12583C1005F73CF/\\$file/CCW_GGE.1_2019_WP.2_final.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/39A4B669B8AC2111C12583C1005F73CF/$file/CCW_GGE.1_2019_WP.2_final.pdf) (last visited Aug 27, 2020).

⁶⁵ One notable proposal here is Rebecca Crootof’s envisaged legal regime of ‘war torts’ to hold states accountable for the injurious wrongs produced by unpredictable safety failures in autonomous weapons. Crootof, *supra* note 56.

⁶⁶ DUSTIN A. LEWIS, GABRIELLA BLUM & NAZ K. MODIRZADEH, *War-Algorithm Accountability* (2016), <https://blogs.harvard.edu/pilac/files/2016/09/War-Algorithm-Accountability-Without-Appendices-August-2016.pdf> (last visited Sep 16, 2016); Garcia, *supra* note 54; Horowitz, *supra* note 54.

⁶⁷ Jürgen Altmann & Frank Sauer, *Autonomous Weapon Systems and Strategic Stability*, 59 SURVIVAL 117–142 (2017); Haner and Garcia, *supra* note 7 at 322. Though note that the same charge was raised against drones, but it is unclear if that has been the case. JAMES IGOE WALSH & MARCUS SCHULZKE, *The ethics of drone strikes: does reducing the cost of conflict encourage war?* (2015), <https://ssi.armywarcollege.edu/pdffiles/PUB1289.pdf>.

⁶⁸ Şerif Onur Bahçecik, *Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames*, 0 GLOB. POLICY (2019), <https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12671> (last visited Jun 17, 2019).

military decision-making processes,⁶⁹ lead to upsets in military and strategic balances of power,⁷⁰ might alter the dynamics of conventional deterrence,⁷¹ or could even erode nuclear deterrence stability.⁷² What is important to note here, is that regardless of the specific application domain—whether the concern is that LAWS could lower the threshold to war, that conflict prediction algorithms could facilitate or tempt greater brinkmanship, or that cyberwarfare systems might blur the relative balance of power (inviting miscalculation) or could be vulnerable to accidental escalation (exacerbating safety risks)—these challenges display common problem logics.

As such, in crafting governance solutions for structural challenges in the area of military AI, it is key to understand how these are intertwined with actors' perceived systemic incentives. It is also useful to chart how these structural changes underpin or exacerbate first-order challenges in terms of ethical challenges, security threats, or safety risks. Finally, in terms of governance logics, military AI capabilities that are entangled in structural shifts highlight the importance of tailored arms control measures, as well as confidence-building arrangements in the near term.

7.1.2 Regime viability: Regime theory and governance foundations

Having established that there are sociotechnical changes that require governance (the rationale), as well as having considered the governance target, the next question for any AI governance regime concerns the question: is governance *viable*?

There are three ways we can approach the question of regime viability or foundations. From an outside or historical view, we can (1) compare the envisaged AI regime against the *history* of past regimes on similar issue areas (7.1.2.1). From an inside view, we can derive insights from Regime Theory—specifically older scholarship focused on the underpinnings of specific regimes. These lenses highlights (2) the *interests* and functions which an AI regime could serve (7.1.2.2), or (3) the degree to which—and the ways by which—various actors could shift international *norms* in its support (7.1.2.3).

⁶⁹ Payne, *supra* note 41.

⁷⁰ Horowitz, *supra* note 54; ALLAN DAFOE, *AI Governance: A Research Agenda* 52 (2018), <https://www.fhi.ox.ac.uk/govaiagenda/>; Ben Garfinkel & Allan Dafoe, *How does the offense-defense balance scale?*, 42 J. STRATEG. STUD. 736–763 (2019); Altmann and Sauer, *supra* note 67; Michael C. Horowitz, *When speed kills: Lethal autonomous weapon systems, deterrence and stability*, 42 J. STRATEG. STUD. 764–788 (2019). Scharre, *supra* note 5.

⁷¹ YUNA WONG ET AL., DETERRENCE IN THE AGE OF THINKING MACHINES (2020), https://www.rand.org/pubs/research_reports/RR2797.html (last visited Feb 3, 2020).

⁷² Keir A. Lieber & Daryl G. Press, *The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence*, 41 INT. SECUR. 9–49 (2017); EDWARD GEIST & ANDREW J LOHN, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* 28 (2018), <https://www.rand.org/pubs/perspectives/PE296.html>; Michael C. Horowitz, Paul Scharre & Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence*, ARXIV191205291 CS (2019), <http://arxiv.org/abs/1912.05291> (last visited Dec 18, 2019); I: Euro-Atlantic Perspectives THE IMPACT OF ARTIFICIAL INTELLIGENCE ON STRATEGIC STABILITY AND NUCLEAR RISK, (Vincent Boulain ed., 2019), <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>; Shahar Avin & S. M. Amadae, *Autonomy and machine learning at the interface of nuclear weapons, computers and people*, in THE IMPACT OF ARTIFICIAL INTELLIGENCE ON STRATEGIC STABILITY AND NUCLEAR RISK (V. Boulain ed., 2019), <https://www.repository.cam.ac.uk/handle/1810/297703> (last visited Oct 16, 2019); Johnson, *supra* note 29. For a discussion with the interface with cyberweapons, see Pavel Sharikov, *Artificial intelligence, cyberattack, and nuclear weapons—A dangerous combination*, 74 BULL. AT. SCI. 368–373 (2018).

Again, international security governance for military AI might offer a productive example here, precisely because the growing concern over the alleged ‘AI arms race’ has in recent years spurred some pessimism regarding the prospects of international regulation for autonomous weapons.⁷³ Governance for military AI might therefore offer a hard case to illustrate these questions.

7.1.2.1 *Historical view: lessons from nuclear arms control*

Firstly, we may be able to draw comparative lessons from past governance successes. The rich and chequered history of arms control initiatives certainly provides ample examples of both failures and successes in international cooperation, and could therefore provide a valuable sum of case studies to draw on in exploring the topic of AI.⁷⁴

Indeed, arms control has been a key instrument in the international quest for the maintenance of peace and security for an extraordinarily long time,⁷⁵ with historical precedent ranging from ancient Hindu, Greek and Roman bans on barbed or poisoned weapons to the attempted medieval prohibition by the Lateran Council on the ‘unchristian’ crossbow and arbalest.⁷⁶ However, the golden age of arms control has arguably bloomed in the last two centuries, as science and technology sped up and delivered a procession of new weapons to the battlefield. The 1868 St. Petersburg Declaration on explosive projectiles banned the use of exploding bullets,⁷⁷ followed by the 1899 and 1907 Hague Peace Conferences which renounced various weapons and means of warfare.⁷⁸ Since then, there has been an extensive array of international treaties banning the development, production, stockpiling or use of a variety of weapons, including poison gas (1925), bacteriological or biological weapons (1972), chemical weapons (1993), blinding lasers (1995), certain types of conventional weapons (1980); certain types of anti-personnel land mines (1996), and cluster munitions (2008).⁷⁹

⁷³ E.g. Edward Moore Geist, *It's already too late to stop the AI arms race—We must manage it instead*, 72 BULL. AT. SCI. 318–321 (2016).

⁷⁴ Scharre, *supra* note 5 at 51. (“...over 40 examples of successful and unsuccessful past efforts at regulating weapons provide a rich reservoir of case studies to draw upon to understand the feasibility of regulating autonomous weapons.”). See also the comparison to nuclear arms control in Paper [I]: Matthijs M. Maas, *How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons*, 40 CONTEMP. SECUR. POLICY 285–311 (2019). For a compilation of treaties, see Amy F Woolf, Paul K Kerr & Mary Beth D Nikitin, *Arms Control and Nonproliferation: A Catalog of Treaties and Agreements* 73 (2018).

⁷⁵ Rosemary Rayfuse, *Public International Law and the Regulation of Emerging Technologies*, in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY (2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22> (last visited Jan 3, 2019); PAUL SCHARRE, ARMY OF NONE: AUTONOMOUS WEAPONS AND THE FUTURE OF WAR 13 (1 edition ed. 2018).

⁷⁶ A. ROBERTS & R. GUELFF, DOCUMENTS ON THE LAWS OF WAR 3 (3rd ed. 2000). Referred to in Rayfuse, *supra* note 75 at 502. Note, the ban on crossbows only applied to use against Christians, and it has been widely held that the prohibition was stillborn. Crootof, *supra* note 50 at 1904. For a history, see also RICHARD DEAN BURNS, THE EVOLUTION OF ARMS CONTROL: FROM ANTIQUITY TO THE NUCLEAR AGE (Reprint edition ed. 2013).

⁷⁷ ST PETERSBURG DECLARATION RENOUNCING THE USE, IN TIME OF WAR, OF EXPLOSIVE PROJECTILES UNDER 400 GRAMMES WEIGHT, LXVI UKPP (1869) 659 (1868).

⁷⁸ THE HAGUE CONVENTIONS AND DECLARATIONS OF 1899 AND 1907, ACCOMPANIED BY TABLES OF SIGNATURES, RATIFICATIONS AND ADHESIONS OF THE VARIOUS POWERS, AND TEXTS OF RESERVATIONS, (James Brown Scott ed., 1915), <http://archive.org/details/hagueconventions00inte0oft> (last visited Sep 12, 2020).

⁷⁹ Rayfuse, *supra* note 75 at 502 (citing sources and instruments).

There have also been horizontal and vertical limits on the proliferation, quantity, or capabilities of nuclear weapons,⁸⁰ anti-ballistic missile systems,⁸¹ weapons on Antarctica,⁸² and weapons of mass destruction in space.⁸³ In some cases, there have been examples of mutual restraint even in the absence of formal agreements, such as over neutron bombs, kinetic anti-satellite weapons, and some forms of bayonets.⁸⁴

That is not to say arms control has always been successful. There have been failures in arms control, where parties genuinely sought mutual restraint—such as early 20th century rules on the use of submarines, air-delivered weapons, or poison gas in warfare—but were unable to uphold these rules during wartime.⁸⁵ Moreover, some treaties have collapsed as more nations gained access to the technology; and even the most successful prohibitions, such as on chemical and biological weapons, have at times failed to rein in rogue regimes.⁸⁶

Nonetheless, for all its flaws, the historical track record of arms control is still quite remarkable. It demonstrates that high-stakes challenges posed by particular new technologies to global security can, under some circumstances, provide the impetus and anchor for surprisingly strong regime complexes.⁸⁷ As Drezner notes with regards to the nuclear non-proliferation regime:

“[t]his kind of regime complex does not resemble the zero-sum description that is often ascribed to security issues. It would appear that the destructive power of nuclear weapons triggered higher levels of international cooperation than would otherwise have been expected. Indeed, the regime complex for nuclear weapons is far more robust than the one for conventional weaponry.”⁸⁸

Could the same impetus be achieved for a governance regime centred on the responsible management or control of military AI? As noted by Payne, “the idea of arms control for AI remains

⁸⁰ E.g. TREATY ON THE NON-PROLIFERATION OF NUCLEAR WEAPONS, 729 UNTS 161 (1968). As well as bilaterally, between the US and the USSR; INTERIM AGREEMENT BETWEEN THE US AND THE USSR ON CERTAIN MEASURES WITH RESPECT TO THE LIMITATION OF STRATEGIC OFFENSIVE ARMS (SALT I INTERIM AGREEMENT), 944 UNTS 3 (1972).

⁸¹ UNITED STATES & USSR, *Treaty on the Limitation of Anti-Ballistic Missile Systems*, 944 UNTS 13 (1972).

⁸² ANTARCTIC TREATY, 402 UNTS 71 (1959).

⁸³ TREATY ON PRINCIPLES GOVERNING THE ACTIVITIES OF STATES IN THE EXPLORATION AND USE OF OUTER SPACE, INCLUDING THE MOON AND OTHER CELESTIAL BODIES, 610 UNTS 205 (1967).

⁸⁴ Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, ARXIV200407213 Cs, 67 (2020), <http://arxiv.org/abs/2004.07213> (last visited Apr 16, 2020).

⁸⁵ *Id.* at 67.

⁸⁶ *Id.* at 67–68.

⁸⁷ Some disagree. For instance, on the basis of three case studies (Iraq’s weapons programs after the Gulf War, great power competition in arms in the interwar period, and superpower military rivalry during the Cold War), Coe and Vaynmann have argued that a ‘transparency-security trade-off’ undercuts many arms control agreements, and has ensured that arms control agreements have been unable to chain these arms races, which together accounted for almost 40% of all global arming in the past 2 centuries. Andrew J. Coe & Jane Vaynman, *Why Arms Control Is So Rare*, 114 AM. POLIT. SCI. REV. 342–355 (2020).

⁸⁸ Daniel W Drezner, *Technological change and international relations*, 33 INT. RELAT. 286–303, 295 (2019). Note, however, that this robustness does not necessarily imply full coherence or regime integration. See John Simpson, *The nuclear non-proliferation regime: back to the future?*, DISARM. FORUM 12, 7 (2004). (arguing that “[a]lthough together they were termed a ‘regime’, in practice some of the elements were seen by many states to be in conflict with each other (e.g. the NPT and export controls). In practice, what had been created was a fragmented system of unilateral and multilateral international governance of nuclear energy, particularly given the IAEA’s much wider promotional role in this area. This system covered a broad range of nuclear-related activities, some of which appeared to have only a peripheral connection to the prevention of nuclear proliferation”).

in its infancy”,⁸⁹ although there have been considerable developments in recent years. Still, there has been uncertainty over the political prospects for an international ban on LAWS, either within or outside of the CCW framework.

Yet the nuclear experience might offer grounds for optimism here. After all, it should be remembered that some early Cold War commentators were pessimistic about the prospects for international regulation of nuclear weapons,⁹⁰ and anticipated a wildfire spread of this ‘ultimate weapon’. For instance, in 1963, U.S. Secretary of Defense Robert McNamara argued to President John F. Kennedy that, because of falling production costs, at least eight new nuclear powers could be expected to emerge within the next decade.⁹¹ As a result, Kennedy, in a public speech later that year, sketched a vision of “the possibility in the 1970s of...a world in which 15 or 20 or 25 nations may have these weapons.”⁹² Given that such early pessimism regarding nuclear governance appears to have been misplaced or at least exaggerated, it is worth considering whether we today are again too pessimistic regarding the prospects of a regime regulating or even banning certain military AI systems.

The question over the *prima facie* political viability of arms control for military AI systems was taken up in Paper [I].⁹³ This paper drew a parallel with the historical track record of governing and controlling nuclear weapons,⁹⁴ arguing that our experience in constraining the spread of the ‘ultimate deterrent’ suggests a surprisingly optimist answer. It implies that even for strategically critical military AI applications, arms races need not be inevitable, but might be

⁸⁹ Payne, *supra* note 41 at 19.

⁹⁰ To be sure, this was not universal. The very early days of the nuclear age saw an active ‘world federalist’ movement, and startlingly ambitious proposals for the far-reaching global control of nuclear weapons, although this did not see further proposals after the failure of the 1946 Baruch Plan. F. Bartel, *Surviving the Years of Grace: The Atomic Bomb and the Specter of World Government, 1945–1950*, 39 DIPL. HIST. 275–302 (2015); Joseph Preston Baratta, *Was the Baruch Plan a Proposal of World Government?*, 7 INT. HIST. REV. 592–621 (1985). Nonetheless, even in later years of the Cold War, arms controls advocates such as Hedley Bull and others of the 1960s ‘Cambridge Community’ would continue to champion a more expansive and ambitious view of strategic arms control, one that was aimed not merely at maintaining deterrence stability, but rather at extensive forms of cooperative security. See Nancy W. Gallagher, *Re-thinking the Unthinkable: Arms Control in the Twenty-First Century*, 22 NONPROLIFERATION REV. 469–498 (2015).

⁹¹ Robert McNamara, “The Diffusion of Nuclear Weapons with and without a Test Ban Agreement” (1963); As quoted in MOEED YUSUF, *Predicting Proliferation: The History of the Future of Nuclear Weapon* 88 15 (2009).

⁹² As quoted in Graham Allison, *Nuclear Disorder*, FOREIGN AFFAIRS, 2010, <https://www.foreignaffairs.com/articles/pakistan/2010-01-01/nuclear-disorder> (last visited Apr 2, 2018).

⁹³ Maas, *supra* note 74.

⁹⁴ Of course, as discussed previously (in Chapter 5.3.6.2), in the governance of technology analogies can be invaluable but also inevitably imperfect, foregrounding some features while backgrounding others. Cf. Rebecca Crootof, *Regulating New Weapons Technology*, in THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 1–25, 17–18 (Eric Talbot Jensen & Ronald T.P. Alcala eds., 2019).. How suitable is the comparison between nuclear weapons and AI? The comparison was a pillar of Paper [I], which argued that even if nuclear weapons are profoundly different from military AI from a scientific or artefactual standpoint, they share certain key strategic, operational and political characteristics that can make the analogy productive. Maas, *supra* note 74 at 288–290. Nonetheless, others have argued that the relevant comparison for the state of AI research today is not ‘nuclear weapons’ but rather ‘nuclear physics research in the early 1930’s’. Marta Halina & Joseph Martin, *Five Ways AI Is Not Like the Manhattan Project (and One Way It Is)*, 3 QUARKS DAILY (2019), <https://www.3quarksdaily.com/3quarksdaily/2019/08/five-ways-ai-is-not-like-the-manhattan-project-and-one-way-it-is.html> (last visited May 6, 2020). Nonetheless, even this analogy can produce lessons for the questions of scientific openness around AI. KATJA GRACE, *Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation* (2015), <https://intelligence.org/files/SzilardNuclearWeapons.pdf>. Either way, this is not to suggest that this is the sole or even best comparator.

managed, channelled or even stopped. This could be facilitated through direct engagement with domestic political coalitions on their interests, or indirectly, by shaping norms top-down through international regimes, or bottom-up, through epistemic communities.⁹⁵

However, while such historical precedent provides an encouraging ‘existence proof’ for international regulation, there are also more pessimistic lessons. In the first place, there is no guarantee that arms control efforts can succeed immediately or in time to contain at least an initial wave of proliferation, as illustrated by the failure of the early attempts, during the late 1940’s, at securing international control of nuclear technology.⁹⁶

Moreover, in a more recent historical perspective, it should be recognised that the broader landscape of arms control has been under sustained stress. Indeed, the past few decades have proven particularly tough on arms control, as key bodies have gone underfunded,⁹⁷ and as long-standing accords have frayed or been abandoned, including the ABM Treaty (in 2004) and the INF Treaty (in 2019).⁹⁸ Indeed there is a very real prospect that if the 2010 U.S.-Russia New Start treaty is not extended before 2021, this would mark the first time since 1972 that both powers’ nuclear arsenals stand entirely unconstrained by treaties.⁹⁹ Moreover, in early 2020, the Trump Administration indicated a willingness to abandon the Open Skies Treaty,¹⁰⁰ and to resume nuclear testing.¹⁰¹ At a more systemic level, new development and digital technologies have also begun to put additional pressure on old and new arms control regimes, by increasing the rate at

⁹⁵ Maas, *supra* note 74 at 292–295, 296–299.

⁹⁶ As noted above, these early years saw the Acheson-Lilienthal proposal and the 1946 Baruch Plan, which urged for far-reaching global control of these weapons, but which soon failed. The seminal work is JOSEPH PRESTON BARATTA, THE POLITICS OF WORLD FEDERATION: FROM WORLD FEDERALISM TO GLOBAL GOVERNANCE (2004).

⁹⁷ Remarkably, the Biological Weapons Convention has just four staffers, and operates on a shoestring budget of just \$1.4 million a year—less than the average budget of a McDonalds restaurant. TOBY ORD, THE PRECIPICE: EXISTENTIAL RISK AND THE FUTURE OF HUMANITY 57 (2020); See also the figures in Gregory C. Koblentz & Paul F. Walker, *Can Bill Gates rescue the Bioweapons Convention?*, BULLETIN OF THE ATOMIC SCIENTISTS (2017), <https://thebulletin.org/2017/04/can-bill-gates-rescue-the-bioweapons-convention/> (last visited Jan 30, 2020); and Hayley Peterson, *Here's what it costs to open a McDonald's restaurant*, BUSINESS INSIDER (2019), <https://www.businessinsider.com/what-it-costs-to-open-a-mcdonalds-2014-11> (last visited Jan 30, 2020).

⁹⁸ Wade Boese, *U.S. Withdraws From ABM Treaty; Global Response Muted*, ARMS CONTROL ASSOCIATION (2002), <https://www.armscontrol.org/act/2002-07/news/us-withdraws-abm-treaty-global-response-muted> (last visited Sep 11, 2020). Amy J. Nelson, *The death of the INF Treaty has lessons for arms control*, BULLETIN OF THE ATOMIC SCIENTISTS (2019), <https://thebulletin.org/2019/11/the-death-of-the-inf-treaty-has-lessons-for-arms-control/> (last visited Jun 3, 2020).

⁹⁹ Deep Cuts Commission, *Urgent Steps to Avoid a New Nuclear Arms Race and Dangerous Miscalculation -- Statement of the Deep Cuts Commission* (2018), https://www.armscontrol.org/sites/default/files/files/documents/DCC_1804018_FINAL.pdf (last visited Apr 26, 2018). Some have argued pessimistically that the stability of many arms control regimes is fickle, and generally more dependent on healthy international relations than the other way around Lionel P. Fatton, *The impotence of conventional arms control: why do international regimes fail when they are most needed?*, 37 CONTEMP. SECUR. POLICY 200–222 (2016).

¹⁰⁰ Bonnie Jenkins, *A farewell to the Open Skies Treaty, and an era of imaginative thinking*, BROOKINGS (2020), <https://www.brookings.edu/blog/order-from-chaos/2020/06/16/a-farewell-to-the-open-skies-treaty-and-an-era-of-imaginative-thinking/> (last visited Jun 22, 2020).

¹⁰¹ Greg Webb, *Trump Officials Consider Nuclear Testing*, ARMS CONTROL ASSOCIATION (2020), <https://www.armscontrol.org/act/2020-06/news/trump-officials-consider-nuclear-testing> (last visited Sep 11, 2020).

which new technologies are produced, the digitalisation of existing weapon technologies, platforms or systems, and the diffusion and ‘latency’ this supports.¹⁰²

As such, taking an ‘outside view’, the history of arms control might support both optimism (arms control for high-stakes destabilizing technologies can be more viable than may initially appear), but also pessimism (nuclear arms control successes were hard-won; and it has been under pressure in recent years). It should then be asked whether, how, or why we would expect governance regimes for military AI to buck these trends. That requires taking a more detailed examination.

7.1.2.2 *Interests: regime foundations of collaboration and coordination*

A second perspective on AI regime viability takes an ‘inside view’. This focuses on the common ‘functions’ which, from a rationalist perspective, could support the establishment of an AI governance regime on a given issue. Regime theory can provide insights here. Classically, theorists of international regimes distinguished between ‘dilemmas of common interests’, and ‘dilemmas of common aversion’.¹⁰³ These provide two distinct bases for international regimes aimed at ensuring, respectively, ‘collaboration’ and ‘coordination’.¹⁰⁴

To take the latter first: regimes to address a ‘*dilemma of common aversion*’ generally require (just) coordination to avoid certain outcomes. Such situations involve multiple possible equilibria—by which is meant that such regimes are created not so much to ensure that one specific outcome is guaranteed, and rather to ensure that particular sub-optimal outcomes are avoided. Examples are questions involving international standardisation, where actors may not care strongly *which* standard is adopted, so long as some uniform standard is implemented amongst them all. While coordination can be difficult to achieve when different actors disagree about the preferred initial equilibrium, once a standard is settled on and a regime established it enables expectations to converge, and becomes relatively ‘self-enforcing’, because actors that defect hurt only themselves.¹⁰⁵ Many debates around global standard-setting in AI can be understood as debates around coordination, in the face of dilemmas of common aversion. Of course, even in such cases, success is not guaranteed, because domestic political considerations can at times lead to contestation, as illustrated by the evolution of cyberspace governance.¹⁰⁶ Nonetheless, the prospects of such regimes can generally be quite good.

In contrast, when actors face a ‘*dilemma of common interests*’, international regimes are oriented towards securing collaboration to ensure a specific outcome. Accordingly, such regimes must specify strict patterns of behaviour, and ensure no one party is tempted to cheat. This often requires formalisation, and sufficient monitoring capabilities to reassure each actor of their

¹⁰² AMY J. NELSON, *Innovation Acceleration, Digitization, and the Arms Control Imperative* 23 (2019), <https://papers.ssrn.com/abstract=3382956> (last visited May 29, 2020) (“arms control is in a terrible state. Violated, discredited and abandoned, arms control is no longer the foreign policy aspiration or lofty global governance tool of days past.”).

¹⁰³ Arthur A. Stein, *Coordination and Collaboration: Regimes in an Anarchic World*, 36 INT. ORGAN. 299–324 (1982).

¹⁰⁴ *Id.* at 311–312.

¹⁰⁵ *Id.* at 314.

¹⁰⁶ Drezner, *supra* note 88 at 296–297 (discussing the evolution of cyberspace governance).

ability to spot others' cheating.¹⁰⁷ Arms control agreements are classic examples of formal institutionalised collaborations in a common interests situations, as they must be able to define 'cheating' quite explicitly, ensure that it is observable, and specify verification and monitoring procedures. Many potential 'adversarial' uses of AI systems, particularly in military or strategic contexts, can be understood as producing dilemmas of common interests.¹⁰⁸ However, the monitoring requirements appear to throw up a particular challenges here, given that monitoring of compliance might create a particularly steep 'transparency-security' trade-off for states.¹⁰⁹ This ensures there is a need for increasing the ability of parties to undertake unilateral monitoring, or alternately to increase their ability to make verifiable, trusted claims about their AI systems.¹¹⁰

This creates additional challenges for arms control for military AI. Because arms control agreements involve the pursuit of mutual restraint and control under a dilemma of common interest, rather than coordination around a dilemma of common aversion, they might be challenging subjects for AI regimes.¹¹¹ Historically there have been considerable difficulties involved in *negotiating* and reaching arms control agreements;¹¹² -in verifying compliance;¹¹³ in maintaining agreements in the face of renewed geopolitical tensions or mistrust;¹¹⁴ or in updating or modernizing agreements in the face of a constantly changing technological state-of the art.¹¹⁵ Accordingly, there are historically only a few cases of preventative arms bans—the exceptions being exploding bullets and blinding lasers.¹¹⁶ That is not to say that, once implemented, arms control regimes have not been successful,¹¹⁷ nor is it to suggest that regimes have not been able to adapt and evolve over time.¹¹⁸ Nonetheless, it remains a challenge.

Moreover, the challenge might be exacerbated this time around, because digital technologies, including AI, may pose a particularly 'hard case'. Indeed, arms control could suffer from a number of drawbacks when pressed into service against prospective military AI systems.

¹⁰⁷ Stein, *supra* note 103 at 312.

¹⁰⁸ See also the previous discussion, in Chapter 2.2.2, on how these challenges could be framed as 'mutual restraint' global public goods.

¹⁰⁹ Coe and Vaynman, *supra* note 87.

¹¹⁰ Brundage et al., *supra* note 84.

¹¹¹ Stein, *supra* note 103 at 313.

¹¹² Richard Dean Burns, *An Introduction to Arms Control and Disarmament*, 1 in ENCYCLOPEDIA OF ARMS CONTROL AND DISARMAMENT 1–12 (Richard Dean Burns ed., 1993). See also Vally Koubi, *International Tensions and Arms Control Agreements*, 37 AM. J. POLIT. SCI. 148–164 (1993).

¹¹³ Allan S. Krass, *Arms Control Treaty Verification*, 1 in ENCYCLOPEDIA OF ARMS CONTROL AND DISARMAMENT 297–316 (Richard Dean Burns ed., 1993).

¹¹⁴ Fatton, *supra* note 99.

¹¹⁵ Nelson, *supra* note 42; NELSON, *supra* note 102.

¹¹⁶ Garcia, *supra* note 57 at 102. However, elsewhere, Garcia suggests that at least twenty-two existing treaties have been set up under a 'preventative framework' to address the (further) militarization of emerging technologies. Garcia, *supra* note 54 at 335.

¹¹⁷ Williamson notes that "while the record of compliance with arms control treaties is far from perfect, it is statistically quite good". Richard Williamson, *Hard Law, Soft Law, and Non-Law in Multilateral Arms Control: Some Compliance Hypotheses*, 4 CHIC. J. INT. LAW, 61 (2003), <https://chicagounbound.uchicago.edu/cjil/vol4/iss1/7>. Notable historical violations include the Soviet Union's production of biological agents in violation of the Biological Weapons Convention; Iraq's violation of the 1925 Geneva Protocol through the use of chemical weapons during the Iran-Iraq war, and Iraq and North Korea's violations of the Nuclear Non-Proliferation Treaty. *Id.* at 60.

¹¹⁸ Caroline Fehl, *Unequal power and the institutional design of global governance: the case of arms control*, 40 REV. INT. STUD. 505–531 (2014).

For instance, Nelson notes that arms control is predominantly designed to manage actual physical weapons; that historically many arms control agreements implicitly ended up bargaining away obsolete last-generation weapons, not next-generation ones, and that many measures rely on ‘what can be seen and counted’.¹¹⁹ More generally, the traditional lens of dual-use technologies may not result in actionable insights, as it is hard to separate the actors into discrete sets of state and non-state actors; and notions of ‘parity’ or ‘balance’ are hard to define. Because of this, Joseph Nye has argued that arms control risks being a traditional response to new technologies.¹²⁰

On the other hand, from a rationalist perspective, the possible role which (AI) technologies can serve in strengthening monitoring capabilities (as a form of *displacement*) can also shed light on the conditions for arms control regimes for military AI systems themselves, and interventions by which we might reduce or mitigate the above-mentioned security-transparency trade-off. That is, for a given agreement between state A and state B, to cap B’s military AI capabilities, how steep might the ‘security-transparency’ trade-off be?

On the one hand, one could argue that the inspection of AI (software) capabilities need not be intrusive on state B’s general security since, unlike missiles, military AI software need not necessarily be ‘co-located’ with other military assets, but could be housed in offsite data centres. Allowing inspections of the software therefore would not necessitate giving access to factories or bases producing or housing many other types of sensitive weapons technologies.

On the other hand, in order to guarantee compliance and assuage state A’s fears of cheating, arms control agreements for military AI might require radical levels of transparency of the AI system itself—although this may depend on whether agreements are struck at the level of AI system techniques (e.g. ‘no reinforcement learning’), capabilities (e.g. ‘no autonomous swarming’), application (e.g. ‘no integration in nuclear command and control’), or their performance or range in the field (e.g. ‘no usage in urban theatres’). Such a high level of transparency could considerably impinge on the security of party B: it might considerably degrade the utility of its military AI systems in the first place—since state A is better able to tailor countermeasures (such as adversarial input). This would suggest that in the context of an arms control agreement for AI systems, party A would need a very high level of transparency for verification—a level that controlled party B would consider unacceptable from a security perspective. This highlights the importance of interventions that enable various parties to make verifiable claims about their AI systems’ properties, without necessarily providing full access.¹²¹

In sum, under a rationalist account of regime theory, arms control for destabilizing or particularly unethical weapons could constitute an example of actors facing a ‘dilemma of common interests’.¹²² From this interests-based perspective, there appear to be some functional grounds

¹¹⁹ Nelson, *supra* note 42; Erica D. Borghard & Shawn W. Lonergan, *Why Are There No Cyber Arms Control Agreements?*, COUNCIL ON FOREIGN RELATIONS (2018), <https://www.cfr.org/blog/why-are-there-no-cyber-arms-control-agreements> (last visited Jan 22, 2018). See also Maas, *supra* note 74 at 289–290.

¹²⁰ Joseph S. Nye, *The World Needs an Arms-control Treaty for Cybersecurity*, THE WASHINGTON POST, October 1, 2015, <https://www.belfercenter.org/publication/world-needs-arms-control-treaty-cybersecurity> (last visited May 31, 2020).

¹²¹ On which, see Brundage et al., *supra* note 84 at 67–69; See also more generally GIACOMO PERSI PAOLI ET AL., *Modernizing Arms Control: Exploring responses to the use of AI in military decision-making* 52 (2020), <https://unidir.org/publication/modernizing-arms-control>; Maas, *supra* note 74.

¹²² Stein, *supra* note 103.

for establishing an institutional regime to restrict at least some types of military AI systems, and secure this ‘mutual restraint’ global public good.¹²³ At the same time, it appears that a number of operational and material features of the technology challenge the strong monitoring and enforcement requirements of such regimes.¹²⁴ This suggests that the foundations of military AI governance regimes could be improved in at least two ways: by improving the abilities to (unilaterally) monitor and enforce such regimes; or by shifting state actors’ *norms* and perceptions.

7.1.2.3 Norms: the role of epistemic communities

Along with existing state interests, constructivists have argued that regimes are grounded in norms and perceptions, which can be shifted or altered in various ways.¹²⁵ In *Paper [I]*, I argued that, far from being narrowly bound to security concerns, the history of nuclear non-proliferation suggest state decision-making can be swayed indirectly, by shaping decision maker norms—and with them the (perception of) interests. Such shifts in norms might be achieved top-down, through international regimes and legal principles. Alternatively, they can be facilitated bottom-up, through the activities of so-called ‘epistemic communities’.¹²⁶

In fact, epistemic communities have played a key role in laying the groundwork for various historical successes in arms control. For instance, Emanuel Adler has provided a study of the road to the 1972 ABM Treaty, which suggests that small communities of experts, appropriately organised and mobilised, can have a disproportionate effect in framing global arms control cooperation through bottom-up norm institutionalisation.¹²⁷ While there are also risks to the direct participation of technical experts in arms control negotiations,¹²⁸ they can play a significant

¹²³ SCOTT BARRETT, WHY COOPERATE?: THE INCENTIVE TO SUPPLY GLOBAL PUBLIC GOODS (2007), <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199211890.001.0001/acprof-9780199211890> (last visited May 24, 2018). See also the discussion in section 2.2.2.

¹²⁴ This question of design and viability will be taken up in more detail in the next section (7.1.3).

¹²⁵ Caroline Fehl, *Explaining the International Criminal Court: A Practice Test’ for Rationalist and Constructivist Approaches*; 10 EUR. J. INT. RELAT. 357–394 (2004).

¹²⁶ Adler defines an epistemic community as “... a network of individuals or groups with an authoritative claim to policy-relevant knowledge within their domain of expertise. The community members share knowledge about the causation of social and physical phenomena in an area for which they have a reputation for competence, and they have a common set of normative beliefs about what will benefit human welfare in such a domain. While members are often from a number of different professions and disciplines, they adhere to the following: (1) shared consummatory values and principled beliefs; (2) shared causal beliefs or professional judgment; (3) common notions of validity based on intersubjective, internally defined criteria for validating knowledge; (4) a common policy project.” Emanuel Adler, *The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control*, 46 INT. ORGAN. 101–145, 1 (1992). For a classic account, see also Peter M. Haas, *Introduction: Epistemic Communities and International Policy Coordination*, 46 INT. ORGAN. 1–35 (1992).

¹²⁷ Maas, *supra* note 74 at 298–299.

¹²⁸ Expert participation in the arms control negotiations may, however, also hold up negotiations on occasion. For instance, Burns recounts how during the early negotiations for the Comprehensive Test Ban Treaty (CTBT), seismic experts were brought on board to develop a verification system that could reliably detect the signs of nuclear tests. However, “[a]fter verification techniques were developed that appeared to be acceptable to many scientists and diplomats, technical experts kept searching for more and more refinements so that the already low error rate could be even smaller. One result of this was that it was impossible to negotiate a comprehensive test ban because critics—who often had a vested interest in seeing underground testing continue—would argue that

role in establishing shared expectations amongst policymakers, setting the political conditions for arms control negotiation, and providing the overarching framework.

Of course, not all epistemic communities are successful, and a few high-profile examples of success provide no guarantee that contemporary campaigns to restrict or ban some military uses of AI will be able to repeat that feat.¹²⁹ Certainly, it is possible that the external historical context has shifted too much. Nonetheless, there are still indicative lessons for the ways in which epistemic communities focused on the governance of military AI might best organize in order to articulate and converge around a shared policy project, or what strategies they could adopt in order to facilitate the key steps of intellectual innovation, domestic norm institutionalisation, and global norm dissemination, to lay the ground for later cooperation.¹³⁰

We can draw on these lessons to evaluate contemporary efforts by various epistemic communities or norm entrepreneurs engaged on establishing governance regimes for military AI. Significantly, the Campaign to Stop Killer Robots has picked up remarkable speed, and the ‘ethical AI community’ has developed considerably within just the past few years,¹³¹ and may play an increased role in shaping global and state perceptions on the issue. Other scholars have highlighted the potential of ‘trusted communities’ in moving forward regulatory frameworks for transnational governance approaches to military AI.¹³² However, to shape norms, these communities may need to make careful decisions around (1) which *framing* of AI technologies to adopt;¹³³ and (2) through which *fora* or avenues to pursue global regulation in the near-term. It may be the case that smaller, informal and like-minded groups potentially are a swifter avenue, even if they fragment the regime complex.¹³⁴

In sum, applying concepts from regime theory helps explore whether, or under what conditions, proposed regimes for certain AI governance targets might be more or less viable. In the case of security governance regimes for AI, an outside-view historical comparison with nuclear arms control suggests that arms control for a high-stakes emerging technology might be achievable. At the same time, regime theory helps identify what specific interests (from a rationalist perspective) or norm shifts (from a constructivist perspective) might be involved in founding such a regime this time around.

no one could not be absolutely certain that no cheating was going on.” Burns, *supra* note 112 at 7. This illustrates the sensitive considerations that such norm entrepreneurs need to confront and navigate.

¹²⁹ Elvira Rosert & Frank Sauer, *How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies*, 0 CONTEMP. SECUR. POLICY 1–26 (2020). (arguing that, at the very least, the current Campaign to Stop Killer Robots may have adopted a suboptimal strategy in trying to emulate the bans on blinding laser weapons and anti-personnel landmines, and that it should accordingly adjust its institutional choices, substance, and regulatory design).

¹³⁰ Maas, *supra* note 74 at 299.

¹³¹ Haydn Belfield, *Activism by the AI Community: Analysing Recent Achievements and Future Prospects*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 15–21 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375814> (last visited Feb 12, 2020).

¹³² KLIJN AND OKANO-HEIJMANS, *supra* note 47 at 19–24.

¹³³ Rosert and Sauer, *supra* note 129.

¹³⁴ Jean-Frédéric Morin et al., *How Informality Can Address Emerging Issues: Making the Most of the G7*, 10 GLOB. POLICY 267–273, 5 (2019) (“[the area of autonomous weapons] – despite its uncertainties – is one in which coordination and integration of governance should be within the grasp of a like-minded group such as the G7, since there is both scientific certainty and consensus regarding the impending dangers”).

7.1.3 Regime design: Governance Disruption and Regime Brittleness

Nonetheless, even if viable, this does not mean all types of regimes might be equally optimal or sufficient in achieving the envisaged interests or norms.

Conventionally, four possible strategies are on the table in the face of new military technologies. The first, unilateral option is for states to forego cooperation and rely on *deterrence*;¹³⁵ the second is gradual (international) *norm development* under the existing laws of war.¹³⁶ Next are two varieties of hard law: a prohibitory treaty (outright ban), or a regulatory treaty that restricts the capabilities of military AI applications (such as the levels of autonomy or the allowed contexts or domains of use).¹³⁷

Of these four approaches, two may appear broadly insufficient to most sets of goals. Relying on deterrence for regulating military AI development and deployment seems insufficient at best, and destabilizing at worst. Likewise, waiting on the gradual development of international norms may be insufficient. This is because in the current normative vacuum, state practice may instil a tacit acceptance of what is considered ‘appropriate’ from a narrowly military point of view.¹³⁸ For instance, as Bode and Huells have suggested, the use of autonomous weapons systems “can have fundamental normative consequences by setting novel standards of appropriate action in international security policy.”¹³⁹

This leaves two varieties of security governance: a full ban or a regulatory ban.

7.1.3.1 A categorical ban

The question of which *type* of policies could be best pursued by a military AI regime turns on the relative efficacy and viability of, on the one hand, a governance regime that aims to secure an outright ban on LAWS (or even military AI broadly), versus a more limited regime that merely aims to regulate certain uses (that is, use cases).¹⁴⁰ There is something to be said for either strategy. However, *prima facie*, we might expect an outright ban to hold up better (provided it can be achieved in the first place).

¹³⁵ Cf. Erin D. Dumbacher, *Limiting cyberwarfare: applying arms-control models to an emerging technology*, 25 NONPROLIFERATION REV. 203–222 (2018) (contrasting deterrence with norm development or a multilateral treaty, in the context of cyberwarfare).

¹³⁶ Kenneth Anderson & Matthew C. Waxman, *Law and ethics for autonomous weapon systems: Why a ban won’t work and how the laws of war can*, LAW ETHICS AUTON. WEAPON SYST. (2013).

¹³⁷ See also Myriam Dunn Cavelty, Sophie-Charlotte Fischer & Thierry Balzacq, “*Killer Robots*” and Preventative Arms Control, in ROUTLEDGE HANDBOOK OF SECURITY STUDIES 15 (2016); Wendell Wallach & Colin Allen, *Framing robot arms control*, 15 ETHICS INF. TECHNOL. DORDR. 125–135 (2013).

¹³⁸ Ingvild Bode & Hendrik Huelss, *Autonomous Weapons Systems and Changing Norms in International Relations*, 44 REV. INT. STUD. 393–413 (2018); Eugenio V. Garcia, *The militarization of artificial intelligence: a wake-up call for the Global South*, 7 (2019), https://www.researchgate.net/publication/335787908_The_militarization_of_artificial_intelligence_a_wake-up_call_for_the_Global_South/references (last visited Mar 20, 2020).

¹³⁹ Bode and Huelss, *supra* note 138.

¹⁴⁰ See also the proposed ‘middle road’ in: Ronald Arkin et al., *A Path Towards Reasonable Autonomous Weapons Regulation*, IEEE SPECTRUM: TECHNOLOGY, ENGINEERING, AND SCIENCE NEWS (2019), <https://spectrum.ieee.org/automation robotics/artificial-intelligence/a-path-towards-reasonable-autonomous-weapons-regulation> (last visited Oct 22, 2019).

A categorical ban derives its functionality from several reasons. In the first place, there are operational reasons why categorical bans may be more effective at facilitating effective coordination, than would a sprawling and complex set of rules around usage. Paul Scharre records how early attempts, during the First World War, to articulate rules of engagement around gas attack ('no offensive use'; 'no delivery by shell') proved less resilient and easily broken.¹⁴¹ This underpins Schelling's old observation that "the most powerful limitations, the most appealing ones, the ones most likely to be observable in wartime, are those that have a conspicuousness and simplicity, that are qualitative and not a matter of degree, that provide recognizable boundaries."¹⁴² Categorical norms work well, in part because they provide a clearer 'focal point' for coordination. This means that it is much easier for third parties (whether observers or other states) to coordinate on when and how to sanction breaches of categorical norms ('no bioweapon use'), than on breaches of continuous norms (e.g. proportionality; 'minimize civilian suffering'), since "small perceptual errors impede coordination on the basis of continuous variables".¹⁴³ These provide compelling reasons why rules or norms—in any context, not just around weapons—are more robust when they are categorical, than if they are continuous or gradual.

In the second place, and more subtly, a categorical ban could shield a regime from future governance disruption, by curbing the driver of such disruptive capability shifts at the very source. As Rebecca Crootof has noted;

"Somewhat counterintuitively, the most extreme form of multilateral treaty regulation of a technology—a complete ban on its creation or use—may be most likely to stand the test of time. This is largely due to the fact that the aim of a ban is not to accommodate shifting state party needs or new technological developments. Rather, it draws a line in the sand and marks a certain technology permanently off limits."¹⁴⁴

Technologies cannot be disruptive, after all, if they are never pursued or further developed.¹⁴⁵ Export control regime 'control lists' cannot fall out of date if no new instances or iterations of a technology are being produced. As such, a categorical ban prunes certain (ostensibly harmful) pathways of technological development at their early stage.¹⁴⁶ This reduces not only the

¹⁴¹ SCHARRE, *supra* note 75 at 341–342.

¹⁴² THOMAS C. SCHELLING, ARMS AND INFLUENCE 164 (1966). As quoted in SCHARRE, *supra* note 75 at 341.

¹⁴³ Moshe Hoffman et al., *Why Norms Are Categorical*, PNAS (2018), https://www.tse-fr.eu/sites/default/files/TSE/documents/sem2018/eco_pub/hoffman.pdf (last visited Jan 23, 2020). See also Scharre, *supra* note 5 at 267. (noting that "[b]ans that attempt to regulate how indiscriminate weapons are used on the battlefield in order to avoid civilian casualties (aerial bombardment, submarine warfare, incendiary weapons, and the Convention on Certain Conventional Weapons (CCW) land mine protocol) have had a poor track record of success. Complete bans on weapons (exploding bullets, expanding bullets, chemical weapons, biological weapons, environmental modification weapons, blinding lasers) have fared better").

¹⁴⁴ Crootof, *supra* note 51 at 114.

¹⁴⁵ This may not be entirely true. The perception or anticipation of a certain technology being possible—and fears that rival parties might pursue it—might still affect a regime indirectly, even if the technology is never realized (or even impossible).

¹⁴⁶ An alternate strategy could be to attempt to slow down or speed up different aspects of technological development in order to affect the *order* of arrival, in the hope that defensive or stabilizing uses of that technology (e.g. vaccines) can be developed and implemented before offensive or destabilizing ones (e.g. bioweapons). Nick Bostrom has called this strategy 'differential technological development'. See NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 229–230 (2014); See also Nick Beckstead, *Differential technological development*:

prospect for the emergence of future problematic capabilities, but also the prospects for legal disruption and the consequent need for legal development or revision.¹⁴⁷

On the other hand, such bans require that they can ‘hold the door’ entirely, as they might still be vulnerable to disruption sparked by un- or less-regulated technological developments in adjacent fields: one might compare the effect of advances in ballistic missiles, or accuracy-enhancing warhead ‘superfuses’¹⁴⁸ on bilateral nuclear arms control agreements dependent on limiting quantities of deployed warheads only. Given the extremely flexible nature of AI techniques and their functions, and the extreme diversity of AI capabilities, developments in the civilian sector would eventually spill back over,¹⁴⁹ driving legal disruption in treaties focused narrowly on military AI. It is unclear if an outright ban could prevent such legal disruption short of attempting to effect an extremely (and undesirably) general restriction on AI research. Nonetheless, even if it were more vulnerable to future disruption, one might argue that a relatively narrowly targeted ban on certain military AI applications might be a valuable goal or function for an AI security regime.

One of the biggest challenges, indeed, relates simply to the difficulty of drawing meaningful and sufficiently categorical ‘red lines’ around military AI (or any digital or virtual technologies). An outright ban on ‘AI’ used in military contexts risks being over-inclusive and misleading; AI is not a single technology, so much as a suite of computational techniques, integrated in existing platforms or operations, aimed at specific decisional tasks. Trying to halt the horizontal proliferation of military AI systems through a categorical ban on any military use of AI down to the simplest machine learning algorithms does not seem viable (and possibly is not desirable). This is unfortunate, given that historical weapons bans have not been as effective if they were partial.¹⁵⁰ That raises the question of how ‘partial’ or targeted regulatory treaties could be improved or made more resilient instead.

7.1.3.2 Partial bans or other stop-gap policies

How then might a more partial ban be better tailored? Articulating suitable categorical bans on the basis of the technological *approaches* and ‘*input*’—the use of certain data; or of certain approaches such as supervised learning—likewise appears to be difficult, given that specific AI approaches (e.g. reinforcement learning) do not map neatly to applications (and at any rate have considerable desirable civilian uses). Alternately, articulating bans on the basis of AI *capabilities* could be more meaningful. This might then enable one to rule out, for instance, any ‘cyberwarfare systems with the ability to autonomously hack back’; or ‘systems that are able to interface with nuclear assets’.

Some early thinking, THE GIVEWELL BLOG (2015), <http://blog.givewell.org/2015/09/30/differential-technological-development-some-early-thinking/> (last visited May 28, 2017).

¹⁴⁷ On the relation between legal disruption and development, see Chapter 5.3.

¹⁴⁸ See also Matthijs M. Maas, *Innovation-Proof Governance for Military AI? How I learned to stop worrying and love the bot*, 10 J. INT. HUMANIT. LEG. STUD. 129–157, 137 (2019). See also the discussion in Section 5.3.3.

¹⁴⁹ Not that this would be an easy or fast process, given various organizational and institutional brakes on the ‘spin-in’ of civilian AI research to military usage. Verbruggen, *supra* note 45.

¹⁵⁰ Scharre, *supra* note 5 at 267.

However, as discussed previously, regimes aimed at restricting *vertical* proliferation have, in the past, relied on setting clearly quantified and comparable caps or limits. Yet, as with cyber weapons, AI systems may be harder to measure or compare,¹⁵¹ especially if there is uncertainty over the relative military efficacy. On the one hand, that would speak in favour of attempting to forecast potential impacts.¹⁵² On the other hand, more accurate forecasting could perversely also undercut the willingness of certain parties to enter into negotiations, if the outcome strongly indicates asymmetric relative stakes.¹⁵³

As such, even if partial bans proved initially politically viable, we should perhaps be more cautious regarding the lasting efficacy of a global, integrated weapons ban. As previously discussed,¹⁵⁴ AI systems might simply prove drivers of governance ‘erosion’, because they might prove hard to be captured under existing international legal solutions for conceptual or political reasons. Arms treaties tend to focus on static weapons, and may need constant updating. They are therefore vulnerable to rapid technological change or obsolescence.¹⁵⁵

This would imply that more resilient angles for security regimes aimed at military AI would involve partial bans that are focused on certain particularly destabilizing or egregious *capabilities*, complemented by standards for *usage*—contexts in which systems should not be deployed; minimum standards for safety and reliability testing—as well as restrictions on use conditions such as computing hardware.¹⁵⁶ These could be complemented by other measures: any AWS systems could be provided with a series of technical safeguards in order to mitigate the risks of inadvertent escalation.¹⁵⁷ Regimes could also draw on past arms control lessons in promoting and implementing a series of Confidence-Building Measures (CBMs)—various information- and transparency-enhancing agreements that enable parties to better understand how distinct AI capabilities would alter the implicit rules of warfare.¹⁵⁸

Nonetheless, many such partial regulations or contextual bans bring us back to a previous problem: whereas a total ban might conceivably insulate a treaty against future technological disruption, a regime that implements a partial ban on usage is vulnerable to new technological developments disrupting its mission—by driving a need for new legal development, or by shifting state incentives. As such, a key consideration, to which we will return to near the end of this chapter, concerns the adoption of strategies or designs in order to render such regimes more resilient to ongoing innovation.

¹⁵¹ Borghard and Lonergan, *supra* note 119; But see also Dumbacher, *supra* note 135.

¹⁵² Subject to all of the difficulties inherent in forecasting, as discussed in chapter 3.2.4.

¹⁵³ Example discussed in Maas, *supra* note 148 at 150.

¹⁵⁴ In section 5.5.1.

¹⁵⁵ See generally Crootof, *supra* note 51.

¹⁵⁶ Brundage et al., *supra* note 84 at 69.

¹⁵⁷ See for instance Leys, *supra* note 63 at 404–410. (discussing various ‘rules of the road’, such as ensuring that swarming AWS absorb a first bow before retaliating, installing a ‘*jus ad bellum* switch’ and an ‘auto-off’ feature, and an ‘orange box’ rule to preserve data around a certain decision to strike).

¹⁵⁸ Michael C. Horowitz, Lauren Kahn & Casey Mahoney, *The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?*, ORBIS (2020), <http://www.sciencedirect.com/science/article/pii/S0030438720300430> (last visited Sep 20, 2020).

7.2 Topology: the state of the AI governance architecture

Given the early state of AI governance at present, and the variety of proposals still on the table, it is for many cases appropriate to focus in on ‘origins’, as this can help in considering the prospects of a proposed or embryonic governance regime for a given AI issue domain.

A second perspective instead puts the emphasis on the *topology* of an emerging or mature AI regime complex, at some given moment in time. Indeed, much regime scholarship to date has occupied itself with mapping the extant ‘governance architecture’ on various issue areas of international concern.¹⁵⁹ Such work shows the value of mapping the ‘topology’ of the governance network of institutions and instruments active on an issue, in order to identify patterns of shared membership, influence, or regime interaction.

Of course, as with many facets of AI governance, it is hard and potentially futile to predict future developments in much detail. Indeed, because they are complex systems, much scholarship on regime complexes argues that it can be difficult or impossible to anticipate institutional overlap or (conflictive) norm interactions far in advance.¹⁶⁰ It could as such be argued that it is especially premature to attempt to map the precise intersections between AI norms and institutions, given the extent to which some of these remain in flux today.¹⁶¹ Yet at the same time, it remains valuable to explore the topology of the emerging AI regime complex, even at this early state, because latent tensions or fault lines amongst policies, norms or institutions may highlight the potential for future tensions or conflict.

I will therefore not aim to provide a comprehensive mapping of the AI regime complex here.¹⁶² However, it is valuable to sketch a number of dimensions along which to map these normative architectures and institutional networks at various points of development. Specifically, an analysis of the topology or state of the AI regime complex can consider three features: *demographics*, *organisation*, and *interactions*. These can be examined at one of three levels of scope (macro, meso, micro).¹⁶³

¹⁵⁹ The literature is extensive, but for overviews, see Frank Biermann et al., *The Fragmentation of Global Governance Architectures: A Framework for Analysis*, 9 GLOB. ENVIRON. POLIT. 14–40, 15 (2009); Laura Gómez-Mera, Jean-Frédéric Morin & Thijs Van De Graaf, *Regime Complexes*, in ARCHITECTURES OF EARTH SYSTEM GOVERNANCE: INSTITUTIONAL COMPLEXITY AND STRUCTURAL TRANSFORMATION 137–157, 147 (Frank Biermann & Rakhyun E. Kim eds., 2020). Rakhyun E. Kim, *Is Global Governance Fragmented, Polycentric, or Complex? The State of the Art of the Network Approach*, INT. STUD. REV., 9–10 (2019), <https://academic.oup.com/isr/advance-article/doi/10.1093/isr/viz052/5571549> (last visited Feb 16, 2020).

¹⁶⁰ As indicated by their nature as ‘complex systems’. Karen J. Alter & Kal Raustiala, *The Rise of International Regime Complexity*, 14 ANNU. REV. LAW SOC. SCI. 329–349, 333 (2018).

¹⁶¹ Notwithstanding also a degree of normative convergence. See Anna Jobin, Marcello Ienca & Effy Vayena, *The global landscape of AI ethics guidelines*, NAT. MACH. INTELL. 1–11 (2019). As well as the discussion in section 2.3.2.2.

¹⁶² However, some of the general trends and developments were sketched in Chapter 2.3. For various mappings of institutional developments, see also James Butcher & Irakli Beridze, *What is the State of Artificial Intelligence Governance Globally?*, 164 RUSI J. 88–96 (2019); Eugenio V Garcia, *Multilateralism and Artificial Intelligence: What Role for the United Nations?*, in THE GLOBAL POLITICS OF ARTIFICIAL INTELLIGENCE 18 (Maurizio Tinnirello ed., 2020); Martina Kunz & Seán Ó hÉigearthaigh, *Artificial Intelligence and Robotization*, in OXFORD HANDBOOK ON THE INTERNATIONAL LAW OF GLOBAL SECURITY (Robin Geiss & Nils Melzer eds., 2020), <https://papers.ssrn.com/abstract=3310421> (last visited Jan 30, 2019).

¹⁶³ This taxonomy is slightly distinct from the levels of analysis used by some regime complex scholars. For instance, departing from systems theory, Rakhyun Kim distinguishes between analyses focusing on ‘nodes’

7.2.1 Demographics

In terms of *demographics*, the focus is on identifying the size and composition of the governance regime complex. The focus here is on the ‘nodes’ in the overall network, in terms of the norms, institutions, or governance initiatives that apply to a given AI issue, without at this stage necessarily considering their internal relations to one another.

Such a mapping can depart from established norms and initiatives, such as the OECD Principles or the G20 Principles on Artificial Intelligence.¹⁶⁴ However, even (or especially) in a regime complex that is still in a very early stage, the mapping of demographics can extend beyond established regimes to ‘candidate’ regimes.

To return to the example of the regime complex on military AI applications, a survey of demographics can depart from ongoing (though to date inconclusive) initiatives, such as the CCW process. However, it can also extend to the wide space of solutions or principles that have been put forward. This includes the prospective application of existing international norms—such as the International Humanitarian Law principles of Distinction, Proportionality, or Humanity, or the Article 36 requirements on the legal review of new weapons.¹⁶⁵ It can also extend to solutions or principles that have been proposed but not yet implemented or embedded, such as governance regimes based on guaranteeing some forms of ‘meaningful human control’, or ‘war tort’ regimes that establish (state) accountability for LAWS.¹⁶⁶

7.2.2 Organisation: density and links

In terms of *organisation*, the focus turns to assessing the density of the institutional network, in terms of the number of institutional membership overlaps or contact points for an AI issue area. This analysis also explores the number and nature of the links between individual instruments, institutions, or initiatives. These links can be found at various levels: in terms of legal *norms*, pursued policy *goals*, *impacts*, or *inter-institutional relations*, such as overlapping memberships; formal partnerships, or operational and informational links or collaborations.

While recent years have seen a range of new governance initiatives, much of the AI regime complex is still relatively fragmented.¹⁶⁷ In particular, key initiatives—at the OECD, G20 and GPAI—have fragmented membership, and still exclude many developing states (see Figure 7.2).

(treaties, regimes, international organizations), ‘links’ (institutional interlinkages or partnerships), ‘clusters’ (regime complexes), and ‘networks’ (structure; dynamics and evolution; system). Kim, *supra* note 159 at 8.

¹⁶⁴ See also Chapter 2.3.

¹⁶⁵ Chavannes, Klonowska, and Sweijs, *supra* note 59 at 12–14.

¹⁶⁶ *Id.* at 17–23. See also Crootof, *supra* note 56.

¹⁶⁷ Peter Cihon, Matthijs M. Maas & Luke Kemp, *Should Artificial Intelligence Governance be Centralised?: Design Lessons from History*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 228–234 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375857> (last visited Feb 12, 2020); See also THORSTEN JELINEK, WENDELL WALLACH & DANIL KERIMI, *Coordinating Committee for the Governance of Artificial Intelligence* (2020), https://www.g20-insights.org/policy_briefs/coordinating-committee-for-the-governance-of-artificial-intelligence/ (last visited Jul 8, 2020).

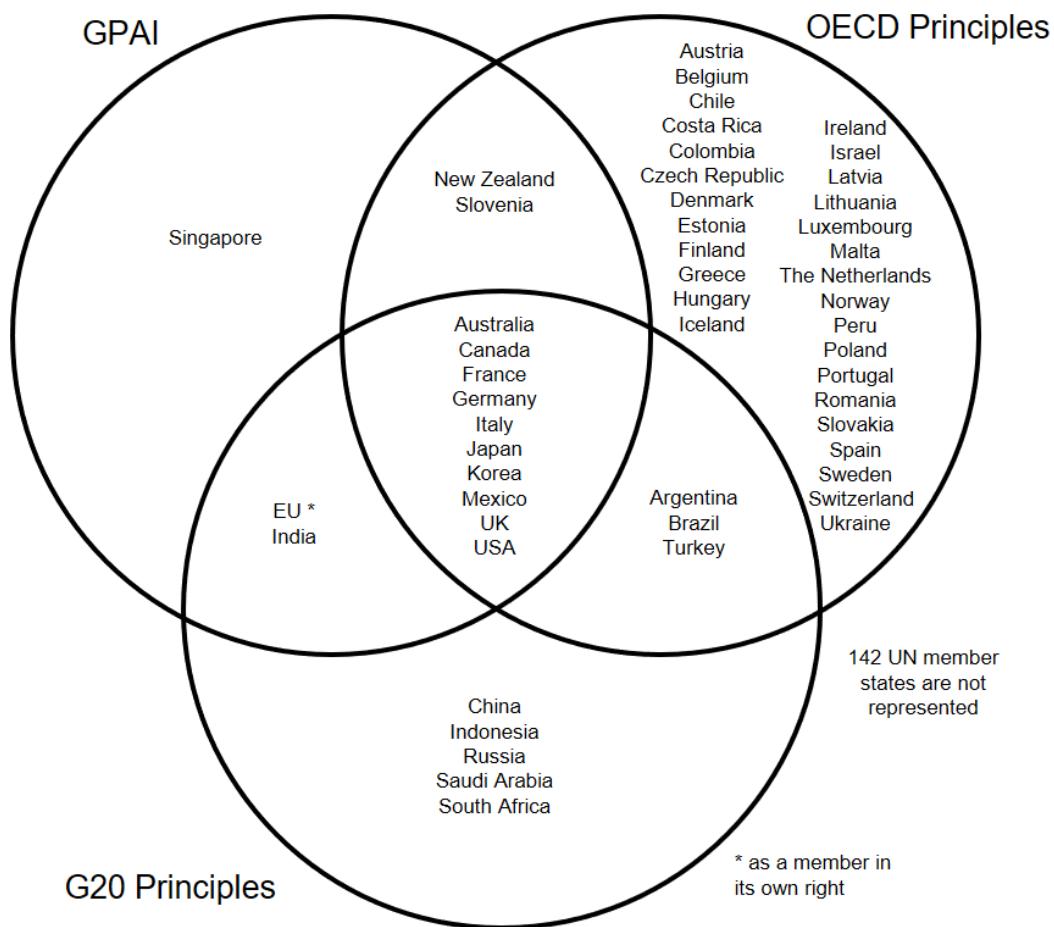


Figure 7.2. Fragmented membership in international AI policy initiatives¹⁶⁸

Nonetheless, amongst these existing AI governance initiatives, there are encouraging signs of emerging coordination as well. This includes incipient *normative* linkages, as seen by the fact that both the G20 Osaka Statement on Trade and Digital Economy, as well as the GPAI, have adopted the OECD's AI principles.¹⁶⁹ There are also growing *institutional* linkages, such as the housing of the GPAI Secretariat under the OECD.¹⁷⁰ Indeed, at the multilateral level there has already been institutional coordination or at least exchange amongst various UN bodies and

¹⁶⁸ Reproduced with permission from forthcoming paper: Peter Cihon, Matthijs M. Maas & Luke Kemp, *Fragmentation and the Future: Investigating Architectures for International AI Governance*, 3 (Forthcoming).

¹⁶⁹ G20, *G20 Ministerial Statement on Trade and Digital Economy* (2019), <https://www.mofa.go.jp/files/000486596.pdf> (last visited May 25, 2020); Global Partnership on Artificial Intelligence, *Joint Statement from founding members of the Global Partnership on Artificial Intelligence* (2020), <https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding>.

¹⁷⁰ OECD, *OECD to host Secretariat of new Global Partnership on Artificial Intelligence* (2020), <https://www.oecd.org/going-digital/ai/OECD-to-host-Secretariat-of-new-Global-Partnership-on-Artificial-Intelligence.htm> (last visited Jul 3, 2020).

agencies, going back several years, on topics around emerging digital technologies and digitalisation. For instance, the ITU's 2019 Summit on AI for Global Good listed 37 UN system entities as partners.¹⁷¹ There are further proposals for further institutions or bodies that would be capable of 'metagovernance' and coordination of AI institutions and regimes.¹⁷² While it remains to be seen to what degree such exchanges will result in aligned policies, they are at least reflective of an increasing institutional density. Even if such efforts do not reduce the institutional fragmentation of the overall AI regime complex, they might at least mitigate its adverse interactions, and ensure potential institutional or norm conflicts are effectively managed.¹⁷³

7.2.3 Interactions: gaps, conflicts, cooperation, synergies

In terms of *interactions*, the focus is on the character of the links between the different institutions or instruments. These linkages can (1) leave *gaps* in governance, where certain issues remain in a 'non-regime' state.¹⁷⁴ They can be (2) *conflictive*, either in the form of outright norm inconsistencies or incompatibilities, or more commonly, in terms of operational externalities,¹⁷⁵ or even inter-institutional 'turf wars'. These linkages can be (3) *cooperative*, when there is only loose integration but the relationship between norms, principles or policies remains unclear; or they can be (4) *synergistic*, with mutually reinforcing norms, or useful institutional divisions of labour.¹⁷⁶

It will be important to examine how and where various AI standards or framework principles are articulated across different AI governance instruments, and whether or not these are aligned. For instance, while in the debate over the international regulation of LAWS, the concept of 'meaningful human control' has seen broad support from the international community, this idea is still subject to a wide diversity in articulation and interpretation.¹⁷⁷ If left unaddressed or unclarified, this ambiguity could spell the emergence of potential norm conflicts amongst later instruments or regimes. By contrast, in other areas, the normative and institutional links between certain bodies, such as the OECD and the GPAI, appear to be more cooperative or even synergistic, at least at this stage. Nonetheless, the current tapestry of institutions and norms also appears to leave a number of potential gaps on certain issue areas, such as the use of AI in cyberwarfare, non-LAWS military decision-making, or the implications of potential future AI capabilities.¹⁷⁸

¹⁷¹ ITU, *AI for Good Global Summit 2019 Insights* (2019), <https://itu.foleon.com/itu/aiforgood2019/home/> (last visited Oct 6, 2019). As cited in Martina Kunz, *Global Artificial Intelligence Governance: Rationale, Architecture and Challenges* 10, 6–7 (2019). (arguing that even if the AI regime is fragmented, there is extensive inter-institutional coordination).

¹⁷² JELINEK, WALLACH, AND KERIMI, *supra* note 167 at 3.

¹⁷³ On the discussion of the management of fragmentation, see also section 7.5.3.

¹⁷⁴ Radoslav S. Dimitrov et al., *International nonregimes: a research agenda*, 9 INT. STUD. REV. 230–258 (2007).

¹⁷⁵ Morin et al., *supra* note 134 at 2–3.

¹⁷⁶ This loosely follows Biermann et al., *supra* note 159 at 19–21. However, it extends their framework with an additional category of 'gaps'.

¹⁷⁷ Chavannes, Klonowska, and Sweijns, *supra* note 59 at 17–19.

¹⁷⁸ Cihon, Maas, and Kemp, *supra* note 167 at 229.

7.2.4 Scope: macro, meso, micro

Finally, these interactions can be examined at three levels of the (AI) governance architecture. (1) At the *macro* level, this considers the interactions of the AI regime complex with other regime complexes (e.g. international security; data privacy; export controls; transport), or with general norms of international law. (2) At the *meso* level, the emphasis is on interactions of different AI regimes which, at least nominally, focus on distinct issue areas (e.g. the interaction of the prospective CCW regime on LAWS with broader AI ethics or standards initiatives, or with hypothetical regimes to prevent the proliferation of malicious AI applications). (3) At the *micro* level, finally, one can focus on the internal dynamics between different institutions focused on the same issue area. This could consider, for instance, the current situation where distinct regimes or fora articulate different and only partially overlapping definitions of ‘meaningful human control’.¹⁷⁹

7.3 Evolution: trends and trajectories in AI regime complexity

However, analysing the topology of the AI regime complex at any given moment yields only a static snapshot. It is important to also address the dynamic interplay and evolution of a regime complex. As has been discussed earlier,¹⁸⁰ the AI governance landscape remains relatively incipient and fragmented today. However, should this be interpreted as the early growing pains of a regime complex on a natural path towards an eventual integrated global (treaty) regime for AI? It can be valuable to explore which factors or developments could determine or influence whether the AI governance architecture continues along in a fragmented regime complex, or whether—or under what conditions—we might expect the eventual emergence (or at least possibility) of an integrated global regime for all or some AI issues.

7.3.1 Caveats: the limits of predicting or designing a regime complex

Before discussing such trajectories, it is important to provide two key caveats. In the first place, *predicting* regime evolution is precarious. Anticipating the future developments of any political system or governance architecture is difficult under the best of circumstances. That goes doubly so for a regime complex, given its nature as a dynamic complex system. As such, it is possible that few well-calibrated predictions can be made in advance. Nonetheless, granting this challenge, it can still be possible to identify factors or drivers that would play a role in supporting distinct conditional trajectories of regime development.

A second and related problem is that deliberately *designing* a regime complex trajectory is also hard. Significantly, many scholars note that a regime complex is rarely if ever the product of rational or intentional design.¹⁸¹ Instead, fragmented regime complexes are the result of a

¹⁷⁹ Chavannes, Klonowska, and Sweijs, *supra* note 59 at 17–19.

¹⁸⁰ In Chapter 2.3.

¹⁸¹ David J. Galbreath & Sascha Sauerteig, *Regime complexity and security governance*, in HANDBOOK OF GOVERNANCE AND SECURITY 82–97, 85 (James Sperling ed., 2014), <https://www.elgaronline.com/view/edcoll/9781781953167/9781781953167.00014.xml> (last visited Oct 15, 2019). Notably, this is the case even if a

combination of local ‘micro’ decisions by diverse actors—such as, say, a state’s decision over whether to support only gradual norm development on an issue, or to amend an existing institution to this purpose, or to pursue the establishment of a new institution to deal with AI’s security impacts. The point is that even if actors’ decisions in many of these areas are based on ‘rational’ calculations towards achieving deliberate goals, they cannot anticipate or determine how this feeds into the emerging regime complex.

Nonetheless, granting these two caveats, it is still valuable to draw on existing regime complex scholarship to anticipate some of the drivers of different trajectories, identify their implications, and if possible identify early warning signs of different scenarios. In doing so, we can take stock of both general background trends in global governance, as well as, more speculatively, consider the effects of AI ‘governance disruption’ on these architectures.

7.3.2 General drivers of regime fragmentation

While some scholars in AI governance envision centralised international AI organisations,¹⁸² one claim of regime complex scholars is that such ambitions may face an uphill battle, given general background trends in the texture and architecture of global governance. Indeed, diverse studies have chronicled how, over the past few decades, and across diverse issue areas, governance has tended towards institutional proliferation, fragmentation, and regime complexity.¹⁸³

At the surface, this trend of governance entropy may appear anomalous. After all, from a rationalist perspective, one might expect states to prefer to avoid the transaction costs and potential policy and norm conflicts involved with creating, navigating and managing many parallel institutions, especially if their work potentially overlaps.¹⁸⁴ In this perspective, one might expect states or other actors to take a cautious response, stick to working with small numbers of centralised institutions, and attempt to preserve the unity or coherence of the international legal system.

Yet transaction costs are not the only considerations or factors that affect decisions around the creation of new institutions. As noted previously, the absence of an international legislative supreme body to coordinate amongst regimes, and the need for distinct regimes to respond to issues in diverse areas has long been seen as driving the institutional and normative

decentralised regime complex of separate, overlapping institutions were in fact the desired, preferable governance arrangement.

¹⁸² JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE (2018); Olivia J Erdelyi & Judy Goldsmith, *Regulating Artificial Intelligence: Proposal for a Global Solution*, in PROCEEDINGS OF THE 2018 AAAI / ACM CONFERENCE ON ARTIFICIAL INTELLIGENCE, ETHICS AND SOCIETY 95–101 (2018), http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf; Luke Kemp et al., *UN High-level Panel on Digital Cooperation: A Proposal for International AI Governance* (2019), https://digitalcooperation.org/wp-content/uploads/2019/02/Luke_Kemp_Submission-to-the-UN-High-Level-Panel-on-Digital-Cooperation-2019-Kemp-et-al.pdf.

¹⁸³ Kenneth W. Abbott, Jessica F. Green & Robert O. Keohane, *Organizational Ecology and Institutional Change in Global Governance*, 70 INT. ORGAN. 247–277 (2016).

¹⁸⁴ Thijs Van de Graaf, *Fragmentation in Global Energy Governance: Explaining the Creation of IRENA*, 13 GLOB. ENVIRON. POLIT. 14–33, 15 (2013). See also Johannes Urpelainen & Thijs Van de Graaf, *Your Place or Mine? Institutional Capture and the Creation of Overlapping International Institutions*, 45 BR. J. POLIT. SCI. 799–827 (2015).

fragmentation of the global governance complex.¹⁸⁵ As such, it has been suggested that regime complexity, even where it appears as a deliberate governance choice, may often be a nearly unavoidable outcome of a set of broader and deeper trends in world politics.¹⁸⁶ For instance, Alter & Raustiala have noted seven interrelated trends which, they argue, together generate and sustain regime complexity. These are: density, accretion, power shifts over time, preference changes, modernity, representation and voice goals, and local governance.¹⁸⁷

We should not, all else being equal, expect AI governance to remain insulated from such broad trends in the global governance architecture. As such, it is valuable to briefly consider why and how these distinct trends generate regime complexity—and whether or not we might expect AI governance to be affected, or to resist these trends towards regime fragmentation.

7.3.2.1 *Institutional density*

In terms of *density*, the argument is that the general growth in the quantity and diversity of global institutions or agreements¹⁸⁸ almost naturally results in an increasing overlap between institutions on many issue areas, ensuring that conflicts are increasingly common. The point is not that this growth necessarily drives norm drift (i.e. making individual agreements more likely to be in conflict with one another), but rather that, through a combinatorial explosion of contact surfaces amongst agreements and institutions, it exacerbates or activates more latent conflicts.

Institutional density could prove particularly strong at driving fragmentation in the AI regime complex. As noted, AI creates a diverse range of challenges. Some of these may be genuinely unprecedented, but many also certainly overlap and interact with issue areas—from privacy to cybercrime, from human rights to security—which are already embedded within existing agreements or regimes. For instance, any new global governance initiatives on privacy or some right to explanation for algorithmically-based decision-making would have to reckon with the EU’s General Data Protection Regulation (GDPR). Even if such conflicts are only partial, or occur in certain domains where regimes are not dense (or even largely absent), the sheer breadth of AI use cases makes it very likely for there to be some overlap.¹⁸⁹

To be certain, there may be some AI techniques or approaches that are more pertinent to some applications or domains than to others. For the sake of argument, ‘Generative Adversarial Networks’ (GANs) underpin the generation of DeepFakes, and have also featured in a range of cybersecurity applications, including systems that can generate likely candidates for human passwords, or synthetic fake ‘fingerprints’ that serve as master keys to biometric identification

¹⁸⁵ See also the discussion in Chapters 3.2.2. and 6.2.

¹⁸⁶ Alter and Raustiala, *supra* note 160 at 337. (“in a deeper sense, a regime complex is often an almost-inevitable result of broader trends in world politics that make regime complexity almost unavoidable.”).

¹⁸⁷ *Id.* at 337.

¹⁸⁸ See also chapter 6.1. However, for an argument that these trends have slowed, see J. Pauwelyn, R. A. Wessel & J. Wouters, *When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking*, 25 EUR. J. INT. LAW 733–763 (2014).

¹⁸⁹ Alter and Raustiala, *supra* note 160 at 339. (“Perhaps most significant, because the density of the global governance ecosystem is so high, any attempt to define a particular problem—to in essence bring together shared sets of issues and create a single regime to govern a given space—is likely to replicate the regime complex problem insofar as new institutions will almost inevitably overlap with preexisting elemental regimes”).

systems,¹⁹⁰ However, these techniques may have less application in facial recognition technologies, or in systems involved in the control of LAWS. However, such cases are rough exceptions, and few generalisations (e.g. ‘reinforcement learning underpins all problematic military applications’) made on this basis would be very meaningful or a good guide to policy.

Instead, the wide versatility of basic AI paradigms, and the intrinsically ‘dual-use’ nature of the underlying computing hardware may ensure that these ‘AI inputs’ become subject to many overlapping and conflicting demands. Indeed, distinct AI governance regimes which rely on policy levers such as technical standards or export controls to regulate AI within their area of operation,¹⁹¹ may be hard-pressed to avoid inadvertently overlap (on the level of norms, goals or impacts) with antecedent regimes that operate in adjacent areas. Already today, export control regimes have struggled to avoid conflict and ensure harmonisation.¹⁹² Digitisation and AI could enhance such trends.¹⁹³ As such, at a first approximation, the condition of institutional density might support fragmentation in AI governance.

7.3.2.2 Accretion

In terms of *accretion*, while states might sometimes seek to vest authority in existing organisations, there may be many situations where they, faced with the path dependence of existing institutions,¹⁹⁴ or potential contestation over new solutions, find it more attractive to create new institutions than to try to amend existing ones.¹⁹⁵ Beyond attempts to sidestep gridlock, institutional accretion might also be the result of symbolic politics—the establishment of a ‘new institution’ potentially signalling to global or domestic publics a greater investment and more meaningful step to combat a high-profile global problem (the ‘turning of a page’) than would alterations of existing institutions, which signal a ‘business-as-usual’. Prima facie, one might accordingly expect the apparent ‘novelty’ of AI technology to spur the creation of new institutions.

Yet at present, it is unclear whether AI issues are sparking institutional accretion. On the one hand, the proliferation of AI ethics codes and principles may speak to a general perception that AI issues remain unaddressed or inadequately governed. Yet institutionally, the currently dominant approach to regulating military uses of AI is the GGE process under the remit of the

¹⁹⁰ Briland Hitaj et al., *PassGAN: A Deep Learning Approach for Password Guessing*, ARXIV170900440 Cs STAT (2017), <http://arxiv.org/abs/1709.00440> (last visited Feb 20, 2019); Philip Bontrager et al., *DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution*, ARXIV170507386 Cs (2017), <http://arxiv.org/abs/1705.07386> (last visited Feb 20, 2019); Discussed in Keith J Hayward & Matthijs M Maas, *Artificial intelligence and crime: A primer for criminologists*, CRIME MEDIA CULT. 1741659020917434, 8 (2020).

¹⁹¹ Jade Leung, Sophie-Charlotte Fischer & Allan Dafoe, *Export Controls in the Age of AI*, WAR ON THE ROCKS (2019), <https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/> (last visited Sep 2, 2019).

¹⁹² KOLJA BROCKMANN, *Challenges to multilateral export controls: The Case for Inter-regime Dialogue and Coordination* 40 (2019), <https://www.sipri.org/publications/2019/other-publications/challenges-multilateral-export-controls-case-inter-regime-dialogue-and-coordination>.

¹⁹³ NELSON, *supra* note 102.

¹⁹⁴ Kathleen Thelen, *Historical Institutionalism in Comparative Politics*, 2 ANNU. REV. POLIT. SCI. 369–404 (1999); Lucio Baccaro & Valentina Mele, *Pathology of Path Dependency? The ILO and the Challenge of New Governance*, 65 ILR REV. 195–224 (2012) (examining the ILO’s inability to undertake certain reforms, in the late 1990s and early 2000s, to increase the involvement of NGOs, or to adopt quantitative indicators of impact).

¹⁹⁵ For instance, Alter & Raustiala give the example of how “even though the World Health Organization has existed since 1948, in 2000 states created the Global Fund to Fight Aids, Tuberculosis, and Malaria to provide a more flexible, private sector-oriented alternative.” Alter and Raustiala, *supra* note 160 at 337.

Convention on Certain Conventional Weapons.¹⁹⁶ This is not the only approach proposed,¹⁹⁷ though it remains the most developed. More generally, while many UN bodies have begun explorations of the impact of AI, this has remained largely within the remit of existing institutions such as the ITU or UNICRI.

That is not to say there have not been attempts at creating new institutions. The most unambiguous example of accretion may be the 2018 French-Canadian proposal for a new ‘International Panel on AI’.¹⁹⁸ While this was initially stalled by US opposition at the August 2019 G7 summit, the US has since come on board to the renamed ‘Global Partnership on AI’, launched in June 2020.¹⁹⁹ However, even here, the new initiative has seen its Secretariat housed at the OECD.²⁰⁰ As such, to a first approximation, it remains unclear to what extent institutional accretion plays a large role in driving AI governance fragmentation one way or the other.

7.3.2.3 Power shifts over time

In terms of *power shifts over time*, the fact that older institutions often reflect outdated power structures may ensure that rising powers or new actors prefer to create new institutions instead. Indeed, Acharya has argued that the very fact of an emerging ‘multiplex’ world order speaks in favour of governance fragmentation as being more reflective of the plurality of values than governance structures that reflect ‘Western’ interests.²⁰¹ It may certainly be possible that parties that seek to pursue AI governance may prefer to do so in new fora which in their view better reflect the new balance of power. Indeed, arguably we are seeing such moves by private industry, with the previous establishment of the Partnership on AI.²⁰² Moreover, it is possible that the use of some AI capabilities could accelerate this driver of fragmentation, insofar as it enables and catalyses shifts in global power relations.²⁰³

At the same time, the set of states that currently lead in AI development overlaps in part with the set of states that have been historically powerful in shaping international institutions. This suggests that the overall impact of this factor—whether or not supporting fragmentation—

¹⁹⁶ See also the discussion in Chapter 2.3.2.3.1.

¹⁹⁷ For an overview, see also Chavannes, Klonowska, and Sweijs, *supra* note 59 at 12–16.

¹⁹⁸ Justin Trudeau, *Mandate for the International Panel on Artificial Intelligence*, PRIME MINISTER OF CANADA (2018), <https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence> (last visited Jul 6, 2019).

¹⁹⁹ Ministère de l’Europe et des Affaires étrangères, *Launch of the Global Partnership on Artificial Intelligence by 15 founding members*, FRANCE DIPLOMACY - MINISTRY FOR EUROPE AND FOREIGN AFFAIRS (2020), <https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding> (last visited Jun 15, 2020); Global Partnership on Artificial Intelligence, *supra* note 169.

²⁰⁰ OECD, *supra* note 170.

²⁰¹ Amitav Acharya, *The Future of Global Governance: Fragmentation May Be Inevitable and Creative Global Forum*, GLOB. GOV. 453–460 (2016).

²⁰² Eric Horvitz & Mustafa Suleyman, *Introduction from the Founding Co-Chairs*, THE PARTNERSHIP ON AI (2016), <https://www.partnershiponai.org/introduction-from-the-founding-co-chairs/> (last visited Sep 10, 2020). However, in June 2020, Baidu quit the partnership, leaving it without any Chinese members. Will Knight, *Baidu Breaks Off an AI Alliance Amid Strained US-China Ties*, WIRED, 2020, <https://www.wired.com/story/baidu-breaks-ai-alliance-strained-us-china-ties/> (last visited Jun 21, 2020).

²⁰³ We will return to this discussion in the next section (7.3.3.), on how Governance Disruption might affect these dynamics.

may depend on the degree to which states such as China (both advanced in AI as well as rising power) will prefer to create new institutions. As such, power shifts may modestly suggest further fragmentation.

7.3.2.4 *Preference changes*

In terms of *preference changes*, it is possible that anticipated shifts in relative power or in domestic politics may mean that it is not just new rising powers but also powerful actors themselves who may become unsatisfied with the existing state of rules and structures.

Given the considerable amount of hope and hype around the potential of AI technology—and widespread allegations of an ‘arms race’ between the US and China—it is possible that the technology’s anticipated strategic gains could lead these powerful actors to forego regulating AI in existing governance fora. Alternatively, increasing automation could lead to shifts and upheavals in the domestic politics of some states, and this can alter their preferences regarding the adequacy of existing institutions. If interests shift, this could lead coalitions of dissatisfied states to pursue ‘regime shifting’—attempts to “alter the status quo ante by moving treaty negotiations, lawmaking initiatives, or standard setting activities from one international venue to another”²⁰⁴—or even ‘competitive regime creation’.²⁰⁵ Either of these would drive fragmentation, or exacerbate its adverse consequences.

7.3.2.5 *The complexity of modernity*

Moreover, some systems theorists have argued that *modernity* by its nature generates complexity, which creates an unavoidable demand for specialisation and functional differentiation in global governance.²⁰⁶ Such scholars argue that many modern challenges—such as climate change, global migration, or AI and robotics—must be approached as ‘complex issues’ that exceed the sectoral remit of pre-existing, formal institutions.²⁰⁷ AI no doubt gives rise to an array of complex societal problems; catalyses new problems in existing fields, and throws up peculiar challenges for easy legal categorisation. If that is so, the technology’s inherent complexity and cross-domain applicability might be prone to driving global regime complexity.

7.3.2.6 *Representation and voice goals*

The expansion of participation in global governance also generates new *representation and voice goals*. The demand for greater voice and inclusion in global governance by many international institutions and non-state actors, especially in areas of internet governance and sustainability, has led to the creation of many new fora where they perceive their voices are better heard. In some cases, this can lead to these actors shifting dialogue to different regimes. In the case of AI, it is possible that AI researcher communities and private tech companies, along with

²⁰⁴ Laurence Helfer, *Regime Shifting: The TRIPs Agreement and New Dynamics of International Intellectual Property Lawmaking*, 29 YALE J. INT. LAW 1–83, 14 (2004).

²⁰⁵ Julia C. Morse & Robert O. Keohane, *Contested multilateralism*, 9 REV. INT. ORGAN. 385–412 (2014).

²⁰⁶ Alter and Raustiala, *supra* note 160 at 338; Michael Zürn & Benjamin Faude, *Commentary: On Fragmentation, Differentiation, and Coordination*, 13 GLOB. ENVIRON. POLIT. 119–130 (2013).

²⁰⁷ Morin et al., *supra* note 134.

a broad array of civil society actors and ‘norm entrepreneurs’²⁰⁸ will be similarly engaged. Indeed, this may already be seen with issues such as LAWS: in response to the (perceived) slow progress by the GGE, some activists have threatened to shift to another forum, as they in the past have with the Ottawa Treaty that banned landmines.²⁰⁹

7.3.2.7 Shift towards local governance

Finally, there may be general shifts towards ‘local governance.’ States may prefer regional organisations and organisations staffed with local officials over ‘remote’ global ones, where it comes to dealing with certain problems. While this may increase procedural legitimacy and the responsiveness of policies, this may also sometimes lead to conflict between local and global organisations.²¹⁰

As such, while some have noted the importance, in international AI regulation, of applying principles of subsidiarity,²¹¹ it remains unclear to what extent or how this shift to local governance will affect the governance of AI. On the one hand, the technology’s impact is transnational, suggesting regional initiatives might not suffice, or cannot achieve important harmonisation. On the other hand, as issues such as data bias have shown, AI systems may affect different regions in subtly different ways, suggesting there may be an increased pressure for diversity rather than harmonisation. In many cases of standard-setting or commercial regulation, these will likely be regional in the first instance.²¹² More generally, there could emerge distinct ‘technocultural’ differences amongst regions on AI issues,²¹³ which might strengthen the trend or preference of various parties towards local governance arrangements.

In sum, reviewing these six drivers, there appear to be many forces or trends that prime the global governance system towards institutional proliferation and regime complexity on many

²⁰⁸ Martha Finnemore & Kathryn Sikkink, *International Norm Dynamics and Political Change*, 52 INT. ORGAN. 887–917 (1998).

²⁰⁹ Janosch Delcker, *How killer robots overran the UN*, POLITICO, 2019, <https://www.politico.eu/article/killer-robots-overran-united-nations-lethal-autonomous-weapons-systems/> (last visited Oct 2, 2019). See also the commentary by the Campaign To Stop Killer Robots, in reaction to the slow progress during the September 2020 session: Campaign to Stop Killer Robots, *Diplomatic Talks in 2020* (2020), <https://www.stopkillerrobots.org/2020/09/diplomatic2020/> (last visited Sep 26, 2020). (“If it is not possible to launch negotiations on a new treaty by the CCW Review Conference in December 2021, then it will be time to find another forum to not only discuss content, but achieve this goal”).

²¹⁰ This may raise particular problems in international criminal law. For instance, the proposed Malabo Protocol, if adopted, would confer an international criminal law jurisdiction on the African Court of Human and Peoples’ Rights, which might intersect with that of the International Criminal Court. Matiangai Sirleaf, *The African Justice Cascade and the Malabo Protocol*, 11 INT. J. TRANSITIONAL JUSTICE 71–91 (2017). As discussed in Alter and Rautiala, *supra* note 160 at 338.

²¹¹ TURNER, *supra* note 182 at 249–250.

²¹² I thank Charlotte Siegmann for this observation.

²¹³ Osonde A. Osoba, *Technocultural Pluralism: A “Clash of Civilizations” in Technology?*, in PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 132–137 (2020), <http://dl.acm.org/doi/10.1145/3375627.3375834> (last visited Feb 12, 2020). See also the discussion of Asian perspectives on AI ethics. Danit Gal, *Perspectives and Approaches in AI Ethics: East Asia*, in OXFORD HANDBOOK OF ETHICS OF ARTIFICIAL INTELLIGENCE (Markus Dubber, Frank Pasquale, & Sunit Das eds., 2019), <https://papers.ssrn.com/abstract=3400816> (last visited Oct 2, 2019). However, others have argued that cultural value differences in AI may be exaggerated, or at least can be overcome; see Séan S. Ó hÉigearthaigh et al., *Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance*, PHILOS. TECHNOL. (2020), <https://doi.org/10.1007/s13347-020-00402-x> (last visited May 17, 2020).

issues. In at least some cases, we could expect these to apply to AI governance, supporting (though certainly not mandating) continued or increasing fragmentation in this regime complex.

7.3.3 AI governance disruption at the intersection with regime complexity

Moreover, these trajectories could be affected by the governance disruption effected by AI applications itself. As detailed above, contemporary regime complex theory has explored many drivers of regime complex evolution and development. However, while scholars have reckoned with the effects of exogenous political or institutional shocks to either the trajectory of a regime complex,²¹⁴ or to the continued viability of the international liberal order as a whole,²¹⁵ there appears to be relatively little examination of the effects of technology-driven change on regime complexity. This is unfortunate because, as the lens of governance disruption has shown, the use of various technologies can in some cases exert considerable effects on the norms, processes, or political scaffolding of the international legal order.

As such, it may be productive to explore the implications of AI-driven governance disruption lens for the processes of regime complex evolution. Such examination is certainly more speculative, and it should be re-emphasised that the following sections are not meant as strong predictions, but as conditional scenarios that anticipate the implications of various types of AI-driven governance disruption on regime complexity. For the sake of argument, this following assumes no- or only very modest further capability developments in the available AI capabilities,²¹⁶ although it does weakly assume the continuing dissemination or application of existing AI capabilities into more areas and domains, in order to anticipate the implications of distinct patterns of dissemination and usage on regimes.

7.3.3.1 *Development and Complexity: AI as generator of regime fault lines*

At a structural level, how will AI's ability to generate a need for development in the substance, laws or norms of international law affect the trajectory of a regime complex?

One key caveat here is that global governance obviously consist of a wider range of norms and governance instruments than solely those embedded in writ 'hard law'. Critically, many of these softer governance instruments are potentially more resilient to 'legal' disruption—because they are already phrased broadly, depend on discretion for implementation, or are more flexible and easy to review and update in the face of changing technological circumstances.²¹⁷ Nonetheless, even if this means that such soft governance instruments are slightly better

²¹⁴ Hanzhi Yu & Lan Xue, *Shaping the Evolution of Regime Complex: The Case of Multiactor Punctuated Equilibrium in Governing Human Genetic Data*, 25 GLOB. GOV. REV. MULTILATERALISM INT. ORGAN. 645–669 (2019).

²¹⁵ See Karen J. Alter, *The Future of International Law*, 101 ICOURTS WORK. PAP. SER. (2017), <https://papers.ssrn.com/abstract=3015177> (last visited Jun 11, 2020). (examining the prospects for the global liberal order if the US turns away from its values). See also the discussion in chapter 5.5.2.

²¹⁶ However, this may be an overtly pessimistic assumption that may well be proven wrong by continued progress in the coming years. See also Chapter 2.1.6.

²¹⁷ This is one reason why Crootof considers 'soft law' a potentially promising alternate avenue for the international regulation of certain new technologies, especially when compared to the propensity for hard-law treaties to be rendered obsolete by advancing technology, and the problems involved with then amending them. Crootof, *supra* note 51 at 124–126.

insulated from—or at least better able to adapt to—AI-driven ‘legal’ shocks, the conceptual and practical changes forced by the use of these systems still bear on other forms of governance, and may (or arguably should) warrant re-examination.

Taking this into account, the prospect of AI systems driving legal development could, paradoxically, exert two divergent effects on regime complexes. On the one hand, situations of clear inadequacy of existing legal instruments or concepts might serve as an opportunity to clarify long-present ambiguities in established global norms, and harmonize latent conflicts between existing regimes, by getting these out in the open, in the context of an urgent governance problem that requires addressing. At the same time, the broad-spectrum legal disruption that is generated by AI systems might illustrate the confusing and fragmenting effects of ‘modernity’, which, as noted, some regime systems theorists point to as a major factor in the trend towards fragmentation and regime complexity.²¹⁸

In this way, the precise effects of AI-driven governance development on the trajectory of a governance architecture will likely depend sensitively on the pre-existing configuration of that architecture. Indeed, it some ways, it could exacerbate pre-existing tendencies.

On the one hand, within an already-integrated regime complex, a centralised, authoritative international institution could in principle seize upon AI’s legal disruption within one area in order to kick-start and facilitate a broader dialogue about long-overdue systemic revision or legal innovation in underlying rules or concepts. For instance, such an institution could seize upon specific problems posed by AI-enabled military surveillance platforms or ‘lethality-enabling technologies’,²¹⁹ to address the broader fact that the distinction between ‘war’ and ‘peacetime’—a cornerstone of IHL—has, in practice already become blurred by new technologies.²²⁰ Such legal development would not necessarily imply an abandonment of these well-proven frameworks; but it would reckon with the ‘transversal’ effects of trends in technology across different fields of (international) law, in order to yield more integrated bodies of law. Of course, even if centralised international institutions could in principle carry out such regulatory innovation and integration, it is not a given that they would also do so.²²¹

On the other hand, starting from an already-fragmented institutional context—and across fragmented application domains—AI technology’s propensity to generate situations of legal Disruption (requiring development) may well exacerbate the forces of (further) fragmentation. Given the extremely diverse set of perspectives through which one can approach AI technology,²²² it seems unlikely that a fragmented regime complex would reliably be able to organically converge towards more integrated or harmonised policy responses. Similar AI architectures will be used across widely different contexts; in each of those issue areas, certain self-similar features of AI systems (such as algorithmic unpredictability or opacity, or the susceptibility to adversarial

²¹⁸ See also Morin et al., *supra* note 134.

²¹⁹ Michel, *supra* note 24. For a recent report that does seek to explore the question of arms control, not just for LAWS, but for broader uses of AI in military decision-making, see PAOLI ET AL., *supra* note 121.

²²⁰ Garcia, *supra* note 57; Braden Allenby, *Are new technologies undermining the laws of war?*, 70 BULL. AT. SCI. 21–31 (2014) (highlighting drones and cyberwarfare). See also Crootof, *supra* note 94 at 10. And see generally Brooks, *supra* note 57.

²²¹ Indeed, the question of whether, or under what conditions, international organizations would be capable of this, remains an open and interesting one.

²²² As discussed in Chapter 2.1.1.

input) can be refracted into seemingly-distinct local problems. If many distinct institutions and organisations are forced to grapple, in parallel and relative isolation, with certain local questions (such as the variable meaning of ‘Meaningful Human Control’ for LAWS, in healthcare, and in other contexts)) of underlying conceptual or legal disruption, it is likely that we may see distinct institutions and regimes resolve and decide these cases in different, and potentially contradictory ways. In that way, the proliferation of AI could trigger a scattering of individual regimes’ norms and policies, in ways that drive regime complexity, and that open up considerable scope for conflicts in terms of regime norms, operations, or impact.

7.3.3.2 *Displacement and Complexity: AI as shield, patch, cure or accelerator of fragmentation*

As one subset of governance disruption, the phenomenon of governance displacement also highlights ways in which the integration or incorporation of AI tools into the practices of global governance might affect regime complexity. In brief, some AI tools, especially if they were accessible (or distributed) to many actors (whether states or non-state observer agencies or NGOs), could play some role in mitigating at least some trends towards regime complexity. They could variably serve as *shield* to the negative consequences of complexity; as *patch* to halt the further fragmentation of regimes, or as *cure* to avert or mitigate latent conflicts and aid in harmonisation of regimes. Yet, left unregulated, the free-for-all use of AI systems could also prove an *accelerator* of processes of legal and governance fragmentation.

For one, various actors might use AI tools to *shield* themselves from the negative operational consequences of a fragmented and distributed regime complex. For instance, the use of these tools in support of many ‘routine’ diplomatic tasks could free up the limited diplomatic resources or staff available to many smaller states, enabling them to participate more fully in various international fora, ‘levelling’ the playing field relative to the large and well-staffed foreign ministries of larger states.²²³ More modestly, through translation and text (e.g. news) summarisation services, such AI tools could facilitate the ability of even less powerful actors to navigate dense regimes complexes, counteracting the democratic deficit identified by some concerned scholars.²²⁴

More actively, actors could also use AI tools as a *patch* to halt or pre-empt the further fragmentation of various regimes. For instance, Ashley Deeks notes how various ‘text-as-data-tools’ might be used to pre-emptively identify treaty conflicts (whether accidental or strategically engineered). She suggests that “[a] state could create a tool that allows it to compare proposed treaty language (either while the negotiations are underway, or before the state has ratified the treaty) to all other treaties to which the state is a party to detect similar language or very similar topics.”²²⁵ This could pre-empt or deter strategic efforts by states to create certain treaty conflicts

²²³ KATHARINA E. HÖNE, *Mapping the challenges and opportunities of artificial intelligence for the conduct of diplomacy* (2019), <https://www.diplomacy.edu/sites/default/files/AI-diplo-report.pdf> (last visited Mar 14, 2019).

²²⁴ Daniel W. Drezner, *The Power and Peril of International Regime Complexity*, 7 PERSPECT. POLIT. 65–70 (2009); Eyal Benvenisti & George W. Downs, *The Empire’s New Clothes: Political Economy and the Fragmentation of International Law*, 60 STANFORD LAW REV. 595–632 (2007).

²²⁵ Ashley Deeks, *High-Tech International Law*, 88 GEORGE WASH. LAW REV. 575–653, 616 (2020).

in order to set the agenda for future negotiations,²²⁶ because any negotiation partners would now be able to spot when such a strategy is being attempted.²²⁷

In principle, such AI tools could not only help spot and patch imminent conflicts produced by specific treaties, but could also be deployed as a modest *cure* to more general pre-existing patterns of fragmentation in international law. For instance, Deeks suggests that states could use such tools even in the absence of specific negotiations, simply to map any tensions amongst their existing treaty commitments, “in order to identify gaps or other potential conflicts before they produce a real world problem.”²²⁸ Accordingly, other actors or international organisations could use such tools to chart, to some degree, major fault lines or gaps amongst international law’s fragmented regimes. This would of course hardly be a panacea: questions of sufficient text data aside, it should be kept in mind that not all (or even most) practical regime conflicts emerge from inconsistent norm commitments recorded in the legal text.²²⁹ Indeed, as noted by Morin and others, “[b]latant legal conflicts [amongst regimes] remain rare and a certain degree of normative ambiguity preserves the unity of the international legal system.”²³⁰ Instead, many regime conflicts may manifest at the level of impacts or institutional policies,²³¹ as a result of real-world inter-institutional interaction, and such negative externalities are not latent in the texts, and therefore not discoverable with such tools.

However, while these uses seem to connote beneficial (or at least, integrative) outcomes of AI displacement on regime complexes, there are many scenarios under which AI tools could also serve as an *accelerator* of regime fragmentation and complexity. After all, as has been discussed, many AI tools could also be used to challenge or subvert international legal processes, in ways that could speed up contestation. Even if AI technology finds productive uptake by many states to facilitate negotiations, this can have unanticipated side effects. After all, as Crootof has noted, speeding up the development of new international legal obligations—by any means—may certainly be necessary to address urgent problems, or help States avoid or resolve disputes; however it also “facilitates legal fragmentation [because] [s]peeding up the development of international legal obligations has expanded both the activities they regulate and opportunities for conflict among them.”²³² That is not a reason to avoid such tools, but should induce some caution about ensuring appropriate scaffolding or guidelines regarding their global use.

²²⁶ SURABHI RANGANATHAN, STRATEGICALLY CREATED TREATY CONFLICTS AND THE POLITICS OF INTERNATIONAL LAW (2014), <https://www.cambridge.org/core/books/strategically-created-treaty-conflicts-and-the-politics-of-international-law/55EACC81A929FC19216E3380D2E9DF69> (last visited Jun 18, 2020).

²²⁷ Deeks, *supra* note 225 at 616. (“[a] state that is concerned that its negotiating partner may be attempting this move could use web scraping and topic analysis on its negotiating partner’s existing treaties to assess whether the partner is trying to play this game”).

²²⁸ *Id.* at 616.

²²⁹ Christian Kreuder-Sonnen & Michael Zürn, *After fragmentation: Norm collisions, interface conflicts, and conflict management*, 9 GLOB. CONST. 241–267 (2020).

²³⁰ Morin et al., *supra* note 134 at 2–3.

²³¹ Thomas Gehring & Sebastian Oberthür, *The Causal Mechanisms of Interaction between International Institutions*, 15 EUR. J. INT. RELAT. 125–156 (2009).

²³² Crootof, *supra* note 94 at 7.

7.3.3.3 *Destruction and Complexity: shifting preferences and institutional proliferation*

Amongst the leading states, perceptions of the technology's very high strategic stakes—whether or not these views are justified—may inhibit willingness to even come to the negotiating table in good faith. Moreover, AI's inherent definitional complexity and effects across many fields, might lend themselves naturally to strategies that seek to obstruct international regulatory action by dragging out debate at international fora. The dual-use nature of many ‘inputs’ of AI capability—computing hardware; training data; talent—might furthermore make international bans of certain applications difficult to enforce.

All of this could result in notable AI issues remaining unsolved, in spite of them featuring (or being raised) on the international agenda. The clear inability of the existing governance arrangement to address these problems would drive processes of erosion in their legitimacy. Moreover, for certain powerful actors, it might drive preference changes—desires to pursue other avenues of governance. In many other cases, this could drive ‘voice and representation’ goals. Either could contribute to processes of institutional proliferation and regime complexity.

Simultaneously, AI's use as tool could offer specific capabilities that result in legal *decline*. This could be because AI systems might offer strategic capabilities which shift interests; they chip away at the rationales for certain powerful states to engage fully in, or comply with, international law regimes. While they would be unlikely to completely erode support, AI applications in areas such as surveillance could arguably reduce states' (perceived) dependence on multilateral security regimes to ensure their security from terrorist threats. Such capabilities could increase actor's willingness to ‘defect’, erode their willingness to support multilateralism, and could as such result in *preference changes* and the resulting increase in the prominence of forum-shopping strategies. That could produce a harmful ‘non-regime-state’. In extreme cases, AI systems could enable new strategies—such as scalable computational propaganda—by which malicious actors could directly challenge the legitimacy of international law or its component regimes.²³³ AI has not created these particular problems, but they may make them worse.

Given the above, we have discussed how the use of AI can shape both the normative and conceptual coherence of the regime complex (through driving *development* or *displacement*), as well as by shifting the incentives, values, or behaviour of various actors within that regime complex (potentially driving *destruction*). From a regime complex perspective, then, the key takeaway is that the governance disruptive impacts of AI may compound (1) trends towards regime fragmentation as a result of increased legitimacy problems or preference shifts; and (2) the problematic interactions of regimes (by increasing the number of contact surfaces).

To be sure, the above discussion of trajectories remains speculative, given the early state of AI governance at present. All this is not to make strong predictions, but rather to identify a range of ways in which AI applications, through governance disruption, could intersect with processes of regime complexity. In aggregate, these current trends and drivers might loosely suggest that, barring major interventions or external shocks, AI governance may remain

²³³ As discussed in Matthijs M. Maas, *International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order*, 20 MELB. J. INT. LAW 29–56, 55 (2019). As well as in Chapter 5.5.2.

fragmented for the time being.²³⁴ However, this is certainly not a foregone conclusion, and it will be key to monitor developments both in the AI regime complex, as well as developments in AI's effects on and in international law, in order to identify in greater detail the overall arc of governance development.

7.4 Consequences: effects of regime complexity on AI governance

Having sketched distinct trajectories for the AI regime complex, and their potential determinants and drivers, the next critical question is about the expected consequences or effects of these trajectories for AI governance. Charting specific outcomes is an intrinsically hard problem. A regime complex is a complex system.²³⁵ According to complex systems theory, changes in one part will interact with others,²³⁶ making prediction difficult. Nonetheless, examining potential consequences of a fragmented or integrated regime complex is of importance in investigating whether—or in what issue areas of AI—either form of governance might be preferred or pursued, all else being equal. Moreover, beyond serving as a crucial consideration regarding what form of AI governance to pursue, it is also useful to grapple with the consequences of either trajectory. If a fragmented AI regime complex might be expected to create many functional problems, it is important to understand these. Conversely, even if one expects—or hopes—that a centralised AI governance regime might be established, it is still important to understand the distinct downside risks this could introduce.

As previously noted, there remains live disagreement amongst scholars over the functional consequences and the normative desirability of regime complexity.²³⁷ For instance, critics anticipate that increasing regime complexity will have negative outcomes, at least if it is left ‘unmanaged’,²³⁸ in the form of a fragmented and eroded form of international law, dysfunction, power inequalities, or because it introduces strategic vulnerabilities for exploitation by certain states. On the other side, proponents of regime complexity argue that such ‘polycentral’ governance can in fact be more flexible and problem-solving, promotes improved inter-institutional dynamics, or creates more inclusive or democratic decision-making processes amongst parties.

7.4.1 Should AI Governance be Centralised or Decentralised? Trade-offs from History

Given the above debates, what could be the consequences of different (integrated or fragmented) governance architectures for AI governance? What would be the legal, operational, and functional

²³⁴ Cihon, Maas, and Kemp, *supra* note 167 at 228; JELINEK, WALLACH, AND KERIMI, *supra* note 167 at 2, 4.

²³⁵ Karen J. Alter & Sophie Meunier, *The Politics of International Regime Complexity*, 7 PERSPECT. POLIT. 13–24, 21 (2009). (“[t]o think in terms of international regime complexity is to study interactive relationships and analyze how the whole shapes the pieces”).

²³⁶ Alter and Raustiala, *supra* note 160 at 333.

²³⁷ See earlier Chapter 6.3.

²³⁸ We will turn to the question of managing regime complexity to ensure ‘coherence’ shortly, in chapter 7.5.3.

results of fragmentation? Even if there are adverse consequences to a fragmented regime complex, might we expect an integrated, centralised regime to perform better instead?

In recent years, some have begun to argue in favour of ‘polycentric’ models of international, collaborative governance for AI.²³⁹ In our paper [IV],²⁴⁰ we approached this debate by examining the trade-offs that might face an AI regime complex, depending on whether it remains fragmented, or whether it were to be(come) more integrated and organised around some centralised international organisation for AI.²⁴¹

In doing so, we drew on indicative lessons from the history of multilateralism. What would the experiences and challenges of other international regime complexes suggest about the costs and benefits of pursuing a centralised governance approach for AI, relative to a more fragmented trajectory? To ground this comparison, we drew lessons from international regimes in three other domain areas: environment, trade, and security. Analytically, these regimes offer imperfect but valuable comparative cases through which to study the challenges and pitfalls of AI governance.²⁴² This is because these three governance domains, while certainly distinct in important ways, are arguably similar to AI governance across key dimensions. For instance, (1) *environmental* governance invokes complex scientific questions that require technical expertise, involves trans-boundary and trans-sector effects, and raises a need for anticipation of future impacts and trends. For its part, (2) *trade* governance spans across a breadth of individual industries and sectors, involve questions of standard-setting, and confronts diverse (North-South) inequalities. Finally, (3) *security* regimes (and particularly arms control) confront large geopolitical stakes and strategic interests, along with a recurring pressure to ‘modernize’ regimes (e.g. updating explicit export control lists, or re-interpreting provisions) in order to track changes in the technology. All of these factors are pertinent in the context of AI governance, raising the hope that some lessons can be transferred. Finally, the environmental-, trade-, and security regimes offer an analytically interesting mix of regime complex types, including particularly different degrees of legalisation and centralisation, from very centralised (trade) to highly fragmented (environment).

²³⁹ For instance, Feijoo et al. have recently argued that “a multi-stakeholder, multi-layer model [...] would be able to accommodate inter-governmental discussions and other existing or new initiatives from industry fora to existing, international institutions and civil society.” Claudio Feijóo et al., *Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy*, TELECOMMUN. POLICY 101988, 10 (2020). See also Han-Wei Liu & Ching-Fu Lin, *Artificial Intelligence and Global Trade Governance: A Pluralist Agenda*, 61 HARV. INT. LAW J. (2020), <https://papers.ssrn.com/abstract=3675505> (last visited Sep 26, 2020). (emphasizing the increasing normative relevance of global legal pluralism, and recommending changes to the WTO in its approach to AI).

²⁴⁰ Cihon, Maas, and Kemp, *supra* note 167.

²⁴¹ Of course, as Gómez-Mera and others note, technically speaking, the moment one introduces such a hierarchic relationship to a central authority, the ‘regime complex’ ceases to exist; Gómez-Mera, Morin, and Van De Graaf, *supra* note 159 at 146. This may be true, although only under definitions that make non-hierarchy a core criterion to the concept—see Alter and Meunier, *supra* note 235; Alter and Raustiala, *supra* note 160. However, my usage of the regime complex concept in this dissertation tracks closer to that by Orsini and others, which does not put such emphasis on non-hierarchical relations; Amandine Orsini, Jean-Frédéric Morin & Oran Young, *Regime Complexes: A Buzz, a Boom, or a Boost for Global Governance?*, 19 GLOB. GOV. REV. MULTILATERALISM INT. ORGAN. 27–39 (2013).

²⁴² The arguments in favour of this comparison are relatively implicit in Paper [IV], but are developed in greater detail in a forthcoming paper: Cihon, Maas, and Kemp, *supra* note 168.

On the basis of these explorations, we identified a series of considerations which, when taken together with the above-discussed, broader debates on the consequences of regime complexity, can frame discussions over the sufficiency of a fragmented regime complex for AI.²⁴³

7.4.2 Political power

The first consideration is around the *political power* that can be projected by an AI governance system.²⁴⁴ Regimes embody power in their authority over rules, norms, and domain knowledge beyond their member states. Generally speaking, the more integrated a regime complex is, the more authority over rules, norms and issue area knowledge can be distilled into fewer (potentially one) institutions.²⁴⁵ By contrast, a fragmented regime complex that is hampered by rule uncertainty, goal contradictions, or even outright conflictual dynamics or ‘turf wars’²⁴⁶ will not carry the same authority as a well-organised flagship institution.

As such, a centralised AI institution would be likely to hold greater sway over its member states (and other actors).²⁴⁷ Moreover, it can also be more influential in the face of other international organisations, strengthening the regime’s ability to head off attempts at ‘competitive regime creation’ by states.²⁴⁸ Furthermore, greater authority may allow a centralised regime to prevail in the face of outright norm or impact conflicts with other regime (complexes) in other areas. For instance, it is arguable that something like this can be seen in the ‘chilling’ effect which the WTO has managed to exert over the fragmented institutional landscape of international environmental law, which has led environmental treaties to self-censor to avoid treading on trade-related issues.²⁴⁹ However, this case also illustrates the sensitive trade-off in centralisation—demonstrating how the benefits of a centralised AI institution should be considered, not only with reference to its internal efficacy in governing its domain area, but also in terms of potential externalities (such as ‘chilling’ effects) across other regimes.

In general terms, considerations of authority might be expected to speak in favour of centralisation of AI governance. Prevailing uncertainty in a domain issue area (such as over technological trajectories or future preferences) has been associated with some increased centralisation in regimes in the past.²⁵⁰ As a high-profile technology with significant vested interests and growing strategic stakes, it is plausible that AI technology could become the subject

²⁴³ This following section broadly follows the thematic organization within Paper [IV]. Cihon, Maas, and Kemp, *supra* note 167.

²⁴⁴ *Id.* at 230.

²⁴⁵ Orsini, Morin, and Young, *supra* note 241 at 36–37.

²⁴⁶ Biermann et al., *supra* note 159.

²⁴⁷ Orsini, Morin, and Young, *supra* note 241 at 36–37.

²⁴⁸ Morse and Keohane, *supra* note 205. One example of this can be found in the 2009 creation, by a coalition of countries including Germany, Spain and Denmark, of the International Renewable Energy Agency (IRENA), after their dissatisfaction with the International Energy Agency. See Graaf, *supra* note 184.

²⁴⁹ Robyn Eckersley, *The Big Chill: The WTO and Multilateral Environmental Agreements*, 4 GLOB. ENVIRON. POLIT. 24–50 (2004). This latter example also illustrates one potential downside, which is the risk that a centralised regime for one issue area becomes so politically powerful, that it begins to exert effects across issue-area governance architectures, imposing negative externalities on regime complexes aimed at different areas. Such questions of regime complex interaction across different regimes are of course a key problem to be examined.

²⁵⁰ Barbara Koremenos, Charles Lipson & Duncan Snidal, *Rational Design: Looking Back to Move Forward*, 55 INT. ORGAN. 1051–1082 (2001). See Cihon, Maas, and Kemp, *supra* note 167 at 230.

of considerable political contestation. In such scenarios, AI would be a thorny object for (global) regulation, with the risk that a fragmented regime complex would not be able to marshal and rely on sufficient political power to address many challenges effectively.

Nonetheless, while the perspective of political power may at first sight seem to speak unambiguously in favour of a centralised regime, that reading may be too simple. The key question might not be ‘what regime complex organisation provides AI institutions with as much authority as possible?’, and rather ‘what organisation provides AI institutions with the necessary level of authority to address the issue under examination?’ In this light, it can be seen that focusing institutional power or authority may not be equally relevant in all contexts, or for all problem logics. On some issues, institutions might be able to sufficiently affect national policies even if they do not carry extreme weight. At the other pole, there might be certain issues so fundamental to state sovereignty that almost no amount of centralisation of global institutional authority—short of world government—would be sufficient to set the desired policies. Yet in between those poles, there can be strategically or politically sensitive issue areas where one might expect a threshold of ‘minimum viable institutional authority’, where one very authoritative institution might be able to carry more clout than a dozen moderately authoritative institutions.

Moreover, from the perspective of sociotechnical change, questions over the ‘appropriate’ level of institutional authority required may also turn on the specific problem logics of AI issues. For instance, global governance for contentious AI-driven ‘ethical challenges’ (e.g. data privacy) might require high levels of institutional political authority in order to set and maintain stakeholder consensus. The same might be the case for governance for AI-produced ‘structural shifts’ (e.g. AI cyberwarfare systems or conflict prediction algorithms), where institutions might need considerable authority in order to compel actors to comply with norms even in the face of countervailing strategic incentives or pressures. Conversely, for governing ‘security threats’ deriving from the potential misuse of AI, high levels of institutional power or authority may be less critical or more assured—given shared state interests in avoiding the proliferation of certain capabilities to non-state actors—with the result that broader coalitions of coordinating institutions could suffice.

7.4.3 Efficiency and participation

The second consideration revolves around *efficiency* and *participation* thresholds. This relates to some of the previously mentioned concerns over participation, and the risk that institutional proliferation may advantage powerful parties more than weaker ones, resulting in a democratic deficit. For instance, the proliferation of hundreds of overlapping multilateral environmental agreements has, in some readings, created a state of ‘treaty congestion’ which imposes various costs and barriers to participation on many parties, whether at the stage of negotiation, implementation or monitoring of outcomes.²⁵¹ Accordingly, fragmented regimes may force parties to spread resources and funding over many distinct institutions, limiting the ability

²⁵¹ Don Anton, “*Treaty Congestion*” in *International Environmental Law*, in ROUTLEDGE HANDBOOK OF INTERNATIONAL ENVIRONMENTAL LAW (Shawkat Alam et al. eds., 2012), <https://www.taylorfrancis.com/books/9780203093474> (last visited Oct 7, 2019).

of less well-resourced states or parties to participate fully.²⁵² As a result, some have argued that the fragmentation of international law tends to strengthen the hand of powerful actors, who are better able to navigate and set the agenda in many parallel institutions, or are better able to play these up against one another.²⁵³

Concerns over inclusive participation might be particularly astute in the context of AI governance, where a small number of powerful states also represent the leading AI developers. Moreover, there is also a strong private sector dynamic. If (as seems likely) tech companies are structurally better resourced to participate in many global fora than are NGOs, the existence of many parallel such fora may not pose as much as a barrier to them as to these civil society actors. This would potentially increase the risk of regulatory capture.

Conversely, a centralised regime for AI could ideally avoid some of those problems, and ensure adequate participation by a wide range of parties. Of course, it can be asked whether full ‘inclusion’ is always required or even net beneficial. There may be rare issue areas where full ‘openness’ and inclusion for governance on AI issues might not be beneficial,²⁵⁴ either because it could deter certain key state parties from engaging with that forum in the first place, or because of the proliferation risk of certain particularly dangerous AI capabilities would imply these should not (at least in great technical detail) be freely discussed in open fora;²⁵⁵ or because of the potential for certain predictive AI uses to drive adverse self-fulfilling structural effects.²⁵⁶ Nonetheless, these may be edge cases, and in general one might expect a strong presumption for both functional and legitimacy reasons towards greater openness and democratic participation being a generally desirable trait of the AI regime complex.²⁵⁷

Finally, while in general regime complex fragmentation might seem to impede efficiency and participation, it is also important to consider here how certain AI tools, if widely distributed, could mitigate or soften this trade-off, by improving the ability of less powerful actors to remain appraised of developments across many fora, and to participate fully even in a fragmented regime complex.²⁵⁸

²⁵² Morin et al., *supra* note 134 at 2.

²⁵³ Benvenisti and Downs, *supra* note 224.

²⁵⁴ Nick Bostrom, *Strategic Implications of Openness in AI Development*, GLOB. POLICY 135–148 (2017). For a general discussion, see also Jess Whittlestone & Aviv Ovadya, *The tension between openness and prudence in AI research*, ARXIV191001170 CS (2020), <http://arxiv.org/abs/1910.01170> (last visited Jan 22, 2020).

²⁵⁵ Shevlane and Dafoe, *supra* note 60. For the general argument of ‘information hazards’, see Nick Bostrom, *Information Hazards: A Typology of Potential Harms from Knowledge*, 10 REV. CONTEMP. PHILOS. 44–79 (2011).

²⁵⁶ See also Agnes Schim van der Loeff et al., *AI Ethics for Systemic Issues: A Structural Approach* (2019), <http://arxiv.org/abs/1911.03216> (last visited Jan 13, 2020). (discussing the risk that a food security prediction system could drive adverse unsustainable agricultural practices or market dynamics, if its predictions were shared without caution). See also the general discussion of ‘structural shifts’ in section 4.4.4.

²⁵⁷ Kunz, *supra* note 171.

²⁵⁸ Deeks, *supra* note 225 at 644–647, 650–651. See also the previous exploration of AI as a ‘shield’, ‘patch’, or ‘cure’ to fragmentation, under 7.3.3.2.

7.4.4 Slowness of establishment or take-off

A third consideration is about the relative *slowness* of establishing different types of regime architectures in response to an emerging issue.²⁵⁹ In general, this appears to speak against centralised regimes.

After all, the process of deliberately negotiating and establishing a new centralised international institution is often a very slow one,²⁶⁰ especially if it departs from a fragmented institutional setting.²⁶¹ For example, the Kyoto Protocol took three years of negotiations to create, and another eight to enter into force. Trade agreement negotiations suggest that such processes can be particularly slow the higher the participation requirements or stakes.²⁶² International law has been quick to respond to some technological changes (spaceflight), but has also been slow in many others (modern antipersonnel mines).²⁶³ Moreover, multilateral treaties can take an extensive time to negotiate, and progress can be slow when a few states perceive that they (may soon) hold large advantages or stakes in the technology's development or deployment.²⁶⁴ For instance, Colin Picker discusses how during the 1980's, the US's expectations that it was not far from deep seabed mining technologies played a role in the Reagan Administration's 1982 refusal to sign or ratify the UN Convention on the Law of the Sea (UNCLOS), which designated that deep sea bed (and its resources) as the common heritage of mankind.²⁶⁵ Likewise, treaty negotiation can be slow whenever there are information and expertise asymmetries between states and private industry players.²⁶⁶ Given that global AI development at present involves both such interstate stake asymmetry, and state-private sector information asymmetries around AI, it can be expected that such hurdles to comprehensive treaties could apply considerably.²⁶⁷

By contrast, the various components of a decentralised regime complexes might be set up at a much shorter time span. For instance, it has been argued that smaller governance initiatives that involve informal institutions with like-minded memberships might be more able to wrangle with the complexity and political sensitivity of 'transversal' issues such as AI.²⁶⁸ If 'fragmented' initiatives such as those by the OECD and the G7 are able to forge ahead with setting norms and achieving important AI governance goals in the near term, then this might be functional. In particular, it might be preferable to the pursuit of a full multilateral treaty or organisation if, in

²⁵⁹ Cihon, Maas, and Kemp, *supra* note 167 at 230–231.

²⁶⁰ *Id.* at 230.

²⁶¹ Which is to say, the pervasive setting facing almost any new governance initiative today, given the existing density and accretion of pre-existing governance initiatives. Alter and Raustiala, *supra* note 160 at 337.

²⁶² For instance, Under the GATT, negotiations for a 26% cut in tariffs between 19 countries took 8 months in 1947. The Uruguay round, beginning in 1986, took 91 months to achieve a tariff reduction of 38% between 125 parties. Cihon, Maas, and Kemp, *supra* note 167 at 230. Citing Will Martin & Patrick Messerlin, *Why is it so difficult? Trade liberalization under the Doha Agenda*, 23 OXF. REV. ECON. POLICY 347–366 (2007).

²⁶³ Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDozo LAW REV. 151–219, 184–185 (2001).

²⁶⁴ This has been identified as one of the barriers to international regulation of cyberwar. Mette Eilstrup-Sangiovanni, *Why the World Needs an International Cyberwar Convention*, 31 PHILOS. TECHNOL. 379–407 (2018).

²⁶⁵ Picker, *supra* note 263 at 13. (citing sources).

²⁶⁶ *Id.* at 187–194.

²⁶⁷ Cihon, Maas, and Kemp, *supra* note 167 at 230.

²⁶⁸ Morin et al., *supra* note 134 at 2–3.

the waiting for a ‘comprehensive’ treaty regime, we would leave urgent AI issue areas effectively ungoverned during a critical period.

7.4.5 Brittleness vs. adaptation

A fourth, related consideration relates to the relative degree of ‘brittleness’ of a centralised regime or integrated regime complex, once established.²⁶⁹ Whereas ‘slowness’ concerns the relatively long process of establishing many centralised institutions, ‘brittleness’ refers to the tendency of centralised institutions to be path-dependent and unable to react to changes in the problem landscape (especially in the form of AI governance disruption). Relative to a fragmented regime complex, a centralised institution faces greater risk from three sources: regulatory capture, unified failure, or lack of adaptability.

The first potential risk is *regulatory capture*. The very qualities that afford a centralised regime with political power and authority may also exacerbate the downside effects of regulatory capture if it were to occur. After all, even if centralised high-profile institutions are more resistant to regulatory capture than are smaller institutions, they are certainly not immune to it.²⁷⁰ Moreover, the risk of such capture (or even the mere appearance of it) might be unusually high in the context of AI governance, given the extraordinary resources and demonstrated political clout of tech companies.²⁷¹ This could considerably undercut the legitimacy or authority of the resulting institution.²⁷²

The second risk of centralisation could be what might be called ‘*unified failure*’. As discussed, given its greater political power, a centralised institution might certainly be in a better position to set global policies on AI. However, this also means that a failure to set the appropriate policies could have more far-reaching consequences for the regime as a whole than would be the case if such mistakes or sub-optimal decisions were made only by one smaller institution amongst many in a broader complex. The argument here is not even necessarily that smaller institutions necessarily come up with better policies than large ones,²⁷³ but rather that in a fragmented regime complex, bad policies might be more easily ‘compartmentalised’ or at least spotted and challenged.

Finally, in the third place there are issues around the ability of a centralised regime to adapt to new external challenges or changes in its issue landscape. Historically, international institutions can be notoriously path-dependent, experiencing difficulty at addressing new problems (or new manifestations of old problems) beyond their remit.²⁷⁴ Ironically, the very path-

²⁶⁹ To avoid confusion, it should be clarified that in Paper [IV], we discussed these two considerations—‘slowness’ and ‘brittleness’—together under one section. Cihon, Maas, and Kemp, *supra* note 167 at 230–231.

²⁷⁰ Cf. Abigail C. Deshman, *Horizontal Review between International Organizations: Why, How, and Who Cares about Corporate Regulatory Capture*, 22 EUR. J. INT. LAW 1089–1113 (2011). (discussing corporate influence on the WHO during the 2009 H1N1 pandemic). See also JENS MARTENS, CORPORATE INFLUENCE ON THE G20: THE CASE OF THE B20 AND TRANSNATIONAL BUSINESS NETWORKS (First edition ed. 2017), https://www.boell.de/sites/default/files/corporate_influence_on_the_g20.pdf.

²⁷¹ Paul Nemitz, *Constitutional democracy and technology in the age of artificial intelligence*, 376 PHIL TRANS R SOC A 20180089, 3 (2018).

²⁷² This point was indeed raised around the initial French-Canadian proposal for an ‘IPAI’. Nature Editors, *International AI ethics panel must be independent*, 572 NATURE 415–415 (2019).

²⁷³ Although this is also a position some take. See also Section 6.3.2 and 7.4.8.

²⁷⁴ See Baccaro and Mele, *supra* note 194 (studying the ILO’s difficulties in reforming participation and rulemaking processes in the 1990s).

dependency that ensures institutional stability and predictability, might also mean that a centralised AI institution might not adapt sufficiently rapidly to unanticipated new problems or outside stressors outside its mandate. In such cases, we have argued, “the regime might break before it bends”²⁷⁵

This is a particularly salient challenge for the prospects of centralised and multilateral treaties for AI governance, because, as the discussion of governance disruption demonstrated, legally disruptive shifts may be unusually common in the context of technology governance. This is because once a treaty is in force, it is hard to amend, given the need to get mutual consent from all partners. After all, when only some partners agree to modify or update a multilateral treaty, this would result in a fracturing of the treaty regime, and a potential erosion of its international pull. As Crootof notes, this prospect can even have chilling effects, since “some states might avoid clearly-needed improvements [to treaties] when there is not complete consensus to avoid undermining the treaty regime’s overall force.”²⁷⁶ This suggests that multilateral treaties may be particularly susceptible to early obsolescence as a result of technological change.²⁷⁷ By contrast, a decentralised regime complex might function better. Indeed, the plural institutional ecologies of regime complexes have been described as ultimately more adaptable over time as well as more flexible across issues.²⁷⁸

However, the point may not be about whether centralised institutions or decentralised regime complexes are more adaptable in the abstract. Rather, in practical terms our estimate of the relative resilience of centralised or fragmented AI regimes to governance disruption, might depend, in part, on the speed or rate at which we expect changes in AI system capability, dissemination, usage, and resulting disruption.

Specifically, key variables in this discussion may be the *rate* and *magnitude* of AI-driven sociotechnical changes which would shift the centre of gravity in the problem landscape beyond current institutional mandates,²⁷⁹ or which create changes that are so significantly legally disruptive that they cannot be accommodated by regular interpretation, and generate a need for legal developments or institutional innovation.²⁸⁰

If we expect this rate-of-change in AI shocks to be relatively low,²⁸¹ then the benefits of the ‘rigidity’ of centralised institutions or treaties (their ability to stand up to political pressure, and pre-empt forum shopping) may well outweigh the costs of their reduced ability to keep pace with occasional changes. Conversely, if we expect that the rate-of-change in AI capability

²⁷⁵ Cihon, Maas, and Kemp, *supra* note 167 at 230–231.

²⁷⁶ Crootof, *supra* note 51 at 109; See also Brian Israel, *Treaty Stasis*, 108 AJIL UNBOUND 63–69 (2014).

²⁷⁷ Crootof, *supra* note 51 at 109.

²⁷⁸ Gómez-Mera, Morin, and Van De Graaf, *supra* note 159 at 145. See also Robert O. Keohane & David G. Victor, *The Regime Complex for Climate Change*, 9 PERSPECT. POLIT. 7–23, 19 (2011).

²⁷⁹ Maas, *supra* note 148 at 139–141 (on how innovation in military AI systems could shift the ‘problem portfolio’). See also the discussion in Section 5.3.5.

²⁸⁰ See also Section 5.3.1.

²⁸¹ Or alternatively, if we accept that AI’s rate of disruptive change has been high in recent years, but believe that most truly legally disruptive capability improvements have by now been produced, so that further conceptual disruption (e.g. the development of swarming drone warfare) will be only modestly more disruptive compared to the more fundamental conceptual disruption (e.g. the development of machine autonomy in robotic systems) which international legal systems have already been reckoning with.

development, applications, and sociotechnical change to be high,²⁸² then this might be taken as an argument for accepting regime fragmentation as an acceptable or even operationally superior trajectory for AI governance, that is better able to rapidly adapt to the frequent appearance of new situations.²⁸³

7.4.6 Breadth vs. depth dilemma

A fifth consideration relates to the so-called '*breadth vs. depth dilemma*'.²⁸⁴ This relates to the insight that while many global challenges (particularly weakest-link public goods) may appear to require universal participation, such universality may be hard to achieve in a treaty or convention. Frequently, this pursuit of centralisation may face an overly high threshold that limits participation from key parties—or is forced to make substantive concessions in order to secure such buy-in. All multilateral agreements face a trade-off between having higher participation ('breadth') or stricter rules and greater ambition of commitments ('depth'). However, this dilemma could be particularly crippling for centralised institutions that are intended to be powerful and require strong commitments from states.²⁸⁵

Indeed, the pursuit of universality in treaties (as opposed to customary international law) can often prove not only an obstacle to the creation and ratification of a treaty, but might also erode the ambition and effectiveness of the eventual instrument that is ultimately adopted.²⁸⁶ For instance, the 2015 Paris Agreement on Climate Change was significantly watered down in order to secure the participation of the US: because negotiators anticipated difficulties in securing ratification through the US Senate, they opted for a 'pledge and review' arrangement with few strong legal obligations, ensuring the US could join through the approval of the executive.²⁸⁷ As such, the overall regime made concessions on the demands made of all parties in order to secure the inclusion of a key state which, in the event, proved fleeting.²⁸⁸

In this way, the breadth vs. depth dilemma is an especially pertinent challenge to centralised institutions. This is because one supposed benefit of a centralised body (relative to a fragmented regime complex) is that it would be able to project greater political authority, such

²⁸² Or higher, at least, than the rate at which comparably disruptive developments have shaken up treaties and institutions in other fields of governance such as trade or ecosystems. Given the rate of development and application of AI, this appears a relatively defensible assumption.

²⁸³ However, this argument in favour of regime complexity would be diminished if there were ways to design more centralised regimes to reckon more effectively with these disruptive changes. See also the discussion of strategies for resilience, in Section 7.5.2.

²⁸⁴ Cihon, Maas, and Kemp, *supra* note 167 at 321.

²⁸⁵ *Id.* at 231.

²⁸⁶ See also Nele Matz-Lück, *Framework Conventions as a Regulatory Tool*, 3 GOETTINGEN J. INT. LAW 439–458, 445 (2009). ("While universal acceptance of multilateral treaties that address issues of global concern seems a necessity to effectively achieve these instruments' objectives, striving for universality may also be a significant obstacle for effectiveness. In a world that is politically divided by the interests of the North and the South, universal legal regulation requires compromise that impedes substantive commitments by the parties. Often the choice is between many states but weak regulation or strong legal obligations but few participants").

²⁸⁷ Luke Kemp, *US-proofing the Paris Climate Agreement*, 17 CLIM. POLICY 86–101 (2017). See also Cihon, Maas, and Kemp, *supra* note 167 at 231.

²⁸⁸ By contrast, it has been argued that during the establishment of the International Criminal Court, when faced with a critical US Administration, states deliberately accepted the option of an independent court without US support over a weaker court backed by the US. Fehl, *supra* note 125.

that it might serve as a powerful anchor that ensures policy coordination and coherence while averting fragmentation in membership or forum-shopping. This dilemma suggests it is unlikely to have both.²⁸⁹ Yet this matters, because securing sufficient state buy-in may be a very thorny challenge for AI governance. Key actors that currently lead in AI development include states that are potentially sceptical of global regulation, or at best have aimed to exclude one another in some fora. Tellingly, the US was initially sceptical of the French-Canadian 2019 IPA proposal; when in early 2020 it chose after all to support the GPAI, it did so in large part in order to counterbalance the role of China.²⁹⁰ This demonstrates the likely difficulty of crafting universal governance instruments that would secure inclusion of all major actors.²⁹¹

By contrast, decentralisation might sometimes allow certain states to undertake at least some much-needed regulatory efforts, when they might be unwilling to sign up to a comprehensive and deep package. For instance, some have suggested that the US-led Asia-Pacific Partnership on Clean Development and Climate may have helped climate governance, insofar as it secured at least some non-binding commitments from actors not bound by the Kyoto Protocol.²⁹² As such, decentralised regime complexes might at times be more suitable than pursuing an all-or-nothing centralised treaty regime, because the latter might be forced to default to a toothless ‘lowest-common-denominator’ set of legal commitments, which might be insufficient.

As with other considerations discussed here, the ‘breadth vs. depth’ dilemma raises interesting trade-offs for AI governance, which may not lend themselves to blanket recommendations, but rather depend sensitively on the specific governance rationale, texture and problem logics of the AI challenge that is being addressed. For instance, if successfully addressing a given AI issue depends more significantly on a regime’s *depth* than on its *breadth* (if, for instance, what is required is for a small number of key states to forego deployment of any AI systems linked to nuclear command and control networks; or at least to subject them to stringent safety and security testing standards), then a centralised approach could be functional because it would only have to reckon with a small number of principals (in this case, only the nuclear-weapon states).²⁹³

Likewise, if successfully addressing an AI issue appears to depend more on a regime’s *breadth* of participation than on its *depth* of requirement (for instance, if what is required is for every actor to implement certain software protocols which inhibit the dissemination or

²⁸⁹ Cihon, Maas, and Kemp, *supra* note 167 at 231.

²⁹⁰ Michael Kratsios, *Artificial Intelligence Can Serve Democracy*, WALL STREET JOURNAL, May 27, 2020, <https://www.wsj.com/articles/artificial-intelligence-can-serve-democracy-11590618319> (last visited Jun 10, 2020).

²⁹¹ Although in a promising step, China has adopted the 2019 ‘Osaka Track’ G20 Ministerial Statement on Trade and Digital Economy, which endorses and adopts the OECD’s AI Principles. G20, *supra* note 169. See also Figure 7.2 in Section 7.2.

²⁹² Fariborz Zelli, *The fragmentation of the global climate governance architecture*, 2 WILEY INTERDISCIP. REV. CLIM. CHANGE 255–270, 259–260 (2011).

²⁹³ This could work on the basis of a ‘minilateral’ approach. Robyn Eckersley, *Moving Forward in the Climate Negotiations: Multilateralism or Minilateralism?*, 12 GLOB. ENVIRON. POLIT. 24–42 (2012). See also discussions on ‘critical mass approaches’ to multilateralism. Luke Kemp, *Critical Mass Governance: Addressing US Participation in Environmental Multilateralism*, May, 2016. See also the discussion of ‘incrementalism, minilateralism and experimentalism’ for AI governance in Liu and Lin, *supra* note 239 at 342–343.

unauthorised use of automated hacking systems with potential criminal uses²⁹⁴) this might, again, be pursued through a centralised regime, since broad (if not universal) buy-in with merely modest (but minimum) standards would be required. Such a challenge could be jeopardised by strategies of forum-shopping, or by inadequate implementation of the required policies by states with little resources, resulting in the usual challenges of securing the ‘weakest link’ goods.²⁹⁵ As such, breadth and inclusion would be needed, suggesting the pursuit of some broad global instruments even if these would have to make substantive concessions.

Paradoxically, however, it is when an AI issue would require *both* depth and breadth for its successful mitigation, that a centralised governance regime might face particular challenges in gaining sufficient consensus from a heterogeneous community of actors. This could be the case for many *ethical challenges*, such as agreements to generally restrict the use of ‘predictive’ algorithms for predicting the behaviour of individuals, populations or states.²⁹⁶ It could include global regimes for ‘structural’ challenges, such as agreements to prevent the development or proliferation of powerful AI capabilities for use in cyberwarfare. It could be the case for global agreement on reducing *safety risks*, such as agreements to set costly minimum standards on ensuring the safety and reliability of AI systems.²⁹⁷

Finally, regimes that combine high depth and breadth requirements might also include cases where the aim is to implement technology-neutral provisions for certain AI capabilities, in order to render these treaties more resilient to later Governance disruption.²⁹⁸ Such neutrality might be key in order to avert or at least delay the eventual obsolescence of treaty regimes.²⁹⁹ However, a challenge here is that states can be generally hesitant to sign up to technologically neutral treaties, because even if the commitments appear light at present, their technological neutrality means these actors may, in the future, find themselves bound to foregoing capabilities which they had not anticipated at the stage of signing on to the treaty.³⁰⁰

In such cases, rather than struggling to secure both depth and breadth in a centralised treaty, it might be worth to pursue ‘second-best’ strategies in the near-term in a fragmented

²⁹⁴ Admittedly, the history of cybersecurity should not leave one optimistic about the viability of solving foundational problems of (malicious) software dissemination or unauthorized use through a simple ‘protocol’, so this is merely illustrative.

²⁹⁵ Anne van Aaken, *Is International Law Conducive To Preventing Looming Disasters?*, 7 GLOB. POLICY 81–96 (2016); Todd Sandler, *Strategic Aspects of Difficult Global Challenges*, 7 GLOB. POLICY 33–44 (2016); BARRETT, *supra* note 123.

²⁹⁶ Deeks, Lubell, and Murray, *supra* note 36; Ashley Deeks, *Predicting Enemies*, 104 VA. LAW REV. 1529–1593 (2018).

²⁹⁷ Amanda Askell, Miles Brundage & Gillian Hadfield, *The Role of Cooperation in Responsible AI Development* 23 (2019).

²⁹⁸ Crootof, *supra* note 51 at 121–123. (though noting also some of the drawbacks of technology-neutral rules, such as the fact that they may not constrain later interpreters as much).

²⁹⁹ As discussed in Chapter 6. For example, some have argued that global restrictions on LAWS should not be grounded in these systems’ presumed inability to discriminate between civilians and combatants (e.g. an assumption that they categorically violate the IHL principle of distinction), which renders such restrictions vulnerable to technological progress—but should rather be grounded in the more fundamental and technologically neutral principles of human dignity. Rosert and Sauer, *supra* note 55.

³⁰⁰ Crootof, *supra* note 94 at 17. (noting how “well-meaning attempts to extend a prohibition on one class of weapons to another might actually undermine the power of weapon bans regimes generally: if such treaties are viewed as overly flexible, States wary of incurring unexpected and unwanted future legal obligations will be more reluctant to join them.”).

regime complex. One approach here could be to accept the benefits of fragmentation, splitting off important provisions into various parallel instruments or regimes, to ensure at least some participation by at least most relevant actors. For instance, for high-stakes military applications of AI, such a strategy could involve parallel treaties—for instance, relating to LAWS testing standards; liability and responsibility; and setting limits to operational usage or range—in order to allow states to ratify and move forward on at least some of those options.³⁰¹

Such a strategy certainly has downside risks, but in optimistic scenarios, grounding initial governance initiative in smaller and informal coalitions of like-minded states could avert sustained governance gridlock, and fast-track much needed policies and norms. For example, in the field of ‘human security’, the 1998 Lysoen Declaration effort initially began life as a bilateral Canadian-Norwegian initiative. However, as it expanded to a coalition of 11 other like-minded and highly committed states along with several NGOs, the resulting ‘Human Security Network’ achieved a significant and outsized global impact, including the Ottawa Treaty ban on anti-personnel mines; the Rome Treaty establishing the International Criminal Court; the Kimberley process aimed at inhibiting the flow of conflict diamonds, and landmark Security Council resolutions on Children and Armed Conflict and Women, Peace and Security.³⁰² Such cases show how AI governance initiatives with a limited membership could set the ground for a ‘critical-mass’ approach that starts from a small group, and which is subsequently able to expand to achieve wider impact.³⁰³

Of course, this strategy also has downsides. While a small group of like-minded actors might be able to settle on more rigorous or stringent terms than might be achieved rapidly in a global or encompassing forum, there are legitimacy challenges if non-parties are only later invited to the table. For instance, Martina Kunz has noted how the 2001 Budapest Cybercrime Convention was initially elaborated under the auspices of the Council of Europe along with only a selected few non-members; all other states had to wait for the convention to enter into force, and could only then be invited for accession upon unanimous agreement by its existing members. This non-inclusive process, she argues, may have contributed to that Convention’s stunted growth.³⁰⁴

7.4.7 Forum shopping and strategic vulnerability

A sixth consideration relates to the potential for *forum shopping*. Whereas a centralised regime could serve as a focal point for AI governance, decentralised or fragmented regime

³⁰¹ John Frank Weaver, *Autonomous Weapons and International Law: We Need These Three International Treaties to Govern “Killer Robots,”* SLATE MAGAZINE, 2014, <https://slate.com/technology/2014/12/autonomous-weapons-and-international-law-we-need-these-three-treaties-to-govern-killer-robots.html> (last visited Apr 16, 2019). As discussed in Cihon, Maas, and Kemp, *supra* note 167 at 231.

³⁰² The relevance of this example to AI governance (and GPAI) is noted by Arindrajit Basu & Justin Sherman, *Two New Democratic Coalitions on 5G and AI Technologies*, LAWFARE (2020), <https://www.lawfareblog.com/two-new-democratic-coalitions-5g-and-ai-technologies> (last visited Aug 27, 2020). See also ANDREW F COOPER, *Stretching the Model of “Coalitions of the Willing”* 31 (2005), https://www.cigionline.org/sites/default/files/working_paper_1_-_stretching_the_model_of_.pdf; Allan Rock, *The Human Security Network, Fifteen Years On*, CENTRE FOR INTERNATIONAL POLICY STUDIES (2013), <https://www.cips-cepi.ca/2013/05/21/the-human-security-network-fifteen-years-on/> (last visited Aug 27, 2020).

³⁰³ Kemp, *supra* note 293; Morin et al., *supra* note 134.

³⁰⁴ Kunz, *supra* note 171 at 3.

complexes create far greater room for strategic manoeuvring by many governance actors. This can take different forms. For instance, there has been extensive scholarship exploring how a fragmented regime complex opens up avenues for states to pursue diverse strategies of ‘forum shopping’, ‘regime shifting’ or ‘competitive regime creation’ in order to induce ‘strategic inconsistency’ between treaty instruments, or even foster direct competition between organisations.³⁰⁵ States may do so in order to maximize influence, frustrate (or deter) certain institutional initiatives, appease domestic pressure, or placate constituencies by shifting discussion to a weaker governance forum.³⁰⁶ Forum-shopping has been extensive in areas such as intellectual property rights in trade, as well as the environment.³⁰⁷ Prima facie, it might pose a problem for AI governance, especially if key private actors choose to participate in- or withdraw from selected fora in order to force certain policies.³⁰⁸

On the other hand, under some circumstances, forum-shopping can also be used by weaker states or non-state actors to (threaten to) bypass a stalled forum in order to achieve greater success elsewhere.³⁰⁹ Indeed, in the context of AI, something like this may well occur in the area of international regulation of LAWS, where some activists, frustrated with the slow process of the CCW Group of Governmental Experts, have increasingly threatened to shift to another forum, as they historically did with the Ottawa Treaty ban on landmines.³¹⁰ Nonetheless, in general, the risk of forum shopping demonstrates a likely downside of high fragmentation in the AI regime, either supporting the establishment of more centralised institutions or, more modestly, highlighting the need for distinct strategies or interventions in order to ‘manage’ the regime complex and avert or mitigate such behaviour.³¹¹

7.4.8 Policy coordination

Finally, a consideration concerns the relative ability of a centralised regime or fragmented regime complex to achieve effective *policy coordination*.³¹² Given the generality of many AI capabilities, their likely cross-sector applications, and their cross-application ‘problem logics’ for governance responses, it appears that there is a high premium on ensuring at least a minimum of coherence and continuity in the principles and standards that are applied to AI—and to avoid conflicts or inconsistencies.

³⁰⁵ Thomas Gehring & Benjamin Faude, *The Dynamics of Regime Complexes: Microfoundations and Systemic Effects*, 19 GLOB. GOV. 119–130, 126 (2013); JOHN BRAITHWAITE & PETER DRAHOS, GLOBAL BUSINESS REGULATION (2000). Morse and Keohane, *supra* note 205; RANGANATHAN, *supra* note 226.

³⁰⁶ Helfer, *supra* note 204; Saadia M. Pekkanen, Mireya Solís & Saori N. Katada, *Trading Gains for Control: International Trade Forums and Japanese Economic Diplomacy*, 51 INT. STUD. Q. 945–970 (2007).

³⁰⁷ Helfer, *supra* note 204.

³⁰⁸ For one example, see Baidu’s 2020 withdrawal from the private-sector initiative Partnership on AI. Knight, *supra* note 202.

³⁰⁹ JOSEPH JUPILLE, WALTER MATTLI & DUNCAN SNIDAL, INSTITUTIONAL CHOICE AND GLOBAL COMMERCE (2013); Laura Gómez-Mera, *Regime complexity and global governance: The case of trafficking in persons*:, 22 EUR. J. INT. RELAT. 566–595 (2016).

³¹⁰ Delcker, *supra* note 209.

³¹¹ Such strategies are discussed in greater detail in Section 7.5.

³¹² Cihon, Maas, and Kemp, *supra* note 167 at 232.

On its face, a centralised institution or relatively integrated regime complex might be expected to better facilitate such policy coordination, by simple virtue of acting as an authoritative focal point that can reduce the occurrence of conflicting mandates, and avert many conflictual regime interactions in the areas of norms, goals, impacts or institutional relations.³¹³

Nonetheless, it is not the case that fragmented regimes are entirely unable to secure policy coordination. Indeed, this is an active area of work in regime complex theory, with scholars studying various models or cases of gradual ‘co-adjustment’, consolidation or convergence of regime complexes over time.³¹⁴ In some accounts, this process results in (1) gradual organic integration of the regime complex. Others anticipate that (2) a fragmented or ‘polycentric’ regime complex can effectively self-organize.

In the first scenario of (1) *gradual integration*, the argument here is that along with drivers of fragmentation, there are also ‘centripetal’ drivers of regime integration.³¹⁵ This would imply that initial fragmentation of a regime complex is not a problem, because it will over time ‘mature’ into a well-coordinated and even integrated regime—one potentially more functional or legitimate than a world organisation grafted from the top. To be sure, such bottom-up coordination and centralisation of regime complexes has occurred; one example is found in the gradual coalescence of the General Agreement on Tariffs and Trade (GATT) and numerous regional, bilateral and sectoral trade treaties into the WTO.³¹⁶ However, while in this case, ‘organic’ integration of the trade regime complex eventually occurred, it involved decades of forum shopping and inaction.³¹⁷

Alternatively, in the second scenario of (2) eventual *self-organisation* of a polycentric regime complex, scholars anticipate that the system can adapt, in the form of a gradual but emergent ‘division of labour’ between institutions.³¹⁸ Such an adaptation of the inter-organisational relationships might be tacit, or it may involve the formal ‘outsourcing’ of certain institutional functions through explicit and formal ‘contracts’ between international organisations.³¹⁹ Such adaptation may either be spurred by internal regime complex conflicts, or by outside shocks.³²⁰

The argument here is that a fragmented regime complex could provide effective AI governance, even (or especially) if it never does integrate. To be certain, there are recent signs of

³¹³ As discussed in 7.2.3.

³¹⁴ Galbreath and Sauerteig, *supra* note 181. See also Jean-Frédéric Morin & Amandine Orsini, *Regime Complexity and Policy Coherency: Introducing a Co-adjustments Model*, 19 GLOB. GOV. 41–51, 42 (2013).

³¹⁵ Kim, *supra* note 159 at 11.

³¹⁶ Cihon, Maas, and Kemp, *supra* note 167 at 232.

³¹⁷ *Id.* at 232.

³¹⁸ Galbreath and Sauerteig, *supra* note 181 at 85–86; Gehring and Faude, *supra* note 305.

³¹⁹ Galbreath and Sauerteig, *supra* note 181 at 90–91. (“a formal arrangement between international organizations. [...] where one international organization seeks to use another for the resources and functions that are needed to address an issue”). They distinguish ‘contract’ from two other (unintentional) outcomes of regime complex dynamics, ‘competition’ and ‘convergence’. *Id.* at 88–90. Note, these terms should be distinguished from the typology which we use in Paper [IV], which focuses instead on potential trends in ‘conflict’, ‘coordination’, or ‘catalyst’ within the AI regime complex. Cihon, Maas, and Kemp, *supra* note 167 at 233.

³²⁰ For instance, Yu and Xue argue regime complexes can experience ‘punctuated equilibria’—long periods of stability, shifted with unstable periods—arguing that the “evolution of regime complex is shaped by interactions between governance issues outside the regime complex, and a combination of actor power and institutional logics”. Yu and Xue, *supra* note 214.

nascent cross-institutional coordination on principles across the AI regime complex. This is illustrated both by institutional coordination at the UN level and within various treaty regimes,³²¹ as well as by the fact that the OECD recommendations on AI explicitly informed both the G20's (June 2019) non-binding AI principles, as well as the Working Groups within the Global Partnership on AI (June 2020). Nonetheless, the question over whether such self-organisation will prove functional in time, or whether a centralised AI regime would still serve better at enabling the required level of policy coordination, ultimately turns on two considerations: one of *timelines*; the other on *required level of cooperation*.

In terms of *timelines*, it is worth considering the degree to which these timelines of regime complex evolution match (that is, are 'synchronised' to) the plausible timelines characterizing the evolution of the AI problem landscape. While it should not be attempted to predict AI developments or the resulting sociotechnical changes, the underlying point is that even where a fragmented regime complex manages to eventually self-organize, the timeframes for doing so appear relatively long—in some cases, one or two decades.³²²

This raises the question of whether we have that time—or what it would cost us in the waiting. If one assumes that a fragmented regime state does not enable sufficient policy coordination for AI governance, then an extended persistence of that state might at best be a waste (involving substantial opportunity costs or duplication of effort), and at worst could hold back global efforts to effectively or coherently govern certain key AI issue areas at a particularly sensitive moment in their development.³²³ Moreover, the regime complex 'lifecycle' may be longer than the timeframes involved in the development, deployment and proliferation of disruptive new AI systems.³²⁴ If new AI developments, uses, or sociotechnical changes routinely short-circuit dynamics of regime adjustment or integration, this would suggest that counting on the regime complex to eventually catch up and settle may not prove a durable strategy.

Finally, the comparison of centralised and decentralised AI regime complexes also turns on the underlying question of *what level of policy coordination* will in fact be required to effectively manage AI issues. For instance, if we expect that the successful governance of AI across many diverse sectors will, in some or all cases, require the application of general principles or invariant rules (e.g. regarding explainability, meaningful human control, or robustness to adversarial input) across diverse application areas (e.g. from autonomous vehicles to autonomous weapons), then achieving sufficient policy coordination across the regime complex would appear especially important, potentially supporting centralisation in AI governance.³²⁵

³²¹ Kunz and Ó hÉigearthaigh, *supra* note 162. Kunz, *supra* note 171. Garcia, *supra* note 162.

³²² Cihon, Maas, and Kemp, *supra* note 167 at 232.

³²³ Specifically, a risk here is that by the time the regime complex has worked through initial conflictual dynamics to achieve more synergistic arrangements, we will, per the Collingridge Dilemma, have traded in the 'information problem' for a 'power problem', inhibiting effective governance.

³²⁴ However, as noted (in section 7.4.4), more traditional instruments of 'integrated' international law—such as multilateral treaties, or the formulation of international customary law—are not necessarily more rapid, and can likewise require many years of negotiation or accretion, respectively. Picker, *supra* note 263; Crootof, *supra* note 51.

³²⁵ This reading is advocated by some scholars. TURNER, *supra* note 182 at 213–221. It may also be weakly implied by the sociotechnical change lens on technology regulation, as this approach tends to focus less on technological boundaries, and more to focus on societal governance rationales that may cut across domains or application areas.

On the other hand, there are certainly scholars who argue that AI governance may not need general rules, or that sector-specific regulations for applications will prove adequate at addressing most new AI problems most of the time.³²⁶ If that is the case, then one might expect outright inter-regime norm conflicts to be either rare or practically irrelevant (that is, remain latent). If one puts a particularly high emphasis on local values and cultural contexts in addressing AI, this might even speak against pushing strong harmonisation initiatives.³²⁷ In such cases, ensuring policy coordination across the larger AI regime complex may not be critical from an instrumental perspective.³²⁸ This would suggest that a fragmented AI regime complex's (assumed) impoverished ability to deliver adequate policy coordination rapidly (or at all) is not a decisive obstacle.

Nonetheless, there are some general reasons to expect that coordination could be valuable, even if regulation itself remains focused on individual AI application areas. In the first place, norm conflicts that remain mostly latent or theoretical at the present level of technological development, may be activated or spill over as a result of technological progress, potentially creating situations of legal ambiguity that are potentially hard to resolve, risking forms of governance erosion.

More practically, effective governance cooperation amongst regimes focused on apparently distinct applications may be especially valuable in the context of AI. This is because while AI is certainly an accessible technology developed by a broad and diverse ecology of actors, it is also true that the lion's share of high-impact and high-stakes AI applications across diverse sectors—the ones that may affect many millions of people on a daily basis—are currently produced by a relatively small set of extremely large private tech actors, most of which operate a similar business models of capturing and selling attention.³²⁹ Moreover, in many cases, the ethical challenges that arise around these actors' AI systems can be at least partially traced back to the problems that emerge around these businesses—or their algorithms—optimizing for a too-narrow set of metrics.³³⁰

Finally, from the perspective of governance levers, these large actors moreover depend on common computing hardware resources provided by a relatively globally concentrated compute

³²⁶ This is notably the position taken by Kunz and Ó hÉigearthaigh, *supra* note 162. As well as, implicitly, by much domain-specific legal scholarship on AI.

³²⁷ See for instance the argument made in Liu and Lin, *supra* note 239.

³²⁸ Though it might be from a global constitutionalist perspective, or a 'legal coherentist' perspective that emphasizes the consistency of a legal system. Roger Brownsword, *Law Disrupted, Law Re-Imagined, Law Re-Invented*, TECHNOL. REGUL. 10–30, 14 (2019).

³²⁹ See SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (1 edition ed. 2019).

³³⁰ Rachel Thomas & David Uminsky, *The Problem with Metrics is a Fundamental Problem for AI*, ARXIV200208512 Cs (2020), <http://arxiv.org/abs/2002.08512> (last visited Aug 27, 2020). See also Rebekah Overdorf et al., *POTs: Protective Optimization Technologies*, ARXIV180602711 Cs (2018), <http://arxiv.org/abs/1806.02711> (last visited May 21, 2019); See also generally David Manheim & Scott Garrabrant, *Categorizing Variants of Goodhart's Law*, ARXIV180304585 Cs Q-FIN STAT (2018), <http://arxiv.org/abs/1803.04585> (last visited Jan 29, 2019). For a more optimistic sketch of how societies could go beyond single-valued optimization systems to explore new and more pluralist applications of AI grounded in a form of 'robust and adaptive decision-making' (RDM), see Edward Parson et al., *Could AI drive transformative social progress? What would this require?*, AI PULSE, 6 (2019), <https://aipulse.org/could-ai-drive-transformative-social-progress-what-would-this-require/> (last visited Sep 28, 2019).

production stack, along with a set of widely used machine learning tools.³³¹ In sum, these constitute a potential common set of regulatees, problem logics, and policy levers which are all relevant to addressing ‘local’ AI problems facing widely distinct regimes or institutions. That implies that ensuring a certain degree of policy coordination would matter to AI governance. It does not mean that only a centralised institution will be adequate, but it does suggest that in its absence, a fragmented AI regime complex should develop and implement distinct strategies to ensure such coordination, as well as to face the broader pressures of change.

7.5 Strategies: AI regime efficacy, resilience, and coherence

In the previous two sections we have considered the potential factors that could affect the *evolution* (7.3) of the AI regime complex towards fragmentation or integration, and we have discussed some of the variable *consequences* (7.4) of either form of regime complex organisation. Nonetheless, it is also clear that, whichever governance trajectory one expects (or prefers), steps can be taken to improve the functioning of the AI governance architecture. As such, we now will explore potential *strategies* for organizing and aligning AI governance.

Which AI governance? My aim in this section is not to advocate categorically in favour of one or the other form of AI regime organisation (e.g. a centralised international AI institution or a decentralised regime complex), but rather to explore strategies applicable (or agnostic) to either form of regime architecture. The hope is that, whether the regime complex is integrated or fragmented, actors engaged in AI governance might gain practical insights from engaging the three conceptual lenses explored in the previous chapters.

What *kind* of strategies? I argue that these three lenses suggest new choices for AI governance at the levels of (1) overall *conceptual approach* (the attitude or mind-set through which we approach or analyse AI governance problems); (2) *instrument choice* (what types of regulatory tools we might prefer or disprefer for AI regimes); and (3) *instrument design* (how we might reconfigure or re-tailor these governance instruments (both old and new) in order to be more fit. These should not be meant as definitive recommendations, but as rough sketches of suggested avenues, in need of much further examination.

Strategies to what *end*? This is a more challenging and controversial question, given how it turns on core debates over what international law or global governance are, or should be, for.³³² My point is not to adjudicate on specific values or goals for AI regimes, given also that regulatory values or goals may change over time. Rather, the implications of the three perspectives on change will be explored relatively narrowly, focusing on potential strategies to improve the *efficacy*, *resilience* and *coherence* of an AI governance system (whether a centralised regime or fragmented regime complex).

By *efficacy* is meant, loosely, the ability of an AI governance system to identify, track, respond to, and effectively (by its own lights) address whichever AI-driven sociotechnical changes create its governance rationale.

³³¹ I thank Luke Kemp for this point. See also Section 2.2.3.2. (on barriers and levers for governance).

³³² Cf. Nico Krisch, *The Decay of Consent: International Law in an Age of Global Public Goods*, 108 AM. J. INT. LAW 1–40 (2014). See also section 2.2.2 (on rationales for governance to secure AI public goods).

By *resilience* is meant, loosely, the ability of an AI governance system to absorb, respond to, or even harness the effects of AI-driven governance disruption, in order to carry out development where it is needed, harness or manage the effects of displacement, and avoid the outcomes of governance destruction (through erosion or –decline).

By *coherence* is meant, loosely, the ability of an AI governance system to manage changes and stresses in the regime architecture, specifically in terms of inter-regime or inter-institutional conflicts over norms, goals, or outcomes.

These are admittedly ‘thin’ functional principles, and they are certainly not the only desiderata we might expect, desire or pursue in AI governance. There is a much richer and broader debate to be had about the substantive principles, purposes or values which AI governance should embody. For a regime to be merely efficacious, resilient, and coherent is not in itself sufficient.

Nonetheless, these three modest goals can be a serviceable core: they can be derived from—or reconciled with existing scholarship on technology and law, legal disruption, or regime complexity, respectively. Moreover, all three goals are instrumentally valuable across a broad range of governance contexts, and for addressing diverse AI problems. Indeed, any AI governance system lacking one or two of these qualities would be hard-pressed to remain ‘fit’ in the face of change.³³³ A regime that lacked all three might well prove more of an obstacle than a tool to global efforts to govern AI. Given this, diverse parties and stakeholders who disagree on other elements or norms around AI governance might nonetheless be able to reach pragmatic agreement or an overlapping consensus on these goals for AI regimes.³³⁴ These are not the only strategies AI governance needs—but they may be the minimum strategies it needs to deal with change. As such, a number of potential strategies and shifts are sketched (see also Table 7.3, in section 7.5.4).

³³³ Roger Brownsword, Eloise Scotford and Karen Yeung have discussed the challenge of ensuring a regulatory environment is adequate or ‘fit for purpose’, which in their view involves “an audit of the regulatory environment that invites a review of: (i) the adequacy of the ‘fit’ or ‘connection’ between the regulatory provisions and the target technologies; (ii) the effectiveness of the regulatory regime in achieving its purposes; (iii) the ‘acceptability’ and ‘legitimacy’ of the means, institutional forms, and practices used to achieve those purposes; (iv) the ‘acceptability’ and ‘legitimacy’ of the purposes themselves; (v) the ‘acceptability’ and ‘legitimacy’ of the processes used to arrive at those purposes; and (vi) the ‘legitimacy’ or ‘acceptability’ of the way in which those purposes and other purposes which a society considers valuable and worth pursuing are prioritized and traded-off against each other.” Roger Brownsword, Eloise Scotford & Karen Yeung, *Law, Regulation, and Technology: The Field, Frame, and Focal Questions*, 1 in THE OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY, 11 (Roger Brownsword, Eloise Scotford, & Karen Yeung eds., 2017), <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-1> (last visited Jan 3, 2019).

³³⁴ This need not even be agreement on deep underlying values, but might rather be a form of ‘overlapping consensus’, or an ‘incompletely theorized agreement’. See John Rawls, *The Idea of an Overlapping Consensus*, 7 OXF. J. LEG. STUD. 1–25 (1987); Cass R. Sunstein, *Incompletely Theorized Agreements*, 108 HARV. LAW REV. 1733–1772 (1995). Some have suggested such consensus or agreement also underpins landmark international legal achievements such as human rights. Charles Taylor, *Conditions of an Unforced Consensus on Human Rights* (1996), <https://www.iij.org/wp-content/uploads/2016/08/Taylor-Conditions-of-an-Unforced-Consensus-on-Human-Rights-1996.pdf>; Cass Sunstein, *Holberg Prize 2018, Acceptance Speech* (2018), <https://www.holbergprisen.uib.no/en/cass-sunsteins-acceptance-speech> (last visited Oct 15, 2018).

7.5.1 AI's Sociotechnical Change and Strategies for Efficacy

As noted previously, the sociotechnical change perspective should not be taken as providing blueprints for AI regimes. Nonetheless, it may offer a valuable complementary thinking tool, and its use might have suggest new directions not only for the underpinning regulatory attitude or approach, but also in terms of instrument selection and instrument design. These strategies can help ensure the *efficacy* of an AI governance—that is, its ability to identify, track, respond to, and effectively (by its own lights) address whichever AI-driven sociotechnical changes create its governance rationale.

7.5.1.1 *Conceptual Approach (Sociotechnical Change)*

The sociotechnical change lens suggests that global governance for AI will have to adapt and adopt particular conceptual approaches to reckon with the way this general-purpose-technology can drive sociotechnical change across diverse domains and implicating diverse problem logics.

(1) [***Govern for sociotechnical change, not technology***]

Rather than immediately focus in on isolated AI applications that are within the remit of certain international legal domains or regime mandates, it can be more productive for AI governance scholars to adopt the lens of sociotechnical change. This need not imply that AI governance regimes must be unmoored from older legal domains. Nonetheless, considering the nature of AI sociotechnical change in this perspective can facilitate in exploring broad questions of regulatory need, timing, -viability, and possible inter-regime externalities such as synergies or conflicts.

Firstly, this entails considering explicitly the AI-enabled capability-change in question: what is the *magnitude* of the change (does it simply lower the threshold to existing behaviours, or does it create completely new behaviours)? What is its *nature* (is the novelty conceptual, operational or political)? What is its *salience*?

Secondly, a next set of questions should focus in on what (if any) are the *governance rationales* this sociotechnical change raises from a general regulatory standpoint—considering how AI-enabled sociotechnical changes could produce (1) possible market failures; (2) risks to human health or safety; (3) risks to moral interests, rights, or values; (4); threats to social solidarity; (5) threats to democratic process; or (6) threats directly to the coherence, efficacy or integrity of the existing governance system charged with mitigating these prior risks—or with reference to specific doctrines under international law.

Finally, this approach calls for an analysis of the structure of AI as a *governance target* for the regime, considering both material-architectural features (e.g. AI weapons' viscerality; dependence on computing hardware), as well as considering the *problem logics* engaged (does the governance problem concern ethical challenges, safety risks, security threats, structural shifts, common benefits, or governance disruption)?

(2) [***Triage governance rationales, and consider indirect challenges***]

A narrow governance focus on ‘exceptional’ technological change may put disproportionate attention on certain hypothetical problems (e.g. self-driving car trolley problems) which might be visceral or conceptually interesting, but which are relatively peripheral. The sociotechnical change perspective can help focus attention on those AI capabilities and uses which pose particularly urgent rationales, and thereby help triage regulatory attention.

Likewise, the sociotechnical change perspective also emphasises that, along with high-profile challenges in the areas of *ethical challenges*, *security threats*, and *safety risks*, there is also a need for AI governance to explore and address ‘indirect’ challenges that have the problem logics of *structural shifts*, *common benefits*, or *governance disruption*.

(3) [*Don’t attempt to predict or wait; anticipate*]

The lens of sociotechnical change acutely highlights the epistemic limits of prediction in technology governance. Even if there are potential improvements in forecasting the near-term technological landscape of AI progress,³³⁵ second-order societal impacts are more complex yet, and more precarious to predict.

Yet curiously, the degree to which a technology’s full sociotechnical impact often does not become clear until well after its introduction, paradoxically highlights how reactive ‘wait-and-see’³³⁶ or ‘firefighting’³³⁷ approach to governing a new technology may often not be in a much better epistemic position than preventative regulation is. Especially at the international level, the time necessary to establish broad agreement that a problem exists, leaves such reactive approaches very vulnerable to the ‘power’ prong of the Collingridge Dilemma. There is no good or clear-cut solution to this problem. Nonetheless, to some degree, strategies of ‘legal foresighting’ could be used,³³⁸ with the aim not being to implement regulation for specific predicted changes, but to ‘prepare the ground’ in the general anticipation of sociotechnical change. This could accordingly help steer AI technologies in more beneficial directions,³³⁹ or help in ‘stress-testing’ various legal concepts in advance, in order to assess the relative degree to which different norms would be vulnerable to different types of disruption (e.g. conceptual ambiguity; inappropriate scope; obsolescence) by certain technological advances and sociotechnical changes.³⁴⁰

³³⁵ Ross Gruetzmacher et al., *Forecasting AI Progress: A Research Agenda*, ARXIV200801848 Cs (2020), <http://arxiv.org/abs/2008.01848> (last visited Aug 11, 2020). See also the discussion in chapter 2.1.6.4. (on Expert Prediction and Governing Under Uncertainty).

³³⁶ Crootof, *supra* note 94 at 21–22.

³³⁷ Maas, *supra* note 148 at 151–152.

³³⁸ Graeme Laurie, Shawn H. E. Harmon & Fabiana Arzuaga, *Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty*, 4 LAW INNOV. TECHNOL. 1–33 (2012).

³³⁹ At the same time, it should also be admitted that there are perverse risks that, if foresighting brings into focus certain relative positions of various actors on the technology, this could erode the political buy-in for some leading actors to bind their hands. Maas, *supra* note 148 at 150. See also the discussion in Chapter 3.3.2.

³⁴⁰ Rosert and Sauer, *supra* note 55. (on whether to ground LAWS regulation on the principle of distinction or principles of human dignity).

7.5.1.2 Instrument Choice (Sociotechnical Change)

The sociotechnical change perspective suggests several instrument choice strategies in AI governance.

(4) [*New AI-application-specific regimes might be too siloed*]

Except in special cases, treaties or regimes for AI that are application-specific might be suboptimal. They could involve wasteful duplication of effort, stumble on the blurred boundaries between underlying inputs (e.g. data, compute, talent); algorithmic paradigms (e.g. ‘reinforcement learning’), domain-transferable capabilities (e.g. ‘prediction’), or specific applications (e.g. ‘predictive profiling’),³⁴¹ while simultaneously failing to reckon with cross-domain patterns of sociotechnical change, or common problem logics.

(5) [*To be functionally extended to AI issues, existing regimes require harmonisation*]

In many circumstances, AI governance should or could work with pre-existing regimes. As noted, there are currently a wide variety of AI initiatives ongoing within existing international organisations such as the ITU, ICAO, IMO, the Global Forum for Road Traffic Safety, and at other UN initiatives.³⁴² For some applications, these may be functional. Nonetheless, AI-driven sociotechnical impacts are unlikely to (fully) track the pre-existing siloes of legal systems or mandates of older international organisations, suggesting that a simple extension of those mandates might produce fragmentation, incoherent policies, or unanticipated externalities. As such, the sociotechnical change suggests that extending older institutions or regimes from other domains should be supported by extensive efforts at regime harmonisation.

(6) [*A definitive global treaty for ‘AI’ might not respond to AI’s governance rationales*]

The sociotechnical change perspective weakly suggests that establishing an exhaustive and comprehensive global ‘AI’ regime in a single treaty might not be appropriate, as it could misconstrue either the *rationale* for global AI governance, which is not the appearance of ‘AI’ as such, but specific sociotechnical changes with distinct problem logics. Likewise, such a regime could misconstrue the adequate *targets* for AI governance, which in many situations might be best situated not at the macro-level of ‘AI’ as a diverse class, or at the micro-level of very local AI applications, but rather at the meso-level of cross-domain AI capabilities or key inputs (computing hardware).

³⁴¹ At least at the general regime level. To be certain, individual rules, standards or provisions could be articulated with reference to applications or designs of AI. See also Jonas Schuett, *A Legal Definition of AI*, ARXIV190901095 CS (2019), <http://arxiv.org/abs/1909.01095> (last visited Jan 6, 2020).

³⁴² See also Chapter 2.3.2.3. (mapping early steps towards institutionalization).

That does not mean that comprehensive or global instruments are categorically inadequate; for instance, framework conventions could serve to articulate a set of general principles (potentially corresponding to different AI problem logics), but leave specific regulation to later amendments or treaties.

7.5.1.3 Instrument Design (Sociotechnical Change)

The sociotechnical change perspective suggests an instrument design strategy for AI governance.

(7) [*Sociotechnical change foregrounds technology-neutral regulation*]

The sociotechnical change lens can generally highlight the value of focusing regulation not on distinct AI architectures or new applications. By considering AI as 'general-purpose technology', this helps emphasize instead the relative (if not universal) utility of articulating governance instruments (whether soft or hard law) to reckon with capabilities.

7.5.2 AI Governance Disruption and Strategies for Resilience

AI governance will increasingly need to adopt new strategies for resilience, in the face of patterns of governance disruption. Such resilience will mean the AI governance system can absorb, respond to, or even harness the effects of AI-driven governance disruption, in order to carry out development where it is needed, and to avoid the outcomes of governance destruction (through erosion or –decline). This goal has implications for approach, instrument choice, and instrument design.

7.5.2.1 Conceptual Approach (Governance Disruption)

The three types of governance disruption emphasize a series of conceptual shifts in technology governance. Significantly, many of these shifts may be important not only when considering the global governance of AI, but in some cases might also matter to ensuring the resilience of the global legal order more broadly.

(8) [*Expect 'Normal Disruption' of the Global Coordination Architecture*]

As noted, not all AI applications should be expected to result in governance disruption. However, the full range of applications of this general-purpose-technology, including its versatile use either in support or in contestation of international law, can spell a larger shift. Accordingly, more than in the past, global governance will have to reckon with impacts of technology *on* the global coordination architecture itself.³⁴³ This suggests that AI governance scholars must anticipate better how AI capabilities can produce patterns of governance disruption that may affect the substance (development), practices (displacement) or political scaffolding (destruction)

³⁴³ See also JELINEK, WALLACH, AND KERIMI, *supra* note 167 at 5.

of global governance and specific regimes. In coming years, given the pre-existing complexity and trends of change from either further AI progress or algorithmic proliferation, such disruption may become less an exception, and increasingly a ‘normal’ (if not ‘predictable’) part of the operation of the global governance system.³⁴⁴

(9) [*Beware unreflexive technology analogies in treaty (re)interpretation*]

New governance disruption can create situations of legal ambiguity, calling for appropriate development of legal or governance instruments. One appealing way to resolve such ambiguities, whether in a domestic or an international law context, may be to invoke plausible technological analogies. Yet the governance disruption lens speaks against an unreflexive reliance on underspecified analogies or legal fictions to achieve ‘quick wins’ against local AI challenges. This is especially the case given the unusually broad and diverse spectrum of analogies that AI technology supports.

That is not to say that analogies cannot offer productive insights, however, a premature convergence to the first plausible analogy may risk locking in less fruitful or sustainable regulatory approaches.³⁴⁵ The additional risk is that such rapid analogies may simply weave in more underlying conceptual ambiguity into the governance system. Finally, a too-quick invocation of analogical reasoning in order to attempt a (re)interpretation or extension of existing treaties or norms to new (AI) technologies may well backfire, if it cannot pass a plausible ‘laugh test’, or if it makes states more hesitant to sign up to technology-specific treaties in general.³⁴⁶

(10) [*Pick your battles, and beware of legal hard-ball*]

To avoid the unnecessary erosion of regimes or the broader authority of international law, AI governance advocates may in some circumstances have a strategic interest in avoiding legal ‘hard-ball’.

³⁴⁴ This notion of ‘normal disruption’ is deliberately meant to echo the idea of ‘normal accidents’—to connote systemic failures or breakdowns which appear exceptional, but which are in fact the emergent and inevitable result of the ‘normal’ way the system has been set up to operate. Maas, *supra* note 62; Maas, *supra* note 74 at 300–302. Indeed, the analogy with ‘normal accidents’ may be more generally enlightening of when, how and why particular (international) legal systems are susceptible to Governance Disruption. Under normal accident theory, sociotechnical systems give rise to ‘normal accidents’ given the following conditions: (1) the system is *opaque and has high interactive complexity* (so no single operator understands it entirely, and effects echo across the system in ways that are not immediately or easily comprehensible); (2) it is *tightly coupled* (so effects can cascade before humans can intervene); (3) organizations have *multiple competing objectives* beyond safety; and (4) there is a *competitive* context. In a certain sense, these conditions have their analogues in international law: it is a *complex system*, one where treaty conflicts are not always identified in time even by the parties acceding to them; while generally slow-moving, international law is *tightly coupled* in the sense that no single actor may be able to unilaterally intervene to amend a multilateral treaty regime in order to halt an impending Governance Disruption cascade; (3) many actors in international law have (and arguably should) more goals than merely ‘maximize resilience to Disruption’; and (4) many regimes and institutions operate in a somewhat competitive ecology. To be certain, this theoretical analogy has its limits, and should not be stretched too far, but it is potentially interesting nonetheless. I thank Peter Cihon for prompting this comparison.

³⁴⁵ Rosert and Sauer, *supra* note 129.

³⁴⁶ Crootof, *supra* note 94 at 17.

That is, the attempt to initiate AI governance by starting with the establishment of treaty regimes on especially conceptually or politically challenging applications of AI (especially if these attempt universal inclusion) carries a risk that such governance efforts may not fail gracefully, but might rather result in governance gridlock or in the erosion of governance more broadly. This could endanger the broader legitimacy of international law, speed up fragmentation, or at least ‘poison the well’ for later, more targeted AI governance initiatives. That is not to say that global governance should avoid a challenge, but the success of AI governance may depend on the sequence in which governance regimes are pursued.

(11) [*Contain Digital Sovereignty and AI Nationalism*]

AI capabilities are of course not the only threat to international law, nor the largest challenge. Nonetheless, there are risks that certain types of AI progress can create increasing strategic incentives for states to violate specific treaty commitments or contest larger regimes. As such, strategies are needed to counteract AI-enabled drivers of governance destruction. This includes efforts to ‘spoil the spoils’ of noncompliance with AI regulations, by shifting state perceptions of the relative risks of rapidly rolling out an immature technology to key strategic or military roles.³⁴⁷ It also entails a need to introduce both regulatory and technological measures to counter the use of AI capabilities as anti-governance weapon, and to preserve the broader epistemic security of the global legal order. Indeed, this will be a key challenge that will apply not just for AI regimes, but also for many others.

7.5.2.2 Instrument Choice (Governance Disruption)

The governance disruption perspective suggests several instrument choice strategies in AI governance.

(12) [*Treaties may be brittle*]

New multilateral treaties are often still seen as the optimal solution to governing new technologies.³⁴⁸ However, they may not remain an optimal instrument for AI governance in many areas. For one, they are very slow to establish, and have to overcome the ‘breadth vs. depth’ dilemma. Moreover, in the absence of specific design features, treaty instrument may be especially vulnerable to the tensions and pressures of AI-driven. That suggests they can be easily disrupted or fragmented by future technological changes that produce legal uncertainty. Indeed, there are many reasons to suspect that the risk of ‘treaty rot’ may be especially acute in the context of AI technologies. This could mean that treaties might be of only temporary effectiveness, or at worst undercut the coherence of governance in the longer-term.

³⁴⁷ For an example of such a framing, see DANZIG, *supra* note 61. (framing the rushed pursuit of narrow technological supremacy as a ‘technology roulette’ that does not improve national security).

³⁴⁸ Crootof, *supra* note 51 at 109. (“treaties are generally viewed as the gold standard, the ideal—if sometimes unattainable—expression of a state’s international legal obligations”).

(13) [***Full bans could be more resilient to disruption but may not hold the door***]

As a special case of the above, the pressures of ‘development’ pose a curious challenge to attempted full bans on certain applications, whether grounded in a precautionary principle,³⁴⁹ or in a ‘preventative security regime’.³⁵⁰ In theory, full bans are more resilient to governance disruption than are partial bans or regulatory treaties, because they can stem all future governance ‘disruption’ at the source. However, as discussed previously,³⁵¹ this assumption may be somewhat ill-suited in regulating a general-purpose technology such as AI, where it might not be able to ‘hold the door’ in this way. For a ban to remain resilient in this way, it might either need to curtail and contain a wide array of AI techniques (both unlikely and undesirable), or face the risk that a narrowly tailored ban on certain AI use cases comes under pressure by ‘spillover’ AI progress in other domains, involving easily transferable capabilities.

That is not to suggest that bans outlawing certain AI applications (e.g. facial recognition; LAWS) could not be functional or useful. Indeed, they might be critical for slowing or halting challenges from AI in the near term, with path-dependent consequences. Nonetheless, such bans must anticipate patterns of regime disruption and erosion.

(14) [***Customary International Law as Fall-back Strategy***]

The lens of governance disruption suggests that relying on gradual norm development under customary international law may not be optimal to address AI issues in the first instance, but may have unexpected merits as a ‘fall-back’ strategy.

On the downside, customary international law is slow and reactive; it would not work well at setting norms for hidden or undisclosed AI uses (such as in cyberwarfare, computational propaganda, or predictive modelling of rival state behaviour or conflict dynamics).³⁵² There is also the risk that the use of AI systems in many contexts may involve the gradual accretion of norms through operational standards of appropriateness, rather than deliberative forums of international lawmaking,³⁵³ and therefore would offer infertile ground for the creation of customary international law.

Nonetheless, customary international law does have the important benefit that it is not grounded in state consent, and that it therefore achieves universal inclusion.³⁵⁴ This is valuable for two reasons: in the first place, it could resolve the ‘breadth vs. depth’ dilemma facing centralised treaty regimes. In the second place, universality could help ensure greater resilience to governance disruption. The global nature of the digital economy suggests that bans or restrictions need to be applied relatively universally, if they are to stem or steer particular lines

³⁴⁹ Crootof, *supra* note 94 at 22–23.

³⁵⁰ Garcia, *supra* note 57.

³⁵¹ See also section 7.1.3.1. (‘Categorical Ban’).

³⁵² The latter case as discussed by Deeks, Lubell, and Murray, *supra* note 36.

³⁵³ Bode and Huelss, *supra* note 138.

³⁵⁴ With the exception of persistent objectors, which consistently contest or deny a rule’s existence or applicability and are therefore exempted from it; however, such cases are rare. Crootof, *supra* note 51 at 127.

of technological development that would otherwise induce later disruption of the regime.³⁵⁵ This suggests that while CIL should not be relied upon as the sole or even principal spearhead of AI governance, it can serve as important pillar.

(15) [***Standards over rules***]

More generally, the need for flexibility in AI regimes suggests that AI governance should often come to place a greater emphasis on the use of standards over rules,³⁵⁶ or on flexible soft law instruments over hard law.³⁵⁷ This might be one way by which to buffer regimes from situations of legal uncertainty, or at least ensure they are better able to carry out required developments (avoiding governance erosion).

At the same time, AI governance by soft law may need to avoid two risks. In the first place, it should go beyond voluntary industry-produced ethics codes,³⁵⁸ and derive from multi-stakeholder institutions. In the second place, steps should be taken to ensure that soft law regimes do not inadvertently undercut older hard law instruments.³⁵⁹

(16) [***Beware the unrestricted automation of international law...***]

With regards to the prospects of *displacement*, scholars of global governance generally should more actively scrutinize the trend towards ‘high-tech international law’.³⁶⁰ In general, they should be wary of the unregulated or unrestricted dissemination and incorporation of (AI) tools in processes of international lawmaking or adjudication. This is because, under plausible conditions, unrestricted usage is likely to further exacerbate power imbalances amongst leading AI states and smaller states, potentially feeding dissatisfaction and contestation.³⁶¹ Even in more unambiguously cooperative uses in support of conflict resolution, the use of such AI tools to speed up law-creation could accelerate overall processes of fragmentation, and therefore should be monitored.

In particular, this lens might urge scrutiny of AI applications which could, directly or indirectly, shift the ‘regulatory modality’ away from international conflict resolution through legal norms or processes such as arbitration. The availability of certain AI tools such as population sentiment analysis, computational propaganda, or lie-detection systems might lead some states to seek to increasingly resolve certain foreign policy conflicts through bilateral state channels—or even through unilateral technological intervention—and only to submit for arbitration those

³⁵⁵ Although given the relatively unequally distributed nature of AI talent and hardware capabilities in the near-term, this may be less of a relative advantage.

³⁵⁶ Picker, *supra* note 263 at 185.

³⁵⁷ Crootof, *supra* note 51 at 124–126.

³⁵⁸ So-called ‘supersoft law’; Thomas Burri, *International Law and Artificial Intelligence*, 60 GER. YEARB. INT. LAW 91–108 (2017).

³⁵⁹ Gregory C Shaffer & Mark A Pollack, *Hard vs. Soft Law: Alternatives, Complements, and Antagonists in International Governance*, 94 MINN. LAW REV. 706–99 (2010).

³⁶⁰ Deeks, *supra* note 225.

³⁶¹ *Id.* at 644–646.

issues where they predict they will win.³⁶² Such shifts could well erode the broader standing or relevance of the global legal order.³⁶³ As such, they should be cautiously monitored and held accountable.

(17) [*...but recognize and promote cooperation-supportive AI tools*]

Nonetheless, while a general degree of caution towards international legal automation may be warranted, it should also be remembered that AI-driven displacement also offers many opportunities to increase the efficacy of the global governance architecture, as well as its resilience not only to future technological disruption but also to other trends. This speaks in favour of accelerating the development and—subject to rigorous and careful assessment and constraints—dissemination of AI tools that could help improve structural conditions for international cooperation. There are various applications of AI that could thus help shift the international 'cooperation-conflict balance'. These could include AI monitoring capabilities that help provide greater assurance of the detection of state noncompliance, and in so doing help resolve situations of pervasive governance gridlock.³⁶⁴

Alternatively, technological and institutional interventions could increase the ability of various parties to make verifiable claims about their own AI capabilities, providing stronger foundations for AI-focused arms control regimes.³⁶⁵ Moreover, various technological interventions (including but not limited to AI tools) could support cooperation and coordination more generally, such as systems that could enable states to reach agreement on issues more quickly.³⁶⁶ Finally, from the perspective of regime complexity, relatively simple AI language models could allow the advance identification of emergent or imminent norm or treaty conflicts, whether accidental or strategically engineered.³⁶⁷ Over time, they could also enable a greater mapping and monitoring of fragmentation both in the AI regime complex,³⁶⁸ as well as in the broader governance architecture, supporting processes of convergence. It would be valuable to offer such AI applications free or low-cost to many parties, and teach them how to deploy and use such systems.³⁶⁹ Such applications could both increase the inclusion of many actors in the fora of international governance,³⁷⁰ as well as help in the management of governance fragmentation more broadly.

³⁶² *Id.* at 628–629.

³⁶³ However this could exacerbate the fragmentation of international law, adding a (vertical) dimension of fragmentation to the existing inter-regime conflicts

³⁶⁴ These could dissolve or at least reduce the 'transparency-security' trade-off, and facilitate 'closed' arms control deals. See also the discussion in Chapter 5.4.2.

³⁶⁵ Brundage et al., *supra* note 84 at 67–69.

³⁶⁶ Deeks, *supra* note 225 at 647–648.

³⁶⁷ *Id.* at 616.

³⁶⁸ For instance, in Paper [IV], we also call for the use of various NLP-tools to improve the monitoring of trends of conflict, coordination or catalyst in the AI regime complex. Cihon, Maas, and Kemp, *supra* note 167 at 233.

³⁶⁹ Deeks, *supra* note 225 at 650–652.

³⁷⁰ Höne, *supra* note 223.

7.5.2.3 Instrument Design (Governance Disruption)

The governance disruption perspective suggests several instrument design strategies in AI governance, some of which could help make even hard law instruments or treaties more resilient to future disruption on various grounds.

(18) [Technology-neutral regulation foregrounded]

While the governance Disruption perspective does not suggest that technology-specific governance is obsolete, it does highlight how and why such regulations or treaties may become especially vulnerable to disruption, leading to regime obsolescence or inter-regime conflicts. Often, they may not be able to carry out required developments, because attempts to re-interpret highly tech-specific treaties to apply to new technological capabilities may not pass the ‘laugh test’.³⁷¹ This suggests a relative increase in the prominence of technological neutrality in drafting instruments and regulations. Of course, this too faces challenges, since states may be more cautious about signing up to expansive technologically-neutral regulations. Moreover, many functions served by AI systems are of a sufficiently high level of generality or abstraction (‘prediction’; ‘pattern recognition’) that resulting regimes could become over-inclusive.³⁷² However, this can be mitigated by instead articulating such norms, not in terms of a ‘technologically neutral’ function (e.g. ‘data processing’), but in terms of a ‘sociotechnically neutral’ impact.

(19) [Pursue more flexible treaty designs]

While from an instrument choice perspective, governance disruption may weakly speak against multilateral treaties, this does not mean that such instruments are not redeemable. Indeed, AI governance proposals could consider more flexible treaty designs, allowing easier or more rapid development in response to situations of governance disruption. This could take the form of an AI-focused framework convention, or modular conventions that enable the incremental addition of new (whether capability- or technology-specific) provisions through Additional Protocols.³⁷³ At the same time, modular treaty designs are not, in and of themselves, sufficient to make up for (e.g. enforcement) shortcomings in the design of the base treaty.³⁷⁴ This implies that provisions must be made to ensure adequate institutional mechanisms for the regular review and expansion of regimes.³⁷⁵

³⁷¹ Crootof, *supra* note 94 at 15. On the other hand, the very complexity, diversity, and opacity of some AI techniques could in some cases render such an approach more viable, because it might be less obvious, to the lay person, that a certain interpretation or analogy is technologically spurious. However, this would also highlight the risk of achieving treaty reinterpretation by adopting unreflexive analogies (see Chapter 5.3.6.2).

³⁷² Maas, *supra* note 148 at 153.

³⁷³ Crootof, *supra* note 51 at 120. For an evaluation of framework conventions as international regulatory tool, see also Matz-Lück, *supra* note 286.

³⁷⁴ For instance, compare the lack, in the Convention on Certain Conventional Weapons, of any verification or enforcement mechanisms, or formal process for resolving compliance concerns.

³⁷⁵ Cihon, Maas, and Kemp, *supra* note 167 at 232–233.

(20) [***Provisions to let the future decide***]

Finally, AI governance could consider reserving more decision-making ability for future actors. In their mandate, international organisations established to deal with AI could be granted greater discretion to alter rules as AI technology or use changed;³⁷⁶ they could also designate 'authoritative interpreters' or institutional bodies charged with reviewing such changes.³⁷⁷ As Crootof notes, one problem currently is that while domestic legal regimes often designate authoritative interpreters who can evaluate the scope and implementation of old laws in new cases, there is no such overarching entity at the international level.³⁷⁸ To be sure, some international treaties do delegate interpretation to adjudicators, though this is not always mandatory, and at times is only an opt-in.³⁷⁹ As a consequence, the legitimacy of new interpretations of multilateral treaties in the face of governance disruption will generally depend on whether the other treaty partners accept this, which can be a long, tortuous, and confused process.³⁸⁰

Designating a mandatory authoritative interpreter for AI-focused treaties or instruments could be one way to help speed up the process by which that regime might resolve the 'development' necessary to manage the pressures of governance disruption. Indeed, as a general rule, 'governance disruption' suggests we may increasingly prefer to design technology governance regimes in ways that shift power of interpretation from present rule-makers to future rule-takers, as the latter will have more granular information available.³⁸¹ At the same time, in such arrangements, care should be taken to include guidelines or provisions to ensure authoritative interpreters are not too easily tempted by the unreflexive use or application of analogies for new technological capabilities, without first rigorously and critically scrutinizing the sociotechnical accuracy, implied regulatory logics, and conceptual risks of each avenue within a larger portfolio of possible analogies.³⁸²

7.5.3 AI Regime Complexity and Strategies for Coherence

Finally, the lens of regime complexity demonstrates how actors in AI governance may need to adopt distinct strategies for regime coherence. As earlier noticed, the focus here is less on the underlying strategic question of whether AI governance (for some or all issues) 'should' preferably be organised as a centralised institution or a decentralised regime complex.³⁸³ It is

³⁷⁶ See also brief discussion by Picker, *supra* note 263 at 185.

³⁷⁷ Crootof, *supra* note 51 at 120.

³⁷⁸ *Id.* at 120.

³⁷⁹ Aaken, *supra* note 295 at 85.

³⁸⁰ Crootof, *supra* note 51 at 120.

³⁸¹ An alternative strategy might be to work through types of sunset clauses or sunset treaties, but this carries risks that certain regimes would not be renewed at a later point. On the other hand, the mere fact of certain treaties being time-bound could conceivably act as a pressure-valve on near-term contestation by certain parties.

³⁸² See also strategy #9.

³⁸³ This may not be a practical question, because scholars generally hold that regime complexes are the emergent results of complex interactions amongst many actors, rather than the result of deliberate design. See also the discussion in 7.3.1. Note, exceptions may be found in deliberately engineered 'treaty conflicts'. E.g. RANGANATHAN,

rather about exploring potential strategies or interventions to ‘manage’ the interaction of the AI regime complex, whatever its form. Exploring strategies to manage institutional interplay would not be about trying to dictate the pre-existing structure or development of a regime complex (in terms of its component institutions and their goals, norms and mandates), but rather about meta-strategies or policies that actors or institutions can adopt to promote beneficial interactions and productive dynamics within the regime complex. An AI governance system can ensure coherence means that actors have an ability to manage changes and stresses in the regime architecture, specifically in terms of inter-regime or inter-institutional conflicts over norms, goals, or outcomes.

Accordingly, the focus here is on articulating potentially valuable strategies conditional on either trajectory. If a centralised global AI institution is established (either in the near term, or after a period of evolution), then policymakers might wish to design that institution in ways that avert or at least minimize some of the identified challenges this brings (in brittleness, breadth vs. depth), because bad design could lock in particularly suboptimal or bad outcomes.³⁸⁴ If the AI regime complex remains fragmented, then this raises questions over how to manage the inter-institutional and inter-regime interactions, in order to minimize conflictual dynamics or problems such as forum shopping.

7.5.3.1 Conceptual Approach (Regime Complexity)

The regime complexity lens suggests AI governance will have to take on board certain conceptual insights.

(21) [Consider AI governance in its broader governance ecology]

For many other issue areas, global governance scholars have found the regime complex lens a valuable framework through which to better reconcile international legal insights into the top-down effects of global norms, with international relations’ insights on actors’ cross-institutional strategies.³⁸⁵ AI governance should not need to reinvent the wheel. Accordingly, more attention should be paid in AI governance to chart not just technological change and applicable norms, but also the messy political, organisational, and institutional factors that are at work in a regime complex. This reinforces the idea that AI governance solutions should not be sketched in a vacuum, but need to be considered in potential relation to other regimes. More generally, this implies that AI governance proposals should also be tailored to trends in the global governance architecture,³⁸⁶ as well as trade-offs in the distribution of the specific institutional landscape on AI governance.

supra note 226. But even here, states may not be able to foresee or control all the downstream normative and institutional consequences for the regime complex.

³⁸⁴ Cihon, Maas, and Kemp, *supra* note 167 at 232.

³⁸⁵ Benjamin Faude & Thomas Gehring, *Regime Complexes as Governance Systems*, in RESEARCH HANDBOOK ON THE POLITICS OF INTERNATIONAL LAW 176–203 (Wayne Sandholtz & Christopher Whytock eds., 2017), <https://www.elgaronline.com/view/9781783473977.xml> (last visited Oct 15, 2019).

³⁸⁶ Including pre-existing institutional density, institutional accretion, power shifts over time, preference changes, modernity, representation and voice goals, and desire for local governance. See also Chapter 7.3.2

(22) [*Explicitly consider avenues to shape underlying regime foundations*]

Relatedly, for a number of AI governance issues, it is valuable to explore not just the viability of international governance given current (state) interests, but also articulate longer-term strategies through which epistemic communities or norm entrepreneurs can shift norms, shaping perceived interests or the structural knowledge conditions for international AI cooperation. This can include considerations in terms of issue selection, issue framing, forum selection, and community organisation.³⁸⁷

7.5.3.2 *Instrument Choice (Regime Complexity)*

The regime complex framework suggests several instrument choice strategies in AI governance.

(23) [*Consider trade-offs of centralisation and decentralisation in selecting instruments*]

Ultimately, a comparison of the relative merits of a fragmented and decentralised regime complex for certain AI issues, should be influenced by a more explicit consideration of the distinct factors and trade-offs that face such a regime—including requisite political power, efficiency and participation, slowness, brittleness, breadth vs. depth dilemma, susceptibility to forum-shopping, and ability to coordinate policy. These factors need to be considered not only in general, but especially in the context of different AI challenges (whether in traditional application domains, or bucketed by sociotechnical change). This cannot produce hard and fast rules, but we can theorize some general heuristics.

To generalize, for AI issues where successful management may require that the regime (1) wields considerable *political power* or (2) involves *participation* from many actors, while heading off (3) *forum shopping*, this appears to speak in favour of attempting to create an integrated institution over settling for the currently fragmented regime complex. This could be the case for ‘weakest-link’ public goods problems around AI, such as the restriction of cyberwarfare or certain military applications.

Conversely, there may be other AI issues where political power, universal participation, or risks from forum shopping are less pivotal to governance, or where we rather (1) expect critical technological turning points to occur in the near term, or at least suddenly with little warning (inflicting a large amount of sociotechnical change before an integrated regime could be established in response); or (2) where we anticipate regular subsequent capability changes and shocks. In such situations, a centralised regime might prove too *slow* or too *brittle*, respectively, and policymakers might instead find it not only sufficient but preferable to instead rely on a range of faster, informal governance initiatives, as these might more rapidly close the near-term gap in governance, and more appropriately or flexibly respond to any future changes in AI usage as and when they emerge or become apparent.

³⁸⁷ See also the discussion of epistemic community considerations in *Paper [I]*. Maas, *supra* note 74.

(24) [***Explore adaptive instruments or strategies that mitigate or bypass trade-offs***]

Along these general considerations, it may be worth to explore regime forms or instruments that could be particularly promising at bypassing or reducing these trade-offs. For instance, *framework conventions* have previously been proposed as one instrument that might be able to mitigate the breadth vs. depth dilemma, while simultaneously supporting a degree of flexibility.³⁸⁸ For other issues, informal or smaller governance initiatives carried out by coalitions of like-minded actors can mature into *critical mass* governance agreements.³⁸⁹

Indeed, in the domain of security regimes for military AI applications, regimes could dovetail with a broader proposed reinvention of the overall non-proliferation and arms control regime architecture, which draw on existing institutions and treaties as (1) ‘toolbox’ for likeminded states’ initiatives (e.g. on military AI); (2) as ‘clearing house’ for partnerships between donors and recipients; (3) or as ‘platform for interaction’ between science and diplomacy.³⁹⁰

7.5.3.3 *Instrument design (Regime Complexity)*

Finally, the regime complex framework suggests several instrument design strategies for AI governance. Of course, as noted, the trajectory of the AI regime complex depends to some degree on external factors, and is not fully under deliberate control. Nonetheless, we can still consider strategies conditional on the trajectory which the AI regime complex takes.

(25) [***Foster regime interplay management across a fragmented AI regime complex***]

For instance, if the AI regime complex remains fragmented, there are many strategies which could be undertaken to minimize the negative consequences. For instance, from a legal perspective, the International Law Commission has articulated a range of strategies for managing regime interaction, which include new specific treaties, rules for treaty interpretation, explicit conflict clauses, or priority rules.³⁹¹

Of course, as noted, adverse regime interactions in a fragmented regime complex are not limited to (legal) norm conflicts, but also include potential strategies of forum-shopping,

³⁸⁸ Matz-Lück, *supra* note 286.

³⁸⁹ Cihon, Maas, and Kemp, *supra* note 167 at 233; Kemp, *supra* note 293.

³⁹⁰ See Ignacio Cartagena Núñez, *Managing Complexity: Three Possible Models for a Future Non-proliferation and Arms Control Regime*, UNIDIR (2019), <https://www.unidir.org/commentary/managing-complexity-three-possible-models-future-non-proliferation-and-arms-control>.

³⁹¹ As articulated in MARTTI KOSKENNIEMI & STUDY GROUP OF THE INTERNATIONAL LAW COMMISSION, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law* (2006), http://legal.un.org/ilc/documentation/english/a_cn4_l682.pdf. BEATRIZ MARTINEZ ROMERA, REGIME INTERACTION AND CLIMATE CHANGE : THE CASE OF INTERNATIONAL AVIATION AND MARITIME TRANSPORT 230–237 (2017), <https://www-taylorfrancis-com.ep.fjernadgang.kb.dk/books/9781315451817> (last visited Jan 11, 2020). See also the legal framework within which to conduct regime interaction, proposed in: Margaret A. Young, *Regime Interaction in Creating, Implementing and Enforcing International Law*, in REGIME INTERACTION IN INTERNATIONAL LAW: FACING FRAGMENTATION 85–110 (Margaret A. Young ed., 2012), <https://www.cambridge.org/core/books/regime-interaction-in-international-law/regime-interaction-in-creating-implementing-and-enforcing-international-law/8F958E230DD068D4E6CE9F1141B9D65B> (last visited Sep 10, 2020).

competitive regime creation, or inter-institutional conflictual dynamics.³⁹² However, regime complexity scholars have suggested that there may be ways by which such interactions within regime complexes can be actively managed.³⁹³ Such decentralised coordination is referred to as ‘*interplay management*’, defined as the “conscious efforts by any relevant actor or group of actors, in whatever form or forum, to address and improve institutional interaction and its effects”.³⁹⁴

There are distinct forms of interplay management, which AI governance scholars should consider: under ‘*orchestration*’, state actors and intergovernmental organisations can aim to work with private actors in the pursuit of regulatory goals;³⁹⁵ in cases of ‘*institutional deference*’, international organisations recognize and cede regulatory authority and jurisdiction to other organisations, in order to reduce overlaps or conflicts and divide labour.³⁹⁶ AI governance scholars should explore such measures in greater detail. Indeed, such interplay management could be supported by diverse proposed initiatives.³⁹⁷

(26) [*Equip a centralised AI institution for inclusion and adaptation*]

Finally, if or as a centralised AI organisation is pursued, policymakers should consider a specific design desiderata to ensure this approach is not counterproductive.³⁹⁸ For one, such an institution should ensure that neither its breadth nor its depth fall below the critical thresholds of participation or substantive provisions necessary to address the AI challenges under concern. Other design steps should focus on avoiding or mitigating the risks of regulatory capture, ‘unified failure’, or lack of adaptability. Accordingly, a global AI organisation should provide for AI monitoring capabilities as well as include expert group, informal meetings or review mechanisms; they might embrace experimentalism, and put less emphasis on the specificity or stability of its rules, and more on their adaptability, consistency and efficacy to the challenge at hand.³⁹⁹

7.5.4 An overview of strategies

In sum, this section has reviewed potential strategies to improve the efficacy, resilience, and coherence of AI governance systems, encompassing diverse shifts at the conceptual level as well as in terms of instrument choice and -design (see Table 7.3).

³⁹² As explored in section 7.4.

³⁹³ Keohane and Victor, *supra* note 278; Morin et al., *supra* note 134.

³⁹⁴ MANAGING INSTITUTIONAL COMPLEXITY: REGIME INTERPLAY AND GLOBAL ENVIRONMENTAL CHANGE, 6 (Sebastian Oberthür & Olav Schram Stokke eds., 2011), <https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262015912.001.0001/upso-9780262015912> (last visited Jan 30, 2020).

³⁹⁵ Kenneth W. Abbott & Duncan Snidal, *International regulation without international government: Improving IO performance through orchestration*, 5 REV. INT. ORGAN. 315–344 (2010).

³⁹⁶ For instance, scholars have tracked numerous such patterns of deference in the regime complexes on counterterrorism, intellectual property, and electoral monitoring. Tyler Pratt, *Deference and Hierarchy in International Regime Complexes*, 72 INT. ORGAN. 561–590 (2018).

³⁹⁷ Such as the recent proposed for a ‘Coordinating Committee for the Governance of Artificial Intelligence’. JELINEK, WALLACH, AND KERIMI, *supra* note 167.

³⁹⁸ Cihon, Maas, and Kemp, *supra* note 167 at 232. (*Paper IV*).

³⁹⁹ Liu and Lin, *supra* note 239 at 340.

These strategies are certainly no blueprints. They are necessarily rough sketches, in need of much further examination, critique, or validation in future work. Nonetheless, it is hoped that they illustrate the ways in which the three theoretical lenses explored in this investigation can provide specific suggestions for how AI governance proposals may remain fit in the face of this complex technology.

	Strategies for efficacy Sociotechnical change	Strategies for resilience Governance disruption	Strategies for coherence Regime complexity
Conceptual Approach	<ol style="list-style-type: none"> 1. Govern for sociotechnical change, not technology 2. Triage governance rationales, consider indirect challenges 3. Don't attempt to predict or wait; anticipate 	<ol style="list-style-type: none"> 8. Expect 'Normal Disruption' of the Global Coordination Architecture 9. Beware unreflexive technology analogies in treaty (re)interpretation 10. Pick your battles, beware legal hard-ball 11. Contain Digital Sovereignty and AI nationalism 	<ol style="list-style-type: none"> 21. Consider AI issues in broader governance ecology 22. Consider avenues to shape regime foundations (interest, norms)
Instrument Choice	<ol style="list-style-type: none"> 4. New AI-application-specific regimes might be too siloed 5. Extending existing regimes to AI requires harmonisation 6. A global AI treaty might mistake AI's governance rationales 	<ol style="list-style-type: none"> 12. Treaties may be brittle 13. Full bans could be resilient, but may not hold the door 14. Customary International Law as Fall-back Strategy 15. Standards over rules 16. Beware the unrestricted automation of international law... 17. ...but recognize and promote cooperation-supportive AI tools 	<ol style="list-style-type: none"> 23. Choice between centralisation and decentralisation depends on trade-offs 24. Explore adaptive instruments or strategies that mitigate or bypass trade-offs
Instrument Design	<ol style="list-style-type: none"> 7. Technology-neutral regulation foregrounded 	<ol style="list-style-type: none"> 18. Technology-neutral regulation foregrounded 19. Pursue more flexible treaty designs (framework conventions; modular treaties) 20. Let the future decide (authoritative interpreters) 1. 	<ol style="list-style-type: none"> 25. Foster regime interplay management across a fragmented AI regime complex 26. Equip a centralised AI institution for inclusion and adaptation 2.

Table 7.3. Overview of Strategies for AI regime efficacy, resilience, and coherence

7.6 Conclusion: AI Governance in Five Parts

This chapter has charted diverse opportunities, challenges and trade-offs for AI governance regimes, as developed through the three conceptual lenses developed in this work. In so doing, we drew on many examples involving international security regimes for military AI applications, although these lenses are relatively agnostic regarding the specific issue domain, and should as such be transposable to debates within other sectors of AI governance.

Specifically, this chapter has argued that an examination of AI governance systems—whether existing, emerging, or proposed—can and should engage with patterns of ‘change’ at five stages of an AI regime. In considering each of these aspects, it was argued that we can draw on the three facets of change explored in this work—sociotechnical change, governance disruption, and changing regime complexity—in order to help enrich our analysis, and improve our choices of policy and strategy.

Firstly, an examination of AI governance can examine the question of a given AI regime’s *origins* or foundations. This involves exploring (a) the regime *purpose*—whether there is a sufficient pattern of sociotechnical change to constitute a governance *rationale*; and what is the texture of the governance *target*. It then involves (b) a consideration of whether a particular governance regime is in fact *viable*. This analysis can take an outside-view or comparative historical perspective. It can also draw on regime theory to consider the *interests* of states and other governance actors. Finally, it can consider whether or how the regime could be rendered more viable given the ability of various actors to shift *norms* of key stakeholders. In the absence of either compatible state interests or sufficient norm-setting on a certain AI issue, regimes or institutions may be very unlikely to emerge (regardless of initial debates across international fora), and the issue may remain beset by gridlock, locked in a non-regime state, or see only very fragmented efforts. Even in cases of some confluence of interests with norms, this may not suffice to render *any* governance regime viable. But in some areas it can have a surprising degree of effectiveness. Finally, a consideration of regime *origins* can consider questions of *design*—what regime strategies would be effective? Which would be resilient to future governance disruption?

Secondly, individual AI governance efforts should reflect more actively on the *topology* of the broader AI regime complex, in order to identify gaps, shortcomings, and latent conflicts in either legal rules, norms, goals, or institutional activities. This involves examining the *demographics* of the growing AI governance system, its *organisation* (in terms of the density of the institutional network, and whether links relate to norms, goals, impacts or institutional relations); and the *interactions* and outcomes of these linkages (whether leaving gaps; conflictive, cooperative, or synergistic). Finally, these interactions can be examined at the macro, meso, or micro level. While such an analysis by itself can offer only a snapshot of an AI governance regime complex—one bound to be outdated by the rapid pace of change—it is still very relevant to undertake such mappings of topology. This is because the AI governance regime today remains in a state of relative fragmentation. Efforts should in particular consider fault-lines or overlaps amongst distinct regulatory instruments—potential conceptual or political tensions, which remain latent today but which could find themselves ‘activated’ or triggered in the wake of later governance disruption produced by advancing AI capabilities or regime complex change.

Thirdly, AI governance regimes should consider drivers of *evolution* over time, both gradual and sudden. To that end, they should engage with both general and AI-specific drivers of regime fragmentation—including pre-existing institutional density and accretion, background shifts in state power or preference over time, and various trends of modernity including the increasing complexity of modern ‘transversal’ challenges, and the increasing call for local governance. Many of these trends may appear to suggest an increasing fragmentation of many regimes, which may constitute a gradually moving boundary condition to governance in general. This should at least be kept in mind by those advocating for centralised AI regimes. Likewise, an examination of possible trajectories for AI regimes will have to increasingly calculate in the effects of AI-driven governance disruption on regime integrity: while some AI capabilities could see use in countering regime fragmentation or treaty conflicts, some AI capabilities could also serve as generators of new legal-conceptual fault lines, or spur increased patterns of contestation. That is not to suggest that these are trends that are inevitable, irresistible or irreversible. Yet those who hope to pursue general and integrated global regimes, may at the very least have to consider more explicitly designs or strategies by which the AI regime could be better insulated from these general trends, or could even be adapted to these developments.

Fourthly, we should consider the *consequences* of regime complexity in greater detail, as this provides key strategic considerations around AI governance. History and regime complex theory suggest that centralizing a governance regime can have various benefits (in terms of authority and political power, efficiency and lower participation thresholds), but also introduces new risks (in terms of slowness and brittleness). Moreover, in many cases, the added benefits of either governance approach depend far more sensitively on their exact institutional design. A powerful centralised regime might be able to avert strategic forum-shopping by states, but that same feature could turn into a liability if this prevents some actors from attempting or threatening to shift debate away in the face of stalled progress. Conversely, a ‘polycentral’ regime complex of many institutions might in principle be capable of generating and experimenting with more solutions, as its advocates claim; but the ability of institutions in a regime complex to carry this out in practice may depend sensitively on how well it is orchestrated. As such, rather than offering a straightforward choice between centralisation or fragmentation, in practice the question over which regime configuration would be more functional or legitimate depends sensitively on diverse factors, including one’s normative criteria, assumptions about the minimum necessary ‘depth’ of institutional authority or ‘breadth’ of inclusion necessary to adequately address a certain AI issue, as well as on assumptions about the pace and functionality of sociotechnical change produced in that domain by further AI progress.

Fifthly, a more nuanced understanding of regime complexity’s consequences can help suggest concrete *strategies* that can preserve or improve the efficacy, resilience and coherence of the AI governance system whatever its form or trajectory. Specifically, I have sketched how all three lenses have implications for our conceptual approach, instrument choice, and instrument design in AI governance. These are not meant to add up to a definitive blueprint for AI governance. However, they do drive home the importance of reckoning with these models. These specific strategies may certainly be contested. However, the larger point is that scholars and policymakers alike should increasingly explore *some* strategies that can ensure the efficacy,

resilience, and coherence of the AI governance system. This is critical if AI governance is to be fit not only today, but also throughout the changing future.

Conclusion

How should global governance for artificial intelligence account for change? Artificial intelligence is a complex technology. Global governance is a complex array of human practices. Both have their constants, but both also are subject to increasing diversity and change. Attempt to put these two together, and one will not have a simple day.

If this work has been successful in its aims, the questions it has raised are more interesting than the limited answers it has provided. AI governance remains an incipient field that is still in a state of flux. This field has seen considerable development through even just the last three years—and will likely see even more in just the next three. In spite of this early state (or in fact because of it), we should not shy back from investigation. Appropriate global AI governance strategies and regimes are urgently needed; there may be a limited window of opportunity to get them ‘right’, and failure to lay robust foundations at this early stage may have long and dependent effects on the trajectory of AI governance, and the character of the world’s relationship with this technology, for years to come.

As such, this project has not sought to provide definitive answers, nor to provide an in-depth study of a global institutional and normative landscape that is still undergoing change. Instead, I have sought to set out conceptual lenses which may enable scholars and practitioners to better understand the role of three facets of ‘change’ in AI governance. In doing so, I drew on three theoretical concepts:

- (1) *Sociotechnical Change*: which explores how (AI) technological capability change can give rise to changes in society; when these ground a rationale for governance; and how such governance strategies should approach their governance target in order to ensure efficacy.
- (2) *Governance Disruption*: which explores how AI can drive change in the substance, assumptions, processes or political scaffolding of (global) governance; how this creates new opportunities or threats; and how we might formulate strategies to ensure AI governance resilience.
- (3) *Regime Complexity*: which explores changes in the conditions for- or dynamics of AI governance, as these arise from broader changes in international governance architecture and regime complex interaction; and which accordingly suggests strategies for AI governance coherence.

Through conceptual analysis, historical cases, and examples from across the AI governance landscape, this work has investigated the strengths and limits of these three perspectives. As a result, we can now return to provide answers to some of the thematic sub-questions which we articulated at the start of this project.

A. Why do we require governance strategies for artificial intelligence? Why do these require new strategies for change?

In Chapter 2, I argued that AI technology may be expected to drive considerable global change, that this will require global governance; and that many existing approaches may be inadequate.

Defining AI definition: I first set out three ways of defining AI (as *science*, as *sociotechnical system in practice*, or as *lever for governance*), and argued in favour of a multi-part analytical definition of AI as sociotechnical system, which emphasised the links between AI *techniques, capabilities, applications* and *sociotechnical change*.

AI's importance today: given this definition, I reviewed the progress, promise and limits of AI technology over recent years, in order to provide a more nuanced understanding of AI's importance. I discussed the remarkable scale and speed of AI advances, but also put these against important and often underappreciated shortfalls or limits of AI technology today. I argued that more interesting than the debate of whether AI technology is categorically transformative or rather completely overhyped, is a perspective on AI as a heterogeneous, 'high-variance' technology.

AI's importance going forward—change from progress or proliferation: I then reviewed and clarified some of the debates over the uncertainties over future AI progress, as well as 'algorithmic proliferation'. I ultimately argued that even under very conservative assumptions about future fundamental progress, we should expect AI capabilities to see increasingly widespread and disruptive application. This is because of falling trends in input thresholds (data, compute and talent), the technology's inherent breadth of application, broader trends and the historical acceleration in the rate of global technology diffusion, and the existence of pre-existing sensor-data, and consumer infrastructures.

AI's governance problems: I argued that many applications of AI capabilities create diverse political, ethical and societal challenges. Drawing on typologies of global public goods, I suggested that not all of these problems demand global cooperation; however, many will benefit from it, and some likely will require forms of international governance for their effective mitigation.

AI's governance landscape today: given this problem, I next turned to the potential options and strategies that have been put forward. I discussed the strengths and limits of traditional international law as well as existing regimes for addressing AI challenges, before surveying recent developments in AI governance. I conclude that the current AI governance landscape has seen some promising steps, but remains fragmented and incipient.

Finally, after clarifying (in Chapter 3) my methodological choices and my underlying ontological and epistemic assumptions, this completed the Foundations of the project.

Accordingly, in Part II (Frameworks), I aimed to answer the three sub-questions that related to the three facets of 'change' in AI governance.

B. Why, when, and how should governance systems approach and respond to AI-driven sociotechnical change?

In Chapter 4, I drew on the work of Lyria Bennett-Moses to argue that the lens of sociotechnical change can serve as a useful heuristic for AI governance. This is because it puts the focus on when, where, and why new AI capabilities actually translates to new widespread behaviours, entities, or relations which create challenges for governance.

Broad definition of technology. This lens foregrounds a broad understanding of technology, as encompassing not just artefacts but broader sociotechnical systems. The focus should be less on technological change per se, and more on technological changes that create new affordances which, if seized upon widely, produce considerable ‘sociotechnical change’. Taking this sociotechnical change perspective is important or useful for AI governance from two perspectives.

Sociotechnical change and the analysis of governance rationales: in the first place, it helps us analyse when and why AI capabilities produce a governance rationale. From a general regulatory perspective, we can distinguish six rationales: (1) possible market failures; (2) new risks to human health or safety; (3) new risks to moral interests, rights, or values; (4) new threats to social solidarity; (5) new threats to democratic process; (6) new threats directly to the coherence, efficacy or integrity of the existing governance system charged with mitigating these prior risks.

Sociotechnical change and the analysis of governance targets: in the second place, this perspective can also help us grapple better with AI challenges as governance targets. This turns on both ‘material’ considerations, but also on the specific ‘problem logic’ at play. I distinguishes six types of sociotechnical change that AI uses could spur: ethical challenges; security threats; safety risks; structural shifts; common benefits; governance disruption. These ground distinct governance rationales, and present different problem logics. I suggested that these each come with distinct problem surfaces that can require or highlight distinct regulatory logics or approaches.

Limits and benefits: Importantly, it was noted that this taxonomy is not meant to be exhaustive, predictive, or prescriptive, nor to present direct blueprints for AI regulation. Instead, it is a thinking tool for understanding cross-domain challenges and features of AI. It also highlighted considerations for formulating distinct strategies to ensure AI governance *efficacy*.

C. Why, how or when might AI applications disrupt global governance, by driving or necessitating changes to its substance and norms, its processes and workings, or its political scaffolding?

In Chapter 5, I aimed to answer this question by sketching the theoretical foundations of various accounts of governance disruption, covering and extending earlier accounts developed in *Papers II and III*, as well as by drawing on conceptual frameworks, concepts, and case studies developed by a range of scholars including Lyria Bennett Moses, Rebecca Crootof, Roger Brownsword, Ashley Deeks, Hin-Yan Liu, and Colin Picker. Accordingly, I sketched a taxonomy of governance ‘development’, ‘displacement’, and ‘destruction’, and explored their historical precedents, preconditions, and limits.

AI governance development creates tensions and uncertainties in existing law. I argued that in cases of *development*, AI capabilities or applications create new legal gaps, reveal uncertainties or ambiguities in existing laws, blur the scope of application of these regimes,

render obsolete core justifying assumptions, or redraw or rebalance the problem landscape in the domain with which the regime is concerned, and in so doing create a pressure for new governance developments to clarify or resolve the shortfalls of the existing governance regime. We also discussed some of the nuances of when the required development can- or cannot be easily carried out, as well as the risks involved in relying too quickly on treaty reinterpretation to extend regimes, if such development relies on the unreflexive use of technology analogies.

AI governance displacement creates challenges and opportunities. In terms of displacement, I discussed the relative degree to which AI applications might or might not be integrated into the *processes* of the global legal order—whether in the automation of rule creation or adjudication; in support of compliance monitoring; or in a wholesale shifting of the ‘modalities’ of international law. I argued that while some of these uses could have beneficial effects, some could also have challenging legal, political and distributive consequences.

Failed development or active AI-enabled contestation could result in governance destruction. Finally, I discussed ways in which AI usage might drive- or contribute to potential destruction or decline of governance regimes. I suggested that this can occur either through indirect ‘erosion’, when AI challenges prove (for conceptual or political reasons) an intractable puzzle for existing governance instruments, such that the appropriate governance ‘development’ cannot be carried out, leaving certain issues visibly unresolved. Alternatively, it could result through ‘decline’, either because developments in AI increase the perceived ‘spoils’ of non-compliance with certain regimes; because some AI systems can serve as ‘weapons’ in the contestation of the global legal order; or because some AI systems may shift values of key actors away from this order.

Not all or even most AI applications produce governance disruption, but it may still put critical pressure on governance regimes. Ultimately, the argument here was not that AI applications will always—or even usually—result in governance disruption, but simply that those situations and areas in which they do may put additional pressure on some norms or regimes of international law—norms which, in some cases are already finding themselves under siege. The iterative generation of new situations demanding legal development may put pressure on a ‘stagnating’ global legal order; displacement can strengthen the enforcement of international law or even the mapping and harmonisation of regime complexity, but may also contribute to exacerbated power differentials and legitimacy problems if tools are used unilaterally. AI’s role in facilitating destruction will create further blockages to regime modernisation, and increase broader trends towards fragmentation and contestation. While the governance effects of AI may therefore be heterogeneous, if not handled well, they appear to contribute towards legal fragmentation. In response, AI governance regimes may need to adopt strategies to anticipate or be more resilient to likely governance disruption. Specifically, it weakly suggests that some traditional tools, such as multilateral treaty regimes, less fit or less resilient to govern AI. Instead, governance may have to shift towards an array of soft law instruments and iterative, adaptive interpretation.

D. Why and how might changes in the broader global governance architecture, as well as amongst individual AI regimes, affect the prospects, development and efficacy of the ‘regime complex’ for AI?

In Chapter 6, I sought to answer this question, by drawing on regime complex theory to provide a theoretical perspective on the institutional ecologies and governance architectures within which AI systems will be embedded. I discussed the background of regime theory and the concept of a regime complex.

Regime complexity has diverse theorised consequences: I then explored the rich previous scholarship on the normative, institutional, and political effects of fragmentation (or ‘decentralisation’) on a governance regime’s functioning, exploring both critiques and defences.

The regime complexity lens grounds a research program into the origins, topology, evolution, consequences, and strategies for AI governance. I argued that regime complexity provides a conceptual lens that enables one to explore five key aspects of a governance system for AI: its *origins* (including its purpose; its viability given prevailing interests and norms; its possible designs); its *topology* at any moment in time; its possible trends of *evolution* over time; the *consequences* of different forms of regime organisation for its legal coherence, effectiveness, or legitimacy; and finally possible *strategies* for mitigating the adverse effects of different types of regime complexes.

This lens provides important strategies, whether one expects continued fragmentation or integration in the AI regime complex. I argued that AI governance initiatives should better engage with the implications of regime complexity, because it can offer productive insights across different future trajectories of AI regime organisation. In a decentralised regime, active efforts must be undertaken to achieve regime harmonisation and modernisation. Even if a centralised institution is viable, its efficacy would depend sensitively on a set of institutional design choices. While it may not be easy or possible to deliberately shift or alter the macro-scale trajectory of a regime complex, whichever form it does take, some strategies will be possible and necessary in order to improve coherence and coordination within the regime.

E. What insights can these three conceptual frameworks provide in exploring the prospects and dynamics of the emerging AI governance regime complex?

In response to this question, I brought the three lenses together in Chapter 7, in order to examine their implications for a range of cases and questions in AI governance, including emerging international security governance architectures for military AI applications. Accordingly, I explored 5 ‘steps’ to AI governance: regime origins; –topology; –evolution; –consequences; and –strategies.

When analysing AI regime origins, the lenses can provide practical insights into regime purpose, viability, and design. I argued these lenses provided valuable insights in terms of AI regime *origins*. The lens of sociotechnical change can provide key insights into a regime’s purpose and design—whether or not that regime responds to governance rationales, and how it constitutes the governance target of AI. The lens of regime complexity can inform a consideration of the

regime's *viability*. This can be done from an outside view—drawing on historical comparisons such as the nuclear non-proliferation regime (articulated in *Paper [I]*), which suggest that states may at times create surprisingly robust regimes even for extremely potent strategic technologies. Such an analysis can also draw on regime theory to explore when or how states might find sufficient common interests in restricting certain AI applications (such as in military use). It can also draw on the processes of *norm-shaping* by various actors and coalitions. Finally, regime theory as well as the lens of governance disruption can help illuminate conditions under which various regime *designs* are more viable, effective (e.g. full bans versus partial bans) or resilient to further technological change. Curiously, this suggests that while 'full bans' might be more resilient to governance disruption than partial bans, this is only if they can actually 'hold the door' on the technology—which appears potentially challenging given the versatility of AI.

When analysing AI governance topology, regime complexity can provide practical insights into regime complex demographics, organisation, and interactions. In terms of the *topology* of the broader AI regime complex, in order to identify gaps, shortcomings, and latent conflicts in either legal rules, norms, goals, or institutional competencies. This involves examining *demographics* of the AI governance system, its *organisation* (in terms of the density of the institutional network, and whether links relate to norms, goals, impacts or institutional relations); and the *interactions* and outcomes of these linkages (whether leaving gaps; conflictive, cooperative, or synergistic). Finally, these interactions can be examined at the macro, meso, or micro level. This is relevant given that the AI governance regime today remains in a state of relative fragmentation. Analysis of AI governance should in particular consider 'fault lines' or overlaps amongst distinct regulatory instruments—potential conceptual or political tensions which remain latent today but which could find themselves 'activated' or triggered in the wake of later governance disruption produced by advancing AI capabilities.

When analysing AI governance evolution, the regime complexity and governance disruption lenses can provide insights into the broader drivers of fragmentation or integration. In terms of *evolution*, I argued that while we should not attempt prediction, a consideration of the developmental trajectories of the AI regime complex can nonetheless consider the role of various factors. These include both general trends that have been identified in the global governance architecture (institutional density, accretion, power and preference shifts, complexity, and shift towards local governance), as well as the effects of governance disruption itself, which show how AI systems could be a generator of regime fault lines; a shield, patch, cure or accelerator of regime complexity and fragmentation, or a driver of contestation and institutional proliferation.

When analysing the consequences of AI regime organisation, the regime complexity lens can provide insights into the potential functional, institutional and normative consequences or trade-offs that might result from the AI regime complex being centralised or fragmented. Drawing on the analysis of *Paper [IV]*, I explored the potential consequences of—or trade-offs between—fragmentation and centralisation in AI governance. On the basis of debates within regime complex theory, as well as various historical experiences from key international regimes in the areas of environment, trade, and security, my co-authors and I suggested that centralizing and integrating a governance regime can have certain benefits in terms of authority and political power, efficiency and lower participation thresholds. However, it also introduces new risks in

terms of slowness because of longer ‘start-up’ periods, and brittleness because of regulatory capture, or in the face of sudden technological or political change. We argued that, in many cases, the added benefits of either governance approach depend far more sensitively on their exact institutional design. Nonetheless, in this chapter, I argued that a lot of these considerations over the relative merits of centralised or decentralised regimes, might depend sensitively on one’s assumptions or expectations about the pace and magnitude of AI-driven sociotechnical change, the relative need for breadth or depth in governing the AI issue, and the relative need for coordination or orchestration of AI policies.

All three lenses suggest valuable insights in terms of conceptual approach, instrument choice, and instrument design, to ensure that AI governance can achieve efficacy, resilience, and coherence in the face of change. As such, rather than offering a definitive recommendation in favour or against a centralised regime—which may be somewhat moot given a relatively constrained ability to deliberately ‘design’ a regime complex—I argued that, along with considering whether and where AI would require a global regime, an important complementary approach is to consider distinct *strategies* to be implemented, either to manage the interactions of an AI regime complex, or in designing a centralised institution. As such, drawing insights from all three lenses, I proposed a series of indicative strategies or shifts which might help enhance the efficacy, resilience, or coherence of AI governance systems.

Given all this, we can now return to this project’s central question:

- “***How should global governance for artificial intelligence account for change?***”

In response, my thesis in this dissertation has been that—to remain fit for change, global governance for artificial intelligence will have to adopt new strategies in order to effectively track the technology’s sociotechnical impacts, remain resilient to further AI-driven disruption in the tools, norms or broader conditions for governance, and coherently manage the interactions of the AI governance regime complex.

Academic contributions and future research: this project has sought to contribute both conceptual lenses as well as concrete strategies and recommendations to the emerging scholarship on global AI governance. However, it is just one step—and there is so much more research that needs to be done in this area.

For one, much more work is needed to fully explore the analytical insights and limits of these lenses for AI governance. Future scholarship as such could include much more in-depth or granular explorations of AI regimes of any one of these lenses. This could include (1) a much more detailed and comprehensive mapping of the *topology* of the AI regime complex, including all its potential regime interactions such as institutional and norm conflicts; (2) case studies of the relative susceptibility or resilience of different international norms or instruments (such as the human rights framework) to different types of governance disruption; (3) the relative role and prominence of ‘material’ or architectural features versus sociotechnical ‘problem logics’ in shaping (AI) technology’s regulatory texture as a governance target; (4) an examination of the distributive

and political effects of different types of AI-driven governance displacement (i.e. automation of rule creation vs. automation of monitoring); (5) a study of the relative prominence or frequency of the different types of governance development spurred by AI capabilities; (6) more detailed case studies of the degree to which various drivers of global governance fragmentation (e.g. modernity, institutional density; local governance goals) may apply to AI governance. Other work could seek to extend the analytical scope of some of these lenses, for instance by (7) drawing on various foresighting methodologies to improve the ability to anticipate even indirect avenues of sociotechnical change. Future scholarship could also explore in greater detail (8) the dyadic interfaces of the three lenses—for instance, the benefits of combining the sociotechnical change and regime complexity lens, or the regime complexity and governance disruption lenses—which have only been lightly explored here.

Along with having sought to contribute to work on AI governance, this project can also contribute lessons back to the three bodies of scholarship—on law, technology and society; technology-driven legal disruption, and regime complex theory—from which we derived the core lenses.

The attempted contribution of this project to scholarship *on law, regulation and technology* has been to apply the sociotechnical change framework—as derived especially from the work of Lyria Bennett Moses—to analysing questions around governing AI-driven change at the global level. In doing so, it has also sought to develop and extend the account of potential governance *rationales* for governance, as well as the role and balance of material to other features in constituting the texture of a governance *target*. Finally, it has also sketched and articulated six clusters of ‘problem logics’ with their own distinct problem origins, barriers, and highlighted governance logics. In practical terms, while this lens is applied to analyse the emerging regime complex for AI, it is in principle also applicable to other domains of global technology governance. Future work in this area could explore (9) the degree to which the six ‘problem logics’ identified around AI-driven sociotechnical change can also shed light on domestic governance questions, or global governance of other technologies.

The attempted contribution of this project to scholarship on *legal ‘disruption’* has been to further develop a taxonomy of *governance disruption*, and in particular to extend existing scholarship in this area to the international legal level. In doing so, it also sought to apply and integrate a number of frameworks and studies from this area of scholarship—including by Lyria Bennett Moses, Rebecca Crootof, Colin Picker, Roger Brownsword and Ashley Deeks. This framework is interesting, since exploring the dynamics of AI governance can reveal interesting and important lessons about the changing nature of technology governance specifically, and of 21st-century international law more broadly. In turn, future work in extending this connection could (10) explore the degree to which the fifth category of legal development which I introduced—dealing with ‘alterations of the problem portfolio’—provides insights to other contexts, or is redundant or could be subsumed with the other categories.

The attempted contribution of this project to scholarship on *regime complex theory* has been to apply this established framework in a novel domain or issue area (AI); to aim to distil previous regime complexity frameworks into a five-fold framework for analysing a regime complex (emphasizing questions around *origins, topology, evolution, consequences, and strategies*); to

highlight the importance, to debates on regime complexity, of considering the effects of technology-driven governance disruption on trends of global governance fragmentation. At this level, future work could go in a huge number of directions; however, key research may involve (11) a wider range of case studies exploring and testing the intersection of technology-driven governance disruption with regime creation, in order to spur a wider theoretical debate over the increasing role of ‘material-contextual’ factors—along with rationalist (interest-based) and constructivist (norms-based) considerations—in grounding or maintaining global regimes.

We face today the question of AI governance under change. How can we govern a changing technology, in a changing world, using regulatory tools and governance systems that may themselves be left changed? This question is a challenge, a responsibility—and an opportunity. I am certain we will meet it as such.

Bibliography

- Aaken, Anne van. "Is International Law Conducive To Preventing Looming Disasters?" *Global Policy* 7, no. S1 (2016): 81–96. <https://doi.org/10.1111/1758-5899.12303>.
- Abade, Andre S., Paulo Afonso Ferreira, and Flavio de Barros Vidal. "Plant Diseases Recognition on Images Using Convolutional Neural Networks: A Systematic Review." *ArXiv:2009.04365 [Cs]*, September 9, 2020. <http://arxiv.org/abs/2009.04365>.
- Abadi, Martin, and David G. Andersen. "Learning to Protect Communications with Adversarial Neural Cryptography," 2016. <https://arxiv.org/pdf/1610.06918v1.pdf>.
- Abbott, Kenneth W., Jessica F. Green, and Robert O. Keohane. "Organizational Ecology and Institutional Change in Global Governance." *International Organization* 70, no. 2 (ed 2016): 247–77. <https://doi.org/10.1017/S0020818315000338>.
- Abbott, Kenneth W., and Duncan Snidal. "Hard and Soft Law in International Governance." *International Organization* 54, no. 3 (2000): 421–56.
- . "International Regulation without International Government: Improving IO Performance through Orchestration." *The Review of International Organizations* 5, no. 3 (September 1, 2010): 315–44. <https://doi.org/10.1007/s11558-010-9092-3>.
- Acharya, Amitav. "The Future of Global Governance: Fragmentation May Be Inevitable and Creative Global Forum." *Global Governance*, no. 4 (2016): 453–60.
- Achiam, Joshua, and Dario Amodei. "Benchmarking Safe Exploration in Deep Reinforcement Learning," 2019. <https://pdfs.semanticscholar.org/4d0f/6a6ffcd6ab04732ff76420fd9f8a7bb649c3.pdf>.
- Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence. "Outcome Document: First Version of a Draft Text of a Recommendation on the Ethics of Artificial Intelligence." UNESDOC Digital Library, 2020. <https://unesdoc.unesco.org/ark:/48223/pf0000373434>.
- Adamczewski, Tom. "A Shift in Arguments for AI Risk." Fragile Credences, May 25, 2019. <https://fragile-credences.github.io/prioritising-ai/>.
- Adams, Sam, Itmar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, J. Storrs Hall, et al. "Mapping the Landscape of Human-Level Artificial General Intelligence." *AI Magazine* 33, no. 1 (March 15, 2012): 25–42. <https://doi.org/10.1609/aimag.v33i1.2322>.
- Adamson, Adewole S., and H. Gilbert Welch. "Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard." *New England Journal of Medicine* 381, no. 24 (December 12, 2019): 2285–87. <https://doi.org/10.1056/NEJMp1907407>.
- Adler, Emanuel. "The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control." *International Organization* 46, no. 1 (1992): 101–45. <https://doi.org/10.1017/S0020818300001466>.
- Aguirre, Anthony, Gaia Dempsey, Harry Surden, and Peter Bart Reiner. "AI Loyalty: A New Paradigm for Aligning Stakeholder Interests." *University of Colorado Law Legal Studies Research Paper* 20.18 (March 25, 2020). <https://papers.ssrn.com/abstract=3560653>.
- AI Impacts. "Evidence against Current Methods Leading to Human Level Artificial Intelligence." AI Impacts, August 12, 2019. <https://aiimpacts.org/evidence-against-current-methods-leading-to-human-level-artificial-intelligence/>.
- . "Friendly AI as a Global Public Good." AI Impacts, August 8, 2016. <https://aiimpacts.org/friendly-ai-as-a-global-public-good/>.
- . "Penicillin and Syphilis." AI Impacts, February 2, 2015. <https://aiimpacts.org/penicillin-and-syphilis/>.
- AI Now Institute. "AI IN 2018: A YEAR IN REVIEW." AI Now Institute (blog), October 24, 2018. <https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e>.
- Ajder, Henry, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. "The State of DeepFakes: Landscape, Threats, and Impact." DeepTrace Labs, September 2019. <https://deeptracelabs.com/mapping-the-deepfake-landscape/>.
- Ajunwa, Ifeoma. "Automated Employment Discrimination." *SSRN Electronic Journal*, 2019. <https://doi.org/10.2139/ssrn.3437631>.
- Aker, Jenny C., and Isaac M. Mbiti. "Mobile Phones and Economic Development in Africa." *The Journal of Economic Perspectives* 24, no. 3 (2010): 207–32.
- Akers, John, Gagan Bansal, Gabriel Cadamuro, Christine Chen, Quanze Chen, Lucy Lin, Phoebe Mulcaire, et al. "Technology-Enabled Disinformation: Summary, Lessons, and Recommendations." *ArXiv:1812.09383 [Cs]*, January 3, 2019. <https://arxiv.org/abs/1812.09383>.
- Alarie, Benjamin. "The Path of the Law: Towards Legal Singularity." *University of Toronto Law Journal* 66, no. 4 (January 1, 2016): 443–55. <https://doi.org/10.3138/UTLJ.4008>.
- Alarie, Benjamin, Anthony Niblett, and Albert H. Yoon. "Law in the Future." *University of Toronto Law Journal*, November 7, 2016. <https://doi.org/10.3138/UTLJ.4005>.
- Albright, David, Sarah Burkhard, and Allison Lach. "Commercial Satellite Imagery Analysis for Countering Nuclear Proliferation." *Annual Review of Earth and Planetary Sciences* 46, no. 1 (2018): 99–121. <https://doi.org/10.1146/annurev-earth-063016-015704>.
- Alcorn, Michael A., Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects." *ArXiv:1811.11553 [Cs]*, April 18, 2019. <http://arxiv.org/abs/1811.11553>.
- Alesina, Alberto, Paola Giuliano, and Nathan Nunn. "On the Origins of Gender Roles: Women and the Plough."

- Quarterly Journal of Economics* 128, no. 2 (2013): 469–530.
- Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoțiuc-Pietro, and Vasileios Lampos. “Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective.” *PeerJ Computer Science* 2 (October 24, 2016): e93. <https://doi.org/10.7717/peerj-cs.93>.
- Allen, Greg. “Understanding AI Technology.” JAIC, April 2020.
- Allen, Greg, and Taniel Chan. “Artificial Intelligence and National Security.” Belfer Center Study. Harvard Kennedy School - Belfer Center for Science and International Affairs, July 2017. <http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf>.
- Allenby, Braden. “Are New Technologies Undermining the Laws of War?” *Bulletin of the Atomic Scientists* 70, no. 1 (2014): 21–31. <https://doi.org/10.1177/0096340213516741>.
- Allison, Graham. “Nuclear Disorder.” *Foreign Affairs*, January 1, 2010. <https://www.foreignaffairs.com/articles/pakistan/2010-01-01/nuclear-disorder>.
- AlQuraishi, Mohammed. “AlphaFold @ CASP13: ‘What Just Happened?’” *Some Thoughts on a Mysterious Universe* (blog), December 9, 2018. <https://moalquraishi.wordpress.com/2018/12/09/alpha-fold-casp13-what-just-happened/>.
- Alschner, Wolfgang, Joost Pauwelyn, and Sergio Puig. “The Data-Driven Future of International Economic Law.” *Journal of International Economic Law* 20, no. 2 (June 2017): 217–31. <https://doi.org/10.1093/jiel/jgx020>.
- Alschner, Wolfgang, and Dmitriy Skougarevskiy. “Can Robots Write Treaties? Using Recurrent Neural Networks to Draft International Investment Agreements.” In *JURIX: Legal Knowledge and Information Systems*, edited by F. Bex and S. Villata, 114–19. IOS Press, 2016. <https://papers.ssrn.com/abstract=2984935>.
- . “Towards an Automated Production of Legal Texts Using Recurrent Neural Networks.” In *Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law*, 229–232. ICAIL ’17. London, United Kingdom: Association for Computing Machinery, 2017. <https://doi.org/10.1145/3086512.3086536>.
- Alter, Adam. *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Reprint edition. New York: Penguin Books, 2018.
- Alter, Karen J. “The Future of International Law.” *ICourts Working Paper Series* 101 (August 2017). <https://papers.ssrn.com/abstract=3015177>.
- Alter, Karen J., James Thuo Gathii, and Laurence R. Helfer. “Backlash Against International Courts in West, East and Southern Africa: Causes and Consequences.” *European Journal of International Law* 27, no. 2 (2016). <https://doi.org/10.2139/ssrn.2591837>.
- Alter, Karen J., and Sophie Meunier. “The Politics of International Regime Complexity.” *Perspectives on Politics* 7, no. 1 (March 2009): 13–24. <https://doi.org/10.1017/S1537592709009033>.
- Alter, Karen J., and Kal Raustiala. “The Rise of International Regime Complexity.” *Annual Review of Law and Social Science* 14, no. 1 (2018): 329–49. <https://doi.org/10.1146/annurev-lawsocsci-101317-030830>.
- Altmann, Jürgen. “Preventive Arms Control for Uninhabited Military Vehicles.” In *Ethics and Robotics*, edited by R. Capurro and M. Nagengberg, 69–82. AKA Verlag Heidelberg, 2009.
- Altmann, Jürgen, and Frank Sauer. “Autonomous Weapon Systems and Strategic Stability.” *Survival* 59, no. 5 (September 3, 2017): 117–42. <https://doi.org/10.1080/00396338.2017.1375263>.
- Alvarez, José E. *International Organizations As Law-Makers*. Oxford Monographs in International Law. Oxford, New York: Oxford University Press, 2005.
- Amazon. “We Are Implementing a One-Year Moratorium on Police Use of Rekognition.” Day One Blog, June 10, 2020. <https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition>.
- Amnesty International, and Access Now. “The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems,” May 2018. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.
- Amodei, Dario, and Jack Clark. “Faulty Reward Functions in the Wild.” *OpenAI* (blog), 2016. <https://openai.com/blog/faulty-reward-functions/>.
- Amodei, Dario, and Danny Hernandez. “AI and Compute.” OpenAI Blog, May 16, 2018. <https://blog.openai.com/ai-and-compute/>.
- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. “Concrete Problems in AI Safety.” *ArXiv:1606.06565 [Cs]*, June 21, 2016. <http://arxiv.org/abs/1606.06565>.
- Andersen, Ross. “The Panopticon Is Already Here.” *The Atlantic*, September 2020. <https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/>.
- Anderson, James M., Nidhi Kalra, Karlyn Stanley, Paul Sorensen, Constantine Samaras, and Tobi A. Oluwatola. “Autonomous Vehicle Technology: A Guide for Policymakers.” Product Page. RAND Corporation, 2016. https://www.rand.org/pubs/research_reports/RR443-2.html.
- Anderson, Kenneth, and Matthew C. Waxman. “Law and Ethics for Autonomous Weapon Systems: Why a Ban Won’t Work and How the Laws of War Can.” *Law and Ethics for Autonomous Weapon Systems*, 2013.
- Antarctic Treaty, 402 UNTS 71 § (1959).
- Anton, Don. “‘Treaty Congestion’ in International Environmental Law.” In *Routledge Handbook of International Environmental Law*, edited by Shawkat Alam, Jahid Hossain Bhuiyan, Tareq M.R. Chowdhury, and Erika J. Techera. Routledge, 2012. <https://www.taylorfrancis.com/books/9780203093474>.
- Arkin, Ronald. *Governing Lethal Behavior in Autonomous Robots*. 1 edition. Boca Raton: Routledge, 2009.
- Arkin, Ronald, Leslie Kaelbling, Stuart Russel, Dorsa Sadigh, Paul Scharre, Bart Selman, and Toby Walsh. “A Path Towards Reasonable Autonomous Weapons Regulation.” IEEE Spectrum: Technology, Engineering, and Science News, October 21, 2019. <https://spectrum.ieee.org/automation robotics/artificial-intelligence/a-path-towards-reasonable-autonomous-weapons-regulation>.

- Armstrong, Stuart, and Kaj Sotala. "How We're Predicting AI-Or Failing To." In *Beyond AI: Artificial Dreams*, edited by Jan Romportl, Pavel Irčing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia, 2012. <https://intelligence.org/files/PredictingAI.pdf>.
- Armstrong, Stuart, Kaj Sotala, and Sean S. OhEigearaigh. "The Errors, Insights and Lessons of Famous AI Predictions – and What They Mean for the Future." *Journal Of Experimental & Theoretical Artificial Intelligence* 26, no. 3 (2014). <https://doi.org/10/gghsn5>.
- Arnett, Eric H. "Science, Technology, and Arms Control." In *Encyclopedia of Arms Control and Disarmament*, edited by Richard Dean Burns, 1:477–90. Charles Scribner's Sons, 1993.
- Arquilla, John, and David Ronfeldt. "Cyberwar Is Coming!" In *Athena's Camp: Preparing for Conflict in the Information Age*, 1992, 1995–1996.
- Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. "Should We Treat Data as Labor? Moving beyond 'Free.'" *AEA Papers and Proceedings* 108 (May 2018): 38–42. <https://doi.org/10.1257/pandp.20181003>.
- Article 36 - New weapons - Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) (1977).
- Ashurt, Carolyn, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. "A Guide to Writing the NeurIPS Impact Statement." Medium, May 13, 2020. https://medium.com/@operations_18894/a-guide-to-writing-the-neurips-impact-statement-4293b723f832.
- Askell, Amanda, Miles Brundage, and Gillian Hadfield. "The Role of Cooperation in Responsible AI Development," July 10, 2019, 23.
- Asselt, Harro van. *The Fragmentation of Global Climate Governance: Consequences and Management of Regime Interactions*. Edward Elgar Publishing, 2014.
- Atherton, Kelsey. "DARPA Wants Wargame AI To Never Fight Fair." *Breaking Defense* (blog), August 18, 2020. <https://breakingdefense.com/2020/08/darpa-wants-wargame-ai-to-never-fight-fair/>.
- Auslin, Michael. "Can the Pentagon Win the AI Arms Race?" *Foreign Affairs*, October 19, 2018. <https://www.foreignaffairs.com/articles/united-states/2018-10-19/can-pentagon-win-ai-arms-race>.
- Austin, Lisa. "We Must Not Treat Data like a Natural Resource." *The Globe and Mail*, July 9, 2018. <https://www.theglobeandmail.com/opinion/article-we-must-not-treat-data-like-a-natural-resource/>.
- Australia. "Australia's System of Control and Applications for Autonomous Weapon Systems." Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, March 20, 2019. [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/39A4B669B8AC2111C12583C1005F73CF/\\$file/CC_W_GGE.1_2019_WP.2_final.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/39A4B669B8AC2111C12583C1005F73CF/$file/CC_W_GGE.1_2019_WP.2_final.pdf).
- Avin, Shahar. "Exploring Artificial Intelligence Futures." *AIHumanities*, January 17, 2019. http://aihumanities.org/en/journals/journal-of-aih-list/?board_name=Enjournal&order_by=fn_pid&order_type=desc&list_type=list&vid=15.
- Avin, Shahar, and S. M. Amadae. "Autonomy and Machine Learning at the Interface of Nuclear Weapons, Computers and People." In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, edited by V. Boulainin. Stockholm International Peace Research Institute, 2019. <https://doi.org/10.17863/CAM.44758>.
- Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563, no. 7729 (November 2018): 59. <https://doi.org/10.1038/s41586-018-0637-6>.
- Axon AI & Policing Technology Ethics Board. "First Report of the Axon AI & Policing Technology Ethics Board." Axon AI & Policing Technology Ethics Board, June 2019. https://static1.squarespace.com/static/58a33e881b631bc60d4f8b31/t/5d13d7e1990c4f00014c0aeb/1561581540954/Axon_Ethics_Board_First_Report.pdf.
- Ayoub, Kareem, and Kenneth Payne. "Strategy in the Age of Artificial Intelligence." *Journal of Strategic Studies* 39, no. 5–6 (September 18, 2016): 793–819. <https://doi.org/10.1080/01402390.2015.1088838>.
- Babuta, Alexander, Marion Oswald, and Ardi Janjeva. "Artificial Intelligence and UK National Security: Policy Considerations." Royal United Services Institute for Defence and Security Studies, April 2020. https://rusi.org/sites/default/files/ai_national_security_final_web_version.pdf.
- Baccaro, Lucio, and Valentina Mele. "Pathology of Path Dependency? The ILO and the Challenge of New Governance." *ILR Review* 65, no. 2 (April 2012): 195–224. <https://doi.org/10.1177/001979391206500201>.
- Bahçecik, Şerif Onur. "Civil Society Responds to the AWS: Growing Activist Networks and Shifting Frames." *Global Policy* 0, no. 0 (2019). <https://doi.org/10.1111/1758-5899.12671>.
- Bakkar, Nadine, Tina Koválik, Ileana Lorenzini, Scott Spangler, Alix Lacoste, Kyle Sponaugle, Philip Ferrante, Elenee Argentinis, Rita Sattler, and Robert Bowser. "Artificial Intelligence in Neurodegenerative Disease Research: Use of IBM Watson to Identify Additional RNA-Binding Proteins Altered in Amyotrophic Lateral Sclerosis." *Acta Neuropathologica* 135, no. 2 (2018): 227–47. <https://doi.org/10.1007/s00401-017-1785-8>.
- Balkin, Jack M. "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *University of California Law Review* 51 (2018): 1149–1210. <https://doi.org/10.2139/ssrn.3038939>.
- . "The Path of Robotics Law." *California Law Review Circuit* 6 (June 2015): 17.
- Ballard, Stephanie, and Ryan Calo. "Taking Futures Seriously: Forecasting as Method in Robotics Law and Policy," 22. Miami, 2019. https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/Calo_Taking-Futures-Seriously.pdf.
- Baratta, Joseph Preston. *The Politics of World Federation: From World Federalism to Global Governance*. Greenwood Publishing Group, 2004.
- . "Was the Baruch Plan a Proposal of World Government?" *The International History Review* 7, no. 4 (November 1, 1985): 592–621. <https://doi.org/10.1080/07075332.1985.9640394>.
- Barlow, John Perry. "A Declaration of the Independence of Cyberspace." Electronic Frontier Foundation,

- February 8, 1996. <https://www.eff.org/cyberspace-independence>.
- Barnes, Julian E., and Josh Chin. "The New Arms Race in AI." *Wall Street Journal*, March 2, 2018, sec. Life. <https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261>.
- Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review* 671 (2016). <https://papers.ssrn.com/abstract=2477899>.
- Barrett, Scott. *Why Cooperate?: The Incentive to Supply Global Public Goods*. Oxford University Press, 2007. <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199211890.001.0001/acprof-9780199211890>.
- Bartel, F. "Surviving the Years of Grace: The Atomic Bomb and the Specter of World Government, 1945–1950." *Diplomatic History* 39, no. 2 (April 1, 2015): 275–302. <https://doi.org/10.1093/dh/dhu005>.
- Bassiouni, M. Cherif. "International Crimes: Jus Cogens and Obligatio Erga Omnes." *Law and Contemporary Problems* 59, no. 4 (October 1, 1996): 63–74.
- Basu, Arindrajit, and Justin Sherman. "Two New Democratic Coalitions on 5G and AI Technologies." *Lawfare*, August 6, 2020. <https://www.lawfareblog.com/two-new-democratic-coalitions-5g-and-ai-technologies>.
- Baum, S.D., Ben Goertzel, and Ted G. Goertzel. "How Long until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting & Social Change* 78 (2011): 185–95.
- Baum, Seth D. "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy." Working Paper. Global Catastrophic Risk Institute, November 12, 2017. <https://papers.ssrn.com/abstract=3070741>.
- . "Medium-Term Artificial Intelligence and Society." *Information* 11, no. 6 (June 2020): 290. <https://doi.org/10.3390/info11060290>.
- . "Reconciliation between Factions Focused on Near-Term and Long-Term Artificial Intelligence." *AI & SOCIETY* 33, no. 4 (2018): 565–72. <https://doi.org/10.1007/s00146-017-0734-3>.
- Baum, Seth D., Stuart Armstrong, Timoteus Ekenstedt, Olle Häggström, Robin Hanson, Karin Kuhlemann, Matthijs M. Maas, et al. "Long-Term Trajectories of Human Civilization." *Foresight* 21, no. 1 (February 11, 2019): 53–83. <https://doi.org/10.1108/FS-04-2018-0037>.
- Bayern, Shawn. "The Implications of Modern Business–Entity Law for the Regulation of Autonomous Systems." *European Journal of Risk Regulation* 7, no. 2 (June 2016): 297–309. <https://doi.org/10.1017/S1867299X00005729>.
- Bayern, Shawn, Thomas Burri, Thomas D. Grant, Daniel M. Häusermann, Florian Mösllein, and Richard Williams. "Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators." *Hastings Science and Technology Law Journal* 9, no. 2 (2017): 135–62.
- Beauchamp, Tom L., and James F. Childress. *Principles of Biomedical Ethics*. Oxford University Press, 2012.
- Beaumont, Guillaume, Kevin Kalomeni, Malcolm Campbell-Verduyn, Marc Lenglet, Serena Natile, Marielle Papin, Daiyi Rodima-Taylor, Arthur Silve, and Falin Zhang. "Global Regulations for a Digital Economy: Between New and Old Challenges." *Global Policy* n/a, no. n/a (2020). <https://doi.org/10.1111/1758-5899.12823>.
- Beckstead, Nick. "Differential Technological Development: Some Early Thinking." The GiveWell Blog, September 30, 2015. <http://blog.givewell.org/2015/09/30/differential-technological-development-some-early-thinking/>.
- Beduschi, Ana. "International Migration Management in the Age of Artificial Intelligence." *Migration Studies*, February 10, 2020. <https://doi.org/10.1093/migration/mnaa003>.
- Beekman, Madeleine, and Tanya Latty. "Brainless but Multi-Headed: Decision Making by the Acellular Slime Mould *Physarum Polycephalum*." *Journal of Molecular Biology* 427, no. 23 (November 2015): 3734–43. <https://doi.org/10.1016/j.jmb.2015.07.007>.
- Belfield, Haydn. "Activism by the AI Community: Analysing Recent Achievements and Future Prospects." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 15–21. New York NY USA: ACM, 2020. <https://doi.org/10.1145/3375627.3375814>.
- Bellemare, Marc G., Will Dabney, and Remi Munos. "A Distributional Perspective on Reinforcement Learning," 2017, 10.
- Benaich, Nathan. "AI Has Disappointed on Covid." *Financial Times*, September 20, 2020. <https://www.ft.com/content/0aafc2de-f46d-4646-acfd-4ed7a7f6fea>.
- Benkler, Yochai, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford, New York: Oxford University Press, 2018.
- Bennett Moses, Lyria. "Agents of Change: How the Law 'Copes' with Technological Change." *Griffith Law Review*, Special Issue: The Laws of Technology and the Technology of Law, 20, no. 4 (January 1, 2011): 763–94. <https://doi.org/10.1080/10383441.2011.10854720>.
- . "Recurring Dilemmas: The Law's Race to Keep Up With Technological Change." *University of New South Wales Faculty of Law Research Series* 21 (April 11, 2007). <http://www.austlii.edu.au/journals/UNSWLRS/2007/21.html>.
- . "Regulating in the Face of Sociotechnical Change." In *The Oxford Handbook of Law, Regulation, and Technology*, edited by Roger Brownsword, Eloise Scotford, and Karen Yeung, 573–96, 2017. <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49>.
- . "Why Have a Theory of Law and Technological Change?" *Minnesota Journal of Law, Science & Technology* 8, no. 2 (2007): 589–606.
- Benvenisti, Eyal. "Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?" *European Journal of International Law* 29, no. 1 (July 23, 2018): 9–82. <https://doi.org/10.1093/ejil/chy031>.
- Benvenisti, Eyal, and George W. Downs. "Comment on Nico Krisch, 'The Decay of Consent: International Law in an Age of Global Public Goods.'" *AJIL Unbound* 108 (ed 2014): 1–3. <https://doi.org/10.1017/S2398772300001744>.
- Benvenisti, Eyal, and George W. Downs. "The Empire's New Clothes: Political Economy and the Fragmentation of International Law." *Stanford Law Review* 60 (2007): 595–632.

- Bergal, Asya. "2019 Recent Trends in GPU Price per FLOPS." AI Impacts, March 25, 2020. <https://aiimprints.org/2019-recent-trends-in-gpu-price-per-flops/>.
- Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*, July 2, 2018, 004912411878253. <https://doi.org/10.1177/0049124118782533>.
- Berman, Paul Schiff. "A Pluralist Approach to International Law." *The Yale Journal of International Law* 32 (2007): 301–29.
- Betts, Alexander. "Regime Complexity and International Organizations: UNHCR as a Challenged Institution." *Global Governance* 19, no. 1 (2013): 69–81.
- Beyleveld, Deryck, and Roger Brownsword. "Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning." *Law, Innovation & Technology* 4, no. 1 (July 2012): 35–65.
- Bhatnagar, Sankalp, Anna Alexandrova, Shahar Avin, Stephen Cave, Lucy Cheke, Matthew Crosby, Jan Feyereisl, et al. "Mapping Intelligence: Requirements and Possibilities." In *Philosophy and Theory of Artificial Intelligence 2017*, edited by Vincent C. Müller, 117–35. Studies in Applied Philosophy, Epistemology and Rational Ethics. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-96448-5_13.
- Bhuta, Nehal, Susanne Beck, Robin Geiss, Hin-Yan Liu, and Klaus Kress, eds. *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press, 2016.
- Bianchi, Andrea. *International Law Theories: An Inquiry into Different Ways of Thinking*. Oxford, New York: Oxford University Press, 2016.
- Biddle, Sam. "ICE's New York Office Uses a Rigged Algorithm to Keep Virtually All Arrestees in Detention. The ACLU Says It's Unconstitutional." *The Intercept* (blog), March 2, 2020. <https://theintercept.com/2020/03/02/ice-algorithm-bias-detention-aclu-lawsuit/>.
- Biermann, Frank. *A World Environment Organization: Solution or Threat for Effective International Environmental Governance?* Edited by Steffen Bauer. 1 edition. Aldershot, Hants, England ; Burlington, VT: Routledge, 2005.
- Biermann, Frank, Philipp Pattberg, Harro van Asselt, and Fariborz Zelli. "The Fragmentation of Global Governance Architectures: A Framework for Analysis." *Global Environmental Politics* 9, no. 4 (October 14, 2009): 14–40. <https://doi.org/10.1162/glep.2009.9.4.14>.
- Bietti, Elettra. "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy." In *Proceedings of ACM FAT* Conference*, 2019. <https://papers.ssrn.com/abstract=3513182>.
- Bird, Eleanor, Jasmin Fox-Skelly, Nicola Jenner, Ruth Larbey, Emma Weitkamp, and Alan Winfield. "The Ethics of Artificial Intelligence: Issues and Initiatives." Scientific Foresight Unit (STOA), EPRS | European Parliamentary Research Service, March 2020. [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU\(2020\)634452_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf).
- Bird, J., and P. Layzell. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, 2:1836–41. Honolulu, HI, USA: IEEE, 2002. <https://doi.org/10.1109/CEC.2002.1004522>.
- Bisin, Alberto. "The Evolution of Value Systems: A Review Essay on Ian Morris's Foragers, Farmers, and Fossil Fuels." *Journal of Economic Literature* 55, no. 3 (September 2017): 1122–35. <https://doi.org/10.1257/jel.20151352>.
- Blain, Loz. "South Korea's Autonomous Robot Gun Turrets: Deadly from Kilometers Away," 2010. <http://newatlas.com/korea-dodam-super-aegis-autonomos-robot-gun-turret/17198/>.
- Blalock, Davis, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Gutttag. "What Is the State of Neural Network Pruning?" In *Proceedings of Machine Learning and Systems 2020 (MLSys 2020)*, 18, 2020. <https://proceedings.mlsys.org/book/296.pdf>.
- Blank, Joshua D., and Leigh Osofsky. "Automated Legal Guidance." *Cornell Law Review* 106 (Forthcoming 2021). <https://papers.ssrn.com/abstract=3546889>.
- Blenkinsop, Philip. "U.S. Trade Offensive Takes out WTO as Global Arbiter." *Reuters*, December 10, 2019. <https://www.reuters.com/article/us-trade-wto-idUSKBN1YE0YE>.
- Bloom, Nicholas. "Are Ideas Getting Harder to Find?" *THE AMERICAN ECONOMIC REVIEW* 110, no. 4 (2020): 1104–44.
- Bode, Ingvild, and Hendrik Huelss. "Autonomous Weapons Systems and Changing Norms in International Relations." *Review of International Studies* 44, no. 3 (February 19, 2018): 393–413.
- Boese, Wade. "U.S. Withdraws From ABM Treaty; Global Response Muted." Arms Control Association, August 2002. <https://www.armscontrol.org/act/2002-07/news/us-withdraws-abm-treaty-global-response-muted>.
- Bolsover, Gillian, and Philip Howard. "Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda." *Big Data* 5, no. 4 (December 2017): 273–76. <https://doi.org/10.1089/big.2017.29024.cpr>.
- Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. "The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars." *Proceedings of the IEEE* 107, no. 3 (March 2019): 502–4. <https://doi.org/10.1109/JPROC.2019.2897447>.
- Bontrager, Philip, Aditi Roy, Julian Togelius, Nasir Memon, and Arun Ross. "DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution." *ArXiv:1705.07386 [Cs]*, May 20, 2017. <http://arxiv.org/abs/1705.07386>.
- Borghard, Erica D., and Shawn W. Lonergan. "Why Are There No Cyber Arms Control Agreements?" Council on Foreign Relations, January 16, 2018. <https://www.cfr.org/blog/why-are-there-no-cyber-arms-control-agreements>.
- Borning, Alan. "Computer System Reliability and Nuclear War." *Communications of the ACM* 30, no. 2 (February 1, 1987): 112–131. <https://doi.org/10.1145/12527.12528>.
- Borrie, John. "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies." UNIDIR Resources. UNIDIR, 2016.

- <http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf>.
- Bosselut, Antoine, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction." *ArXiv:1906.05317 [Cs]*, June 14, 2019. <http://arxiv.org/abs/1906.05317>.
- Bostrom, Nick. "Existential Risk Prevention as a Global Priority." *Global Policy* 4, no. 1 (2013): 15–31.
- . "Information Hazards: A Typology of Potential Harms from Knowledge." *Review of Contemporary Philosophy* 10 (2011): 44–79.
- . "Strategic Implications of Openness in AI Development." *Global Policy*, February 1, 2017, 135–48. <https://doi.org/10.1111/1758-5899.12403>.
- . *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
- . "The Vulnerable World Hypothesis." *Global Policy*, September 6, 2019, 1758-5899.12718. <https://doi.org/10.1111/1758-5899.12718>.
- Bostrom, Nick, and Milan M. Cirkovic. *Global Catastrophic Risks*. 1 edition. Oxford; New York: Oxford University Press, 2008.
- Bostrom, Nick, Allan Dafoe, and Carrick Flynn. "Public Policy and Superintelligent AI: A Vector Field Approach." In *Ethics of Artificial Intelligence*, edited by S.M. Liao. Oxford University Press, 2019. <http://www.nickbostrom.com/papers/aipolicy.pdf>.
- Boulanin, Vincent, ed. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Vol. I: Euro-Atlantic Perspectives. SIPRI, 2019. <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>.
- Boulanin, Vincent, Lora Saalman, Petr Topychkanov, Fei Su, and Moa Peldán Carlsson. "Artificial Intelligence, Strategic Stability and Nuclear Risk." SIPRI, June 2020. https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.
- Boulanin, Vincent, and Maaika Verbruggen. "Mapping the Development of Autonomy in Weapon Systems." Stockholm International Peace Research Institute, 2017. https://www.sipri.org/sites/default/files/2017-11/sipri-report_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.
- Boutin, Berenice. "Technologies for International Law & International Law for Technologies." *Groningen Journal of International Law* (blog), October 22, 2018. <https://grojil.org/2018/10/22/technologies-for-international-law-international-law-for-technologies/>.
- Box, George E. P. "Science and Statistics." *Journal of the American Statistical Association* 71, no. 356 (December 1, 1976): 791–99. <https://doi.org/10.1080/01621459.1976.10480949>.
- Boyle, Alan, and Christine Chinkin. *The Making of International Law*. Foundations of Public International Law. Oxford, New York: Oxford University Press, 2007.
- Braithwaite, John, and Peter Drahos. *Global Business Regulation*. Cambridge University Press, 2000.
- Branch, Jordan. "What's in a Name? Metaphors and Cybersecurity." *International Organization*, 2020, 1–32. <https://doi.org/10.1017/S002081832000051X>.
- Branwen, Gwern. "Complexity No Bar to AI," June 1, 2014. <https://www.gwern.net/Complexity-vs-AI>.
- . "GPT-3 Creative Fiction," June 19, 2020. <https://www.gwern.net/GPT-3>.
- . "On GPT-3 - Meta-Learning, Scaling, Implications, And Deep Theory," May 2020. <https://www.gwern.net/newsletter/2020/05>.
- Bresnahan, Timothy, and Manuel Trajtenberg. "General Purpose Technologies 'Engines of Growth'?" *Journal of Econometrics* 65, no. 1 (1995): 83–108.
- Brief, Arthur P., Janet M. Dukerich, Paul R. Brown, and Joan F. Brett. "What's Wrong with the Treadway Commission Report? Experimental Analyses of the Effects of Personal Values and Codes of Conduct on Fraudulent Financial Reporting." *Journal of Business Ethics* 15, no. 2 (February 1, 1996): 183–98. <https://doi.org/10.1007/BF00705586>.
- Brockmann, Kolja. "Challenges to Multilateral Export Controls: The Case for Inter-Regime Dialogue and Coordination." SIPRI, December 2019. <https://www.sipri.org/publications/2019/other-publications/challenges-multilateral-export-controls-case-inter-regime-dialogue-and-coordination>.
- Brodsky, Jessica S. "Autonomous Vehicle Regulation: How an Uncertain Legal Landscape May Hit the Brakes on Self-Driving Cars Cyberlaw and Venture Law." *Berkeley Technology Law Journal* 31 (2016): 851–78.
- Bromley, Mark, and Giovanna Maletta. "The Challenge of Software and Technology Transfers to Non-Proliferation Efforts: Implementing and Complying with Export Controls." SIPRI, April 2018. <https://www.sipri.org/publications/2018/other-publications/challenge-software-and-technology-transfers-non-proliferation-efforts-implementing-and-complying>.
- Brooks, Michael. "Forever 20 Years Away: Will We Ever Have a Working Nuclear Fusion Reactor?" *New Statesman*, November 6, 2014. <https://www.newstatesman.com/scitech/2014/11/forever-20-years-away-will-we-ever-have-working-nuclear-fusion-reactor>.
- Brooks, Rodney. "A Better Lesson," March 19, 2019. <https://rodneybrooks.com/a-better-lesson/>.
- Brooks, Rosa. "War Everywhere: Rights, National Security Law, and the Law of Armed Conflict in the Age of Terror." *University of Pennsylvania Law Review* 153 (January 1, 2004): 675–761.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models Are Few-Shot Learners." *ArXiv:2005.14165 [Cs]*, May 28, 2020. <http://arxiv.org/abs/2005.14165>.
- Brownsword, Roger. "From Erewhon to AlphaGo: For the Sake of Human Dignity, Should We Destroy the Machines?" *Law, Innovation and Technology* 9, no. 1 (January 2, 2017): 117–53. <https://doi.org/10.1080/17579961.2017.1303927>.
- . "In the Year 2061: From Law to Technological Management." *Law, Innovation and Technology* 7, no. 1 (January 2, 2015): 1–51. <https://doi.org/10.1080/17579961.2015.1052642>.
- . "Law and Technology: Two Modes of Disruption, Three Legal Mind-Sets, and the Big Picture of Regulatory Responsibilities." *Indian Journal of Law and Technology* 14 (2018): 1–40.
- . "Law Disrupted, Law Re-Imagined, Law Re-Invented." *Technology and Regulation*, May 20, 2019, 10–30. <https://doi.org/10.26116/techreg.2019.002>.

- . *Law, Technology and Society: Re-Imagining the Regulatory Environment*. 1 edition. Abingdon, Oxon ; New York, NY: Routledge, 2019.
- . "Technological Management and the Rule of Law." *Law, Innovation and Technology* 8, no. 1 (January 2, 2016): 100–140. <https://doi.org/10.1080/17579961.2016.1161891>.
- Brownsword, Roger, Eloise Scotford, and Karen Yeung. "Law, Regulation, and Technology: The Field, Frame, and Focal Questions." In *The Oxford Handbook of Law, Regulation and Technology*, edited by Roger Brownsword, Eloise Scotford, and Karen Yeung, Vol. 1. Oxford University Press, 2017. <https://doi.org/10.1093/oxfordhb/9780199680832.013.1>.
- Brundage, Miles. "Artificial Intelligence and Responsible Innovation." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 543–54. Synthese Library 376. Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-26485-1_32.
- . "Modeling Progress in AI." *ArXiv:1512.05849 [Cs]*, December 17, 2015. <http://arxiv.org/abs/1512.05849>.
- Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *ArXiv:1802.07228 [Cs]*, February 20, 2018. <http://arxiv.org/abs/1802.07228>.
- Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *ArXiv:2004.07213 [Cs]*, April 15, 2020. <http://arxiv.org/abs/2004.07213>.
- Brunn, Axel. *Are Filter Bubbles Real?* 1 edition. Medford, MA: Polity, 2019.
- Bryant, Ross. "Google's AI Becomes First Non-Human to Qualify as a Driver." *Dezeen*, February 12, 2016. <https://www.dezeen.com/2016/02/12/google-self-driving-car-artificial-intelligence-system-recognised-as-driver-usa/>.
- Bryson, Joanna J., Mihailis E. Diamantis, and Thomas D. Grant. "Of, for, and by the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25, no. 3 (September 1, 2017): 273–91. <https://doi.org/10.1007/s10506-017-9214-9>.
- Buchanan, Ben. "The AI Triad and What It Means for National Security Strategy." Center for Security and Emerging Technology, August 2020. <https://cset.georgetown.edu/research/the-ai-triad-and-what-it-means-for-national-security-strategy/>.
- Bud, Robert. *Penicillin: Triumph and Tragedy*. Oxford, New York: Oxford University Press, 2009.
- Buitenhuis, Miriam C. "Towards Intelligent Regulation of Artificial Intelligence." *European Journal of Risk Regulation* 10, no. 1 (March 2019): 41–59. <https://doi.org/10.1017/err.2019.8>.
- Bull, Hedley. *The Anarchical Society: A Study of Order in World Politics*. New York: Columbia University Press, 1977.
- Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of Machine Learning Research*, 81:1–15, 2018. <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.
- Burley, Anne-Marie Slaughter. "International Law and International Relations Theory: A Dual Agenda." *The American Journal of International Law* 87, no. 2 (1993): 205–39. <https://doi.org/10.2307/2203817>.
- Burns, Richard Dean. "An Introduction to Arms Control and Disarmament." In *Encyclopedia of Arms Control and Disarmament*, edited by Richard Dean Burns, 1:1–12. Charles Scribner's Sons, 1993.
- . *The Evolution of Arms Control: From Antiquity to the Nuclear Age*. Reprint edition. Lanham u.a.: Rowman & Littlefield Publishers, 2013.
- Burr, Christopher, and Nello Cristianini. "Can Machines Read Our Minds?" *Minds and Machines*, March 27, 2019. <https://doi.org/10.1007/s11023-019-09497-4>.
- Burrell, Jenna. "How the Machine Thinks": Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (June 1, 2016): 2053951715622512. <https://doi.org/10.1177/2053951715622512>.
- Burri, Thomas. "Free Movement of Algorithms: Artificially Intelligent Persons Conquer the European Union's Internal Market." In *Research Handbook on the Law of Artificial Intelligence*, edited by Woodrow Barfield and Ugo Pagallo. Edward Elgar, 2017. <https://papers.ssrn.com/abstract=3010233>.
- . "International Law and Artificial Intelligence." *German Yearbook of International Law* 60 (October 27, 2017): 91–108.
- Burt, David, Aaron Kleiner, J. Paul Nicholas, and Kevin Sullivan. "Cyberspace 2025: Today's Decisions, Tomorrow's Terrain." Microsoft, 2014. <https://www.microsoft.com/en-us/cybersecurity/content-hub/cyberspace-2025>.
- Butcher, James, and Irakli Beridze. "What Is the State of Artificial Intelligence Governance Globally?" *The RUSI Journal* 164, no. 5–6 (September 19, 2019): 88–96. <https://doi.org/10.1080/03071847.2019.1694260>.
- Caldwell, M., J. T. A. Andrews, T. Tanay, and L. D. Griffin. "AI-Enabled Future Crime." *Crime Science* 9, no. 1 (August 5, 2020): 14. <https://doi.org/10.1186/s40163-020-00123-8>.
- Calo, Ryan. "Artificial Intelligence Policy: A Primer and Roadmap." *UC Davis Law Review* 51 (2017): 37.
- . "Peeping HALs: Making Sense of Artificial Intelligence and Privacy." *EJLS - European Journal of Legal Studies*, The Future of... Law & Technology in the Information Society, 2, no. 3 (2010). <http://www.ejls.eu/6/83UK.htm>.
- . "Robotics and the Lessons of Cyberlaw." *California Law Review* 103 (2015): 513–64.
- Calo, Ryan, Ivan Evtimov, Earlene Fernandes, Tadayoshi Kohno, and David O'Hair. "Is Tricking a Robot Hacking?" *University of Washington School of Law Research Paper*, March 27, 2018. <https://papers.ssrn.com/abstract=3150530>.
- Campaign to Stop Killer Robots. "About Us." Accessed September 5, 2020. <https://www.stopkillerrobots.org/about/>.
- . "Diplomatic Talks in 2020," September 25, 2020. <https://www.stopkillerrobots.org/2020/09/diplomatic2020/>.
- Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. "AI Now 2017 Report." AI Now Institute, October 2017. https://assets.contentful.com/8wprhhvnpfc0/1A9c3ZTCza2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf.

- Candès, Emmanuel, John Duchi, and Chiara Sabatti. "Comments on Michael Jordan's Essay 'The AI Revolution Hasn't Happened Yet'." *Harvard Data Science Review* 1, no. 1 (June 20, 2019). <https://doi.org/10.1162/99608f92.ff7ca64e>.
- Casetext. "CARA A.I." Accessed September 8, 2020. <https://casetext.com/cara-ai/>.
- Carey, Ryan. "Interpreting AI Compute Trends." AI Impacts, July 10, 2018. <https://aiimpacts.org/interpreting-ai-compute-trends/>.
- Carlsmith, Joseph. "How Much Computational Power Does It Take to Match the Human Brain?" Open Philanthropy Project, August 14, 2020. <https://www.openphilanthropy.org/brain-computation-report>.
- Carpenter, Charli. *"Lost" Causes, Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Ithaca: Cornell University Press, 2014. <https://doi.org/10.7591/9780801470363>.
- Carvin, Stephanie. "Normal Autonomous Accidents: What Happens When Killer Robots Fail?" SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 1, 2017. <https://papers.ssrn.com/abstract=3161446>.
- Casati, Roberto, and Achille Varzi. "Holes." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University, 2019. <https://plato.stanford.edu/archives/sum2019/entries/holes/>.
- Casey, Anthony J., and Anthony Niblett. "Self-Driving Laws." *University of Toronto Law Journal* 66, no. 4 (October 2016): 429–42. <https://doi.org/10.3138/UTLJ.4006>.
- Casey, Anthony J., and Anthony Niblett. "The Death of Rules and Standards." *Indiana Law Journal* 92, no. 4 (Fall 2017): 1401–47.
- Castel, J.G., and Mathew E. Castel. "The Road to Artificial Superintelligence - Has International Law a Role to Play?" *Canadian Journal of Law & Technology* 14 (2016). <https://ojs.library.dal.ca/CJLT/article/download/7211/6256>.
- Cave, Stephen, and Seán S. Ó hÉigearaigh. "An AI Race for Strategic Advantage: Rhetoric and Risks." In *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*, 2018. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf.
- Cave, Stephen, and Seán S. Ó hÉigearaigh. "Bridging Near- and Long-Term Concerns about AI." *Nature Machine Intelligence* 1, no. 1 (January 2019): 5. <https://doi.org/10.1038/s42256-018-0003-2>.
- Cavelty, Myriam Dunn, Sophie-Charlotte Fischer, and Thierry Balzacq. "Killer Robots' and Preventative Arms Control." In *Routledge Handbook of Security Studies*, 15. Routledge, 2016.
- Chahal, Husanjot, Ryan Fedasiuk, and Carrick Flynn. "Messier than Oil: Assessing Data Advantage in Military AI." Center for Security and Emerging Technology, July 2020. <https://cset.georgetown.edu/research/messier-than-oil-assessing-data-advantage-in-military-ai/>.
- Chalmers, David J. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (2010): 7–65.
- Chavannes, Esther, Klaudia Klonowska, and Tim Sweijns. "Governing Autonomous Weapon Systems: Expanding the Solution Space, from Scoping to Applying." *HCSS Security*, 2020, 39.
- Checkel, Jeffrey T. "International Norms and Domestic Politics: Bridging the Rationalist—Constructivist Divide." *European Journal of International Relations* 3, no. 4 (December 1, 1997): 473–95. <https://doi.org/10.1177/1354066197003004003>.
- Chen, Ricky T. Q., Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. "Neural Ordinary Differential Equations." *ArXiv:1806.07366 [Cs, Stat]*, June 19, 2018. <http://arxiv.org/abs/1806.07366>.
- Cherry, John, and Christopher Korpela. "Enhanced Distinction: The Need for a More Focused Autonomous Weapons Targeting Discussion at the LAWS GGE." Humanitarian Law & Policy Blog, March 28, 2019. <https://blogs.icrc.org/law-and-policy/2019/03/28/enhanced-distinction-need-focused-autonomous-weapons-targeting/>.
- Chesney, Robert, and Danielle Keats Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107 (2019): 1753–1820.
- Chief Executives Board for Coordination. "Summary of Deliberations: A United Nations System-Wide Strategic Approach and Road Map for Supporting Capacity Development on Artificial Intelligence." United Nations System, June 17, 2019. <http://digitallibrary.un.org/record/3811676>.
- China's State Council. "A Next Generation Artificial Intelligence Development Plan." Translated by Rogier Creemers, Graham Webster, Paul Triolo, and Elsa Kania. New America Cybersecurity Initiative, August 1, 2017. <https://naproduction.s3.amazonaws.com/documents/translation-fulltext-8.1.17.pdf>.
- Chinoy, Sahil. "We Built an 'Unbelievable' (but Legal) Facial Recognition Machine." *The New York Times*, April 16, 2019, sec. Opinion. <https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html>.
- Christen, Markus, Thomas Burri, Joseph Chapa, Raphael Salvi, Filippo Santoni de Sio, and John Sullins. "An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications." Digital Society Initiative White Paper Series. University of Zurich, November 1, 2017. <https://papers.ssrn.com/abstract=3063617>.
- Christiano, Paul. "Takeoff Speeds." *The Sideways View* (blog), February 24, 2018. <https://sideways-view.com/2018/02/24/takeoff-speeds/>.
- Cihon, Peter. "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development." Technical Report. Oxford: Center for the Governance of AI, Future of Humanity Institute, University of Oxford, April 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.
- Cihon, Peter, Moritz J Kleinaltenkamp, Jonas Schuett, and Seth Baum. "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries," 10. Sandra Day O'Connor College of Law, Arizona State University, 2020.
- Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Fragmentation and the Future: Investigating

- Architectures for International AI Governance,” Forthcoming.
- . “Should Artificial Intelligence Governance Be Centralised?: Design Lessons from History.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–34. New York NY USA: ACM, 2020. <https://doi.org/10.1145/3375627.3375857>.
- Cirkovic, Milan M., Nick Bostrom, and Anders Sandberg. “Anthropic Shadow: Observation Selection Effects and Human Extinction Risks.” *Risk Analysis* 30, no. 10 (2010). <http://www.nickbostrom.com/papers/anthropicsshadow.pdf>.
- Citron, Danielle, and Robert Chesney. “Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics.” *Foreign Affairs*, March 1, 2019. <https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>.
- Clark, Jack. “Import AI 214: NVIDIA+ARM; a 57-Subject NLP Test; and AI for Plant Disease Identification,” September 14, 2020. <https://us13.campaign-archive.com/?u=67bd06787e84d73db24fb0aa5&id=a88c355f6a>.
- . “Import AI 215: The Hardware Lottery; Micro GPT3; and, the Peace Computer.” Import AI, September 21, 2020. <https://mailchi.mp/jack-clark/import-ai-215-the-hardware-lottery-micro-gpt3-and-the-peace-computer?e=524806dd16>.
- Clark, Jack, and Gillian K Hadfield. “Regulatory Markets for AI Safety,” 9, 2019.
- Clark, Nathan Edward. “Blurred Lines: Multi-Use Dynamics for Satellite Remote Sensing.” *Journal of International Humanitarian Legal Studies* 10, no. 1 (June 9, 2019): 171–83. <https://doi.org/10.1163/18781527-01001003>.
- Cleek, Margaret Anne, and Sherry Lynn Leonard. “Can Corporate Codes of Ethics Influence Behavior?” *Journal of Business Ethics* 17, no. 6 (April 1, 1998): 619–30. <https://doi.org/10.1023/A:1017969921581>.
- Clifton, Jesse. “Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda.” Effective Altruism Foundation, March 4, 2020. <https://longtermrisk.org/research-agenda?fbclid=IwAR0TeDk9PPrPzCOi4v7t2RcSwmb0xhC-I4mEYQodMAyBnJe2pmfBF0Mjvlo>.
- Coalition for Critical Technology. “Abolish the #TechToPrisonPipeline.” Medium, June 29, 2020. <https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techttoprisonpipeline-9b5b14366b16>.
- Coe, Andrew J., and Jane Vaynman. “Why Arms Control Is So Rare.” *American Political Science Review* 114, no. 2 (May 2020): 342–55. <https://doi.org/10.1017/S000305541900073X>.
- Cognilytica. “Data Engineering, Preparation, and Labeling for AI 2019.” Cognilytica, March 6, 2019. <https://www.cognilytica.com/2019/03/06/report-data-engineering-preparation-and-labeling-for-ai-2019/>.
- Cohen, Rachel S. “Air Force to Test Weapons Swarming Software in October.” *Air Force Magazine* (blog), September 21, 2020. <https://www.airforcemag.com/air-force-to-test-weapons-swarming-software-in-october/>.
- Cojazzi, Giacomo G. M., Erik Van Der Goot, Marco Verile, Erik Wolfart, Marcy Rutan Fowler, Yana Feldman, William Hammond, John Schweighardt, and Mattew Ferguson. “Collection and Analysis of Open Source News for Information Awareness and Early Warning in Nuclear Safeguards.” *ESARDA Bulletin* 50 (2013): 94–105.
- Cole, Samantha. “For the Love of God, Not Everything Is a Deepfake.” *Vice* (blog), April 27, 2020. https://www.vice.com/en_us/article/7kzgg9/joe-biden-tongue-gif-twitter-deepfake.
- Colgan, Jeff D., Robert O Keohane, and Thijs Van de Graaf. “Punctuated Equilibrium in the Energy Regime Complex.” *Review of International Organizations* 7, no. 2 (2012): 117–43.
- Collingridge, David. *The Social Control of Technology*. New York: Palgrave Macmillan, 1981.
- Comin, Diego, and Martí Mestieri. “If Technology Has Arrived Everywhere, Why Has Income Diverged?” *American Economic Journal: Macroeconomics* 10, no. 3 (July 2018): 137–78. <https://doi.org/10.1257/mac.20150175>.
- Comin, Diego, and Martí Mestieri. “Technology Adoption and Growth Dynamics,” 2014, 38.
- Constantin, Sarah. “Hoe Cultures: A Type of Non-Patriarchal Society.” *Otium* (blog), September 13, 2017. <https://srconstantin.wordpress.com/2017/09/13/hoe-cultures-a-type-of-non-patriarchal-society/>.
- Convention on Cybercrime, ETS No. 185 § (2001). <https://www.coe.int/en/web/conventions/full-list>.
- Convention on International Civil Aviation, 15 UNTS 295 § (1944).
- Convention on Rights and Duties of States adopted by the Seventh International Conference of American States (1933). <https://treaties.un.org/doc/Publication/UNTS/LON/Volumen%20165/v165.pdf>.
- Convention on Road Traffic, 125 UNTS 22 § (1949).
- Convention on Road Traffic, 1042 UNTS 15705 § (1968).
- Convention on the Suppression of Unlawful Acts Relating to International Civil Aviation, 50 ILM 144 § (2011).
- Cooper, Andrew F. “Stretching the Model of ‘Coalitions of the Willing.’” The Centre for International Governance Innovation, 2005. https://www.cigionline.org/sites/default/files/working_paper_1_-stretching_the_model_of_.pdf.
- Corea, Francesco. “AI Knowledge Map: How to Classify AI Technologies.” Medium, June 28, 2020. https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020.
- Cornebise, Julien, John Worrall, Micah Farfour, and Milena Marin. “Witnessing Atrocities: Quantifying Villages Destruction in Darfur with Crowdsourcing and Transfer Learning,” 2018. https://aiforsocialgood.github.io/2018/pdfs/track1/80_aisg_neurips2018.pdf.
- Corrêa, Nicolas Kluge. “Blind Spots in AI Ethics and Biases in AI Governance,” August 2020. https://www.researchgate.net/publication/342821723_Blind_Spots_in_AI_Ethics_and_Biases_in_AI_governance?channel=doi&linkId=5f47ea4fa6fdcc14c5d0da27&showFulltext=true.
- Council of Europe Convention on preventing and combating violence against women and domestic violence, CETS No. 210 § (2001).
- Cowen, Tyler, and Ben Southwood. “Is the Rate of Scientific Progress Slowing Down?,” 2019, 43.
- Crawford, James. “The Current Political Discourse Concerning International Law.” *The Modern Law Review* 81, no. 1 (2018): 1–22. <https://doi.org/10.1111/1468-2230.12314>.

- Crawford, Kate, and Ryan Calo. "There Is a Blind Spot in AI Research." *Nature News* 538, no. 7625 (October 20, 2016): 311. <https://doi.org/10.1038/538311a>.
- Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kazunias, Amba Kak, et al. "AI Now 2019 Report." AI Now Institute, 2019. https://ainowinstitute.org/AI_Now_2019_Report.pdf.
- Crawford, Kate, and Vladan Joler. "Anatomy of an AI System." AI Now Institute, 2018. <http://www.anatomyof.ai>.
- Cremer, Carla Zoe, and Jess Whittlestone. "Canaries in Technology Mines: Warning Signs of Transformative Progress in AI," 7. Santiago de Compostela, Spain, 2020.
- Crevier, Daniel. *AI: The Tumultuous Search for Artificial Intelligence*. New York: BasicBooks, 1993.
- Crootof, Rebecca. "Autonomous Weapon Systems and the Limits of Analogy." *Harvard National Security Journal* 9 (2018): 51–83. <https://doi.org/10.2139/ssrn.2820727>.
- _____. "Change Without Consent: How Customary International Law Modifies Treaties." *Yale Journal of International Law* 41, no. 2 (2016): 65.
- _____. "'Cyborg Justice' and the Risk of Technological-Legal Lock-In." *Columbia Law Review Forum*, October 5, 2019. <https://papers.ssrn.com/abstract=3464724>.
- _____. "International Cybertorts: Expanding State Accountability in Cyberspace." *Cornell Law Review* 103 (2018): 565–644.
- _____. "Jurisprudential Space Junk: Treaties and New Technologies." In *Resolving Conflicts in the Law*, edited by Chiara Giorgetti and Natalie Klein, 106–29, 2019. <https://brill.com/view/book/edcoll/9789004316539/BP00015.xml>.
- _____. "Regulating New Weapons Technology." In *The Impact of Emerging Technologies on the Law of Armed Conflict*, edited by Eric Talbot Jensen and Ronald T.P. Alcala, 1–25. Oxford University Press, 2019.
- _____. "The Killer Robots Are Here: Legal and Policy Implications." *CARDOZO LAW REVIEW* 36 (January 2015): 80.
- _____. "The Killer Robots Are Here: Legal and Policy Implications." *Cardozo Law Review* 36, no. 5 (June 2015): 1837–1915.
- _____. "War Torts: Accountability for Autonomous Weapons." *University of Pennsylvania Law Review* 164, no. 6 (May 2016): 1347–1402.
- _____. "Why the Prohibition on Permanently Blinding Lasers Is Poor Precedent for a Ban on Autonomous Weapon Systems." Lawfare, November 24, 2015. <https://www.lawfareblog.com/why-prohibition-permanently-blinding-lasers-poor-precedent-ban-autonomous-weapon-systems>.
- Crootof, Rebecca, and B. J. Ard. "Structuring Techlaw." *Harvard Journal of Law & Technology* 34 (forthcoming 2021). <https://papers.ssrn.com/abstract=3664124>.
- Cross, Stephen E., and Edward Walker. "DART: Applying Knowledge Based Planning and Scheduling to CRISIS Action Planning." In *Intelligent Scheduling*, edited by Monte Zweben and Mark Fox, 711–29. San Francisco, CA: Morgan Kaufmann, 1994.
- CTBTO Preparatory Commission. "Overview of the Verification Regime." Comprehensive Nuclear-Test-Ban Treaty Organization. Accessed September 9, 2020. <https://www.ctbto.org/verification-regime/background/overview-of-the-verification-regime/>.
- Cummings, Mary L., Heather Roff, Kenneth Cukier, Jacob Parakilas, and Hannah Bryce. *Artificial Intelligence and International Affairs: Disruption Anticipated*. Chatham House, 2018. <https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>.
- Cunneen, Martin, Martin Mullins, Finbarr Murphy, Darren Shannon, Irini Furxhi, and Cian Ryan. "Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics." *Cybernetics and Systems* 51, no. 1 (January 2, 2020): 59–80. <https://doi.org/10.1080/01969722.2019.1660541>.
- Currie, Adrian. "Existential Risk, Creativity & Well-Adapted Science." In *Studies in the History & Philosophy of Science*, 76:39–48, 2019. <https://www.sciencedirect.com/science/article/abs/pii/S0039368117303278?via%3Dihub>.
- _____. "Geoengineering Tensions." *Futures*, February 15, 2018. <https://doi.org/10.1016/j.futures.2018.02.002>.
- Currie, Adrian, and Shahar Avin. "Method Pluralism, Method Mismatch, & Method Bias." *Philosopher's Imprint* 19, no. 13 (2019). <http://hdl.handle.net/2027/spo.3521354.0019.013>.
- Cussins, Jessica. "National and International AI Strategies." Future of Life Institute, February 2020. <https://futureoflife.org/national-international-ai-strategies/>.
- Dabney, Will, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. "A Distributional Code for Value in Dopamine-Based Reinforcement Learning." *Nature*, January 15, 2020, 1–5. <https://doi.org/10.1038/s41586-019-1924-6>.
- Dacrema, Maurizio Ferrari, Paolo Cremonesi, and Dietmar Jannach. "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches." In *Proceedings of the 13th ACM Conference on Recommender Systems*, 101–109. RecSys '19. Copenhagen, Denmark: Association for Computing Machinery, 2019. <https://doi.org/10.1145/3298689.3347058>.
- Dafaoe, Allan. "AI Governance: A Research Agenda." Oxford: Center for the Governance of AI, Future of Humanity Institute, 2018. <https://www.fhi.ox.ac.uk/govaiagenda/>.
- _____. "AI Governance: Opportunity and Theory of Impact." EA Forum, September 17, 2020. <https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact>.
- _____. "On Technological Determinism: A Typology, Scope Conditions, and a Mechanism." *Science, Technology, & Human Values* 40, no. 6 (November 1, 2015): 1047–76. <https://doi.org/10.1177/0162243915579283>.
- Daly, Angela, Thilo Hagendorff, Li Hui, Monique Mann, Vidushi Marda, Ben Wagner, Wei Wang, and Saskia Witteborn. "Artificial Intelligence Governance and Ethics: Global Perspectives," June 28, 2019.

- <https://arxiv.org/ftp/arxiv/papers/1907/1907.03848.pdf>
- Danagoulian, Areg. "Verification of Arms Control Treaties with Resonance Phenomena." *Nuclear Physics News* 30, no. 1 (January 2, 2020): 25–30.
<https://doi.org/10.1080/10619127.2020.1717271>.
- Danaher, John. *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press, 2019.
- . "The Threat of Algoracry: Reality, Resistance and Accommodation." *Philosophy & Technology* 29, no. 3 (September 2016): 245–68.
<https://doi.org/10.1007/s13347-015-0211-1>.
- . "Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism." *Science and Engineering Ethics*, June 20, 2019.
<https://doi.org/10.1007/s11948-019-00119-x>.
- Danzig, Richard. "An Irresistible Force Meets a Moveable Object: The Technology Tsunami and the Liberal World Order." *Lawfare Research Paper Series* 5, no. 1 (August 28, 2017).
<https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf>.
- . "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority." Center for a New American Security, June 2018.
<https://www.cnas.org/publications/reports/technology-roulette>.
- Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." *Reuters*, October 10, 2018. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.
- Dauvergne, Peter. "The Globalization of Artificial Intelligence: Consequences for the Politics of Environmentalism." *Globalizations* 0, no. 0 (June 30, 2020): 1–15.
<https://doi.org/10.1080/14747731.2020.1785670>.
- Davis, Ernest, and Gary Marcus. "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence." *Communications of the ACM* 58, no. 9 (August 24, 2015): 92–103.
<https://doi.org/10.1145/2701413>.
- Davis, Joshua P. "Law Without Mind: AI, Ethics, and Jurisprudence." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, May 1, 2018.
<https://papers.ssrn.com/abstract=3187513>.
- De Spiegeleire, Stephan, Matthijs M. Maas, and Tim Sweijs. *Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-Sized Force Providers*. The Hague, The Netherlands: The Hague Centre for Strategic Studies, 2017. <http://hcss.nl/report/artificial-intelligence-and-future-defense>.
- De Spiegeleire, Stephan, Matthijs Maas, and Tim Sweijs. *Artificial Intelligence and the Future of Defense: Strategic Implications for a Small Force Provider*. The Hague, The Netherlands: HCSS, 2017.
- Deeks, Ashley. "Coding the Law of Armed Conflict: First Steps." In *The Law of Armed Conflict in 2040*, edited by Matthew C. Waxman. New York: Oxford University Press, 2020.
<https://papers.ssrn.com/abstract=3612329>.
- . "High-Tech International Law." *George Washington Law Review* 88 (2020): 575–653.
- . "Introduction to the Symposium: How Will Artificial Intelligence Affect International Law?" *AJIL Unbound* 114 (ed 2020): 138–40.
<https://doi.org/10.1017/aju.2020.29>.
- . "Predicting Enemies." *Virginia Law Review* 104 (2018): 1529–93.
- . "The International Legal Dynamics of Encryption." *Hoover Institute Aegis Paper Series* 1609 (2017): 28.
- Deeks, Ashley, Noam Lubell, and Daragh Murray. "Machine Learning, Artificial Intelligence, and the Use of Force by States." *Journal of National Security Law & Policy* 10 (2019): 1–25.
- Deep Cuts Commission. "Urgent Steps to Avoid a New Nuclear Arms Race and Dangerous Miscalculation -- Statement of the Deep Cuts Commission," March 19, 2018.
https://www.armscontrol.org/sites/default/files/files/documents/DCC_1804018_FINAL.pdf.
- DeepMind. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II." DeepMind, January 25, 2019.
<https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/>.
- Defense Science Board. "Defense Science Board Summer Study on Autonomy." US Department of Defense, 2016. <https://www.hsdl.org/?abstract&did=794641>.
- DeGusta, Michael. "Are Smart Phones Spreading Faster than Any Technology in Human History?" *MIT Technology Review*, May 9, 2012.
<https://www.technologyreview.com/2012/05/09/186160/are-smart-phones-spreading-faster-than-any-technology-in-human-history/>.
- Delcker, Janosch. "How Killer Robots Overran the UN." *POLITICO*, February 12, 2019.
<https://www.politico.eu/article/killer-robots-overran-united-nations-lethal-autonomous-weapons-systems/>.
- . "The AI Treaty You've Never Heard of — A Tech Industry Reckoning — Sabre-Rattling." *POLITICO AI: Decoded*, July 15, 2020.
<https://www.politico.eu/newsletter/ai-decoded/politico-ai-decoded-the-ai-treaty-youve-never-heard-of-a-tech-industry-reckoning-sabre-rattling/>.
- Deputy Secretary of Defense. "Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)." United States Department Of Defense, April 26, 2017.
<https://www.scribd.com/document/346681336/Establishment-of-the-AWCFT-Project-Maven>.
- Deshman, Abigail C. "Horizontal Review between International Organizations: Why, How, and Who Cares about Corporate Regulatory Capture." *European Journal of International Law* 22, no. 4 (November 1, 2011): 1089–1113.
<https://doi.org/10.1093/ejil/chr093>.
- DeSilver, Drew. "Chart of the Week: The Ever-Accelerating Rate of Technology Adoption." *Pew Research Center* (blog), March 14, 2014.
<https://www.pewresearch.org/fact-tank/2014/03/14/chart-of-the-week-the-ever-accelerating-rate-of-technology-adoption/>.
- Deudney, Daniel. *Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity*. Oxford, New York: Oxford University Press, 2020.
- . "Turbo Change: Accelerating Technological Disruption, Planetary Geopolitics, and Architectonic Metaphors." *International Studies Review* 20, no. 2 (June 1, 2018): 223–31.
<https://doi.org/10.1093/isr/viy033>.

- DeVries, Kelly. "Catapults Are Still Not Atomic Bombs: Effectiveness and Determinism in Premodern Military Technology." *Vulcan* 7, no. 1 (December 5, 2019): 34–44. <https://doi.org/10.1163/22134603-00701004>.
- DeVries, Phoebe M. R., Fernanda Viégas, Martin Wattenberg, and Brendan J. Meade. "Deep Learning of Aftershock Patterns Following Large Earthquakes." *Nature* 560, no. 7720 (August 2018): 632–34. <https://doi.org/10.1038/s41586-018-0438-y>.
- Dhariwal, Prafulla, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. "Jukebox: A Generative Model for Music." *ArXiv:2005.00341 [Cs, Eess, Stat]*, April 30, 2020. <http://arxiv.org/abs/2005.00341>.
- Diffie, Whitfield, and Susan Landau. "The Export of Cryptography in the 20th Century and the 21st," April 19, 2005. https://privacyink.org/pdf/export_control.pdf.
- Dimitrov, Radoslav S., Detlef F. Sprinz, Gerald M. DiGiusto, and Alexander Kelle. "International Nonregimes: A Research Agenda." *International Studies Review* 9, no. 2 (2007): 230–58. <https://doi.org/10.1111/j.1468-2486.2007.00672.x>.
- Ding, Jeffrey. "Deciphering China's AI Dream: The Context, Components, Capabilities, and Consequences of China's Strategy to Lead the World in AI." Future of Humanity Institute, Governance of AI Program, March 2018. https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf?platform=hootsuite.
- Ding, Jeffrey, and Allan Dafoe. "The Logic of Strategic Assets: From Oil to Artificial Intelligence." *ArXiv:2001.03246 [Cs, Econ, q-Fin]*, January 9, 2020. <http://arxiv.org/abs/2001.03246>.
- Dingwerth, Klaus, and Philipp Pattberg. "Global Governance as a Perspective on World Politics." *Global Governance* 12, no. 2 (2006): 185–203.
- Domíngos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. 1 edition. New York: Basic Books, 2015.
- Donahoe, Eileen, and Megan MacDuffee Metzger. "Artificial Intelligence and Human Rights." *Journal of Democracy* 30, no. 2 (2019): 115–26. <https://doi.org/10.1353/jod.2019.0029>.
- Dosi, Giovanni. "Technological Paradigms and Technological Trajectories." *Research Policy* 11 (1982): 147–62.
- Dressel, Julia, and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4, no. 1 (January 1, 2018): eaao5580. <https://doi.org/10.1126/sciadv.aao5580>.
- Drexler, K Eric. "Reframing Superintelligence: Comprehensive AI Services as General Intelligence." Technical Report. Oxford: Future of Humanity Institute, University of Oxford, January 2019. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.
- Dreyfus, Hubert. *On the Internet*. 1st ed. Routledge, 2001. <https://www.abebooks.com/first-edition/Internet-Thinking-Action-HUBERT-DREYFUS-Routledge/30612233740/bd>.
- Dreyfus, Hubert L. "Alchemy and Artificial Intelligence." Paper. RAND Corporation, 1965. <https://www.rand.org/pubs/papers/P3244.html>.
- Drezner, Daniel W. "Technological Change and International Relations." *International Relations* 33, no. 2 (June 1, 2019): 286–303. <https://doi.org/10.1177/0047117819834629>.
- Drezner, Daniel W. "The Power and Peril of International Regime Complexity." *Perspectives on Politics* 7, no. 1 (2009): 65–70.
- . "The Tragedy of the Global Institutional Commons." In *Back to Basics: State Power in a Contemporary World*, edited by Martha Finnemore and Judith Goldstein. Oxford University Press, 2013. <https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199970087.001.0001/acprof-9780199970087-chapter-13>.
- Drum, Kevin. "Tech World: Welcome to the Digital Revolution." *Foreign Affairs*, August 2018.
- Dumbacher, Erin D. "Limiting Cyberwarfare: Applying Arms-Control Models to an Emerging Technology." *The Nonproliferation Review* 25, no. 3–4 (May 4, 2018): 203–22. <https://doi.org/10.1080/10736700.2018.1515152>.
- Dunlap, Charles. "Lawfare Today: A Perspective." *Yale Journal of International Affairs*, January 1, 2008, 146–54.
- Dunoff, Jeffrey L., and Mark A. Pollack. "Reviewing Two Decades of IL/IR Scholarship." In *Interdisciplinary Perspectives on International Law and International Relations*, edited by Jeffrey L. Dunoff and Mark A. Pollack, 626–62. Cambridge: Cambridge University Press, 2012. <https://doi.org/10.1017/CBO9781139107310.031>.
- Dutton, Tim, Brent Barron, and Gaga Boskovic. "Building an AI World: Report on National and Regional AI Strategies." CIFAR, December 2018. https://www.cifar.ca/docs/default-source/ai-society/buildinganaiworld_eng.pdf?sfvrsn=fb18d129_4.
- Eaton, A. W. "Artifacts and Their Functions." In *The Oxford Handbook of History and Material Culture*, edited by Ivan Gaskell and Sarah Anne Carter, 2020. <https://doi.org/10.1093/oxfordhb/9780199341764.013.26>.
- Ebben, Maureen. "Automation and Augmentation: Human Labor as Essential Complement to Machines." In *Maintaining Social Well-Being and Meaningful Work in a Highly Automated Job Market*, by Shalin Hai-Jew, 2020. <https://doi.org/10.4018/978-1-7998-2509-8.ch001>.
- Eckersley, Peter, and Yomna Nasser. "EFF AI Progress Measurement Project (2017-)." Electronic Frontier Foundation, June 12, 2017. <https://www.eff.org/ai/metrics>.
- Eckersley, Robyn. "Moving Forward in the Climate Negotiations: Multilateralism or Minilateralism?" *Global Environmental Politics* 12, no. 2 (2012): 24–42.
- . "The Big Chill: The WTO and Multilateral Environmental Agreements." *Global Environmental Politics* 4, no. 2 (May 2004): 24–50. <https://doi.org/10.1162/152638004323074183>.
- Edelman. "2019 Edelman AI Survey." Edelman, March 2019. https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_AI_Survey_Whitepaper.pdf.
- . "2019 Edelman Trust Barometer: Trust in Technology." Edelman, 2019. <https://www.edelman.com/sites/g/files/aatuss191/files/>

- 2019-
- 04/2019_Edelman_Trust_Barometer_Technology_Report.pdf.
- Eichensehr, Kristen E. "Cyberwar & International Law Step Zero." *TEXAS INTERNATIONAL LAW JOURNAL* 50, no. 2 (2015): 24.
- Eilstrup-Sangiovanni, Mette. "Why the World Needs an International Cyberwar Convention." *Philosophy & Technology* 31, no. 3 (September 2018): 379–407. <https://doi.org/10.1007/s13347-017-0271-5>.
- Ekelhof, Merel. "Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation." *Global Policy* 10, no. 3 (March 2019). <https://doi.org/10.1111/1758-5899.12665>.
- Elish, M. C. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (We Robot 2016)." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 20, 2016. <https://papers.ssrn.com/abstract=2757236>.
- Ellsberg, Daniel. *The Doomsday Machine: Confessions of a Nuclear War Planner*. Bloomsbury USA, 2017.
- Els, Andrea Scripa. "Artificial Intelligence as a Digital Privacy Protector." *Harvard Journal of Law & Technology* 31, no. 1 (2017): 19.
- Engler, Alex. "Fighting Deepfakes When Detection Fails." *Brookings* (blog), November 14, 2019. <https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/>.
- Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. "Runaway Feedback Loops in Predictive Policing." *ArXiv:1706.09847 [Cs, Stat]*, June 29, 2017. <http://arxiv.org/abs/1706.09847>.
- Erdelyi, Olivia J., and Judy Goldsmith. "Regulating Artificial Intelligence: Proposal for a Global Solution." In *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*, 95–101, 2018. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf.
- Ernest, Nicholas, David Carroll, Corey Schumacher, Matthew Clark, Kelly Cohen, and Gene Lee. "Genetic Fuzzy Based Artificial Intelligence for Unmanned Combat Aerial Vehicle Control in Simulated Air Combat Missions." *Journal of Defense Management* 06, no. 01 (2016). <https://doi.org/10.4172/2167-0374.1000144>.
- Est, Q. C. van, J. Gerritsen, and L. Kool. "Human Rights in the Robot Age: Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality." The Hague: Rathenau Instituut, May 11, 2017. <https://research.tue.nl/en/publications/human-rights-in-the-robot-age-challenges-arising-from-the-use-of->
- Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542, no. 7639 (February 2017): 115–18. <https://doi.org/10.1038/nature21056>.
- Eth, Daniel. "The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes." *Informatica* 41, no. 4 (December 27, 2017). <http://www.informatica.si/index.php/informatica/article/view/1874>.
- Etzioni, Oren. "How to Know If Artificial Intelligence Is about to Destroy Civilization." *MIT Technology Review*, February 25, 2020. <https://www.technologyreview.com/2020/02/25/906083/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/>.
- Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.
- European Commission. "Ethics Guidelines for Trustworthy AI." The European Commission's High-Level Expert Group on Artificial Intelligence, April 8, 2019. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- Evans, R., J. Jumper, J. Kirkpatrick, L. Sifre, T.F.G. Green, C. Qin, A. Zidek, et al. "De Novo Structure Prediction with Deep-Learning Based Scoring." In *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts)*, 2018. <https://deepmind.com/blog/alphafold/>.
- Everitt, Tom, Gary Lea, and Marcus Hutter. "AGI Safety Literature Review." *ArXiv:1805.01109 [Cs]*, May 3, 2018. <http://arxiv.org/abs/1805.01109>.
- Fallis, Don. "The Epistemic Threat of Deepfakes." *Philosophy & Technology*, August 6, 2020. <https://doi.org/10.1007/s13347-020-00419-2>.
- Farrell, Henry, and Bruce Schneier. "Common-Knowledge Attacks on Democracy." Research Publication. Berkman Klein Center, October 1, 2018. <https://papers.ssrn.com/abstract=3273111>.
- Fatton, Lionel P. "The Impotence of Conventional Arms Control: Why Do International Regimes Fail When They Are Most Needed?" *Contemporary Security Policy* 37, no. 2 (May 3, 2016): 200–222. <https://doi.org/10.1080/13523260.2016.1187952>.
- Faude, Benjamin, and Thomas Gehring. "Regime Complexes as Governance Systems." In *Research Handbook on the Politics of International Law*, edited by Wayne Sandholtz and Christopher Whytock, 176–203. Edward Elgar Publishing, 2017. <https://doi.org/10.4337/9781783473984>.
- Fearon, James, and Alexander Wendt. "Rationalism v. Constructivism: A Skeptical View." In *Handbook of International Relations*, 52–72. London: SAGE Publications Ltd, 2002. <https://doi.org/10.4135/9781848608290>.
- Fehl, Caroline. "Explaining the International Criminal Court: A 'Practice Test' for Rationalist and Constructivist Approaches." *European Journal of International Relations* 10, no. 3 (2004): 357–94. <https://doi.org/10.1177/1354066104045541>.
- . "Unequal Power and the Institutional Design of Global Governance: The Case of Arms Control." *Review of International Studies* 40, no. 3 (July 2014): 505–31. <https://doi.org/10.1017/S026021051300034X>.
- Fehl, Caroline, and Elvira Rosert. "It's Complicated: A Conceptual Framework for Studying Relations and Interactions between International Norms." PRIF Working Paper. Peace Research Institute Frankfurt, September 2020.
- Feijoo, Claudio, Younsun Kwon, Johannes M. Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. "Harnessing Artificial Intelligence (AI) to Increase Wellbeing for All: The Case for a New Technology Diplomacy." *Telecommunications Policy*, May 6, 2020, 101988. <https://doi.org/10.1016/j.telpol.2020.101988>.

- Feldstein, Steven. "The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression." *Journal of Democracy* 30, no. 1 (January 9, 2019): 40–52. <https://doi.org/10.1353/jod.2019.0003>.
- Field, Matt. "Strangelove Redux: US Experts Propose Having AI Control Nuclear Weapons." *Bulletin of the Atomic Scientists* (blog), August 30, 2019. <https://thebulletin.org/2019/08/strangelove-redux-us-experts-propose-having-ai-control-nuclear-weapons/>.
- Finlayson, Samuel G., John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. "Adversarial Attacks on Medical Machine Learning." *Science* 363, no. 6433 (March 22, 2019): 1287–89. <https://doi.org/10.1126/science.aaw4399>.
- Finnemore, Martha, and Kathryn Sikkink. "International Norm Dynamics and Political Change." *International Organization* 52, no. 4 (1998): 887–917.
- Fischer, Sophie-Charlotte, and Andreas Wenger. "A Politically Neutral Hub for Basic AI Research." Policy Perspectives. Zurich: CSS, ETH Zurich, March 2019. http://www.css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.
- Fischer-Lescano, Andreas, and Gunther Teubner. "Regime-Collisions: The Vain Search for Legal Unity in the Fragmentation of Global Law." *Michigan Journal of International Law* 25, no. 4 (2004): 999–1046.
- Fitzpatrick, Mark. "Artificial Intelligence and Nuclear Command and Control." *Survival* 61, no. 3 (May 4, 2019): 81–92. <https://doi.org/10.1080/00396338.2019.1614782>.
- Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhu Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," January 15, 2020. <https://dash.harvard.edu/handle/1/42160420>.
- Fjeld, Jessica, Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Joshua Feldman, and Sally Kagay. "Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches." Berkman Klein Center for Internet & Society at Harvard University, 2019. <https://ai-hr.cyber.harvard.edu/images/primpviz.pdf>.
- Floridi, Luciano. "AI and Its New Winter: From Myths to Realities." *Philosophy & Technology* 33, no. 1 (March 1, 2020): 1–3. <https://doi.org/10.1007/s13347-020-00396-6>.
- . "Energy, Risks, and Metatechnology." *Philosophy & Technology* 24, no. 2 (June 1, 2011): 89–94. <https://doi.org/10.1007/s13347-011-0026-7>.
- Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28, no. 4 (December 1, 2018): 689–707. <https://doi.org/10.1007/s11023-018-9482-5>.
- Floridi, Luciano, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. "How to Design AI for Social Good: Seven Essential Factors." *Science and Engineering Ethics*, April 3, 2020. <https://doi.org/10.1007/s11948-020-00213-5>.
- Flynn, Carrick. "Recommendations on Export Controls for Artificial Intelligence." Center for Security and Emerging Technology, February 2020. <https://cset.georgetown.edu/recommendation-s-on-export-controls-for-artificial-intelligence/>.
- Foreign Affairs. "Does Technology Favor Tyranny?" Foreign Affairs, February 13, 2019. <https://www.foreignaffairs.com/ask-the-experts/2019-02-12/does-technology-favor-tyranny>.
- Fountain, Henry. "A Dream of Clean Energy at a Very High Price." *The New York Times*, March 27, 2017, sec. Science. <https://www.nytimes.com/2017/03/27/science/fusion-power-plant-iter-france.html>.
- France Diplomatic Ministry for Europe and Foreign Affairs. "11 Principles on Lethal Autonomous Weapons Systems (LAWS)." France Diplomacy - Ministry for Europe and Foreign Affairs, 2019. <https://www.diplomatie.gouv.fr/en/french-foreign-policy/united-nations/multilateralism-a-principle-of-action-for-france/alliance-for-multilateralism-63158/article/11-principles-on-lethal-autonomous-weapons-systems-laws>.
- Freedberg, Sydney J. "A Slew To A Kill: Project Convergence." *Breaking Defense* (blog), September 16, 2020. <https://breakingdefense.com/2020/09/a-slew-to-a-kill-project-convergence/>.
- . "Target Gone In 20 Seconds: Army Sensor-Shooter Test." *Breaking Defense* (blog), September 10, 2020. <https://breakingdefense.com/2020/09/target-gone-in-20-seconds-army-sensor-shooter-test/>.
- Frey, Carl Benedikt. *The Technology Trap: Capital, Labor, and Power in the Age of Automation*. Princeton University Press, 2019.
- Frey, Carl Benedikt, and Michael A. Osborne. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114 (January 2017): 254–80. <https://doi.org/10.1016/j.techfore.2016.08.019>.
- Friedman, Daniel A., and Eirik Sovik. "The Ant Colony as a Test for Scientific Theories of Consciousness." *Synthese*, February 12, 2019. <https://doi.org/10.1007/s11229-019-02130-y>.
- Friedman, David D. "Does Technology Require New Law?" *Public Policy* 71 (2001): 16.
- Fuller, Lon L. *The Morality of Law*. Yale University Press, 1969.
- Funke, Christina M., Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas S. A. Wallis, and Matthias Bethge. "The Notorious Difficulty of Comparing Human and Machine Perception." *ArXiv:2004.09406 [Cs, q-Bio, Stat]*, April 20, 2020. <http://arxiv.org/abs/2004.09406>.
- Future of Life Institute. "AI Safety Myths." Future of Life Institute, 2016. <https://futureoflife.org/background/aimyths/>.
- . "Asilomar AI Principles." Future of Life Institute, 2017. <https://futureoflife.org/ai-principles/>.
- . "Open Letter on Autonomous Weapons." Future of Life Institute, 2015. <https://futureoflife.org/open-letter-autonomous-weapons/>.
- Fuzaylova, Elizabeth. "War Torts, Autonomous Weapon Systems, and Liability: Why a Limited Strict Liability Tort Regime Should Be Implemented." *Cardozo Law Review* 40, no. 3 (March 5, 2019): 1327–66.
- G20. "G20 Ministerial Statement on Trade and Digital Economy," June 28, 2019. <https://www.mofa.go.jp/files/000486596.pdf>.

- . “Ministerial Declaration: G20 Digital Economy Ministers Meeting, 2020 Riyadh Summit,” July 22, 2020. <http://www.g20.utoronto.ca/2020/2020-g20-digital-0722.html>.
- Gabriel, Iason. “Artificial Intelligence, Values and Alignment.” *DeepMind Research*, January 13, 2020. https://deepmind.com/research/publications/Artificial-Intelligence-Values-and-Alignment?fbclid=IwAR1gmUGQSnHD-ly56XDKde4Cvo3OTwMddtHOX3aBxo_xTJiBgGEO_OSe_mlk.
- Gal, Danit. “Perspectives and Approaches in AI Ethics: East Asia.” In *Oxford Handbook of Ethics of Artificial Intelligence*, edited by Markus Dubber, Frank Pasquale, and Sunit Das. Oxford University Press, 2019. <https://papers.ssrn.com/abstract=3400816>.
- Galbreath, David J., and Sascha Sauerteig. “Regime Complexity and Security Governance.” In *Handbook of Governance and Security*, edited by James Sperling, 82–97. Edward Elgar, 2014. [https://www.elgaronline.com/view/edcoll/9781781953167.00014.xml](https://www.elgaronline.com/view/edcoll/9781781953167/9781781953167.00014.xml).
- Gallagher, Nancy W. “Re-Thinking the Unthinkable: Arms Control in the Twenty-First Century.” *The Nonproliferation Review* 22, no. 3–4 (October 2, 2015): 469–98. <https://doi.org/10.1080/10736700.2016.1149279>.
- Gallie, W. B. “Essentially Contested Concepts.” *Proceedings of the Aristotelian Society* 56 (1955): 167–98.
- Gamble, John, and Charlotte Ku. “International Law - New Actors and New Technologies: Center Stage for NGOs.” *Law and Policy in International Business* 31 (January 1, 2000): 221.
- Garcia, Denise. “Future Arms, Technologies, and International Law: Preventive Security Governance.” *European Journal of International Security* 1, no. 1 (February 2016): 94–111. <https://doi.org/10.1017/eis.2015.7>.
- . “Lethal Artificial Intelligence and Change: The Future of International Peace and Security.” *International Studies Review* 20, no. 2 (June 2018): 334–41. <https://doi.org/10.1093/isr/viy029>.
- Garcia, Eugenio V. “Multilateralism and Artificial Intelligence: What Role for the United Nations?” In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Taylor & Francis, 2020.
- Garcia, Eugenio V. “The Militarization of Artificial Intelligence: A Wake-up Call for the Global South,” September 2019. https://www.researchgate.net/publication/335787908_The_militarization_of_artificial_intelligence_a_wake-up_call_for_the_Global_South/references.
- Gardner, Allison F. “Environmental Monitoring’s Undiscovered Country: Developing a Satellite Remote Monitoring System to Implement the Kyoto Protocol’s Global Emissions-Trading Program.” *New York University Environmental Law Journal* 9 (2001 2000): 152.
- Garfinkel, Ben. “Reinterpreting ‘AI and Compute.’” AI Impacts, December 18, 2018. <https://aiimpacts.org/reinterpreting-ai-and-compute/>.
- Garfinkel, Ben, Miles Brundage, Daniel Filan, Carrick Flynn, Jelena Luketina, Michael Page, Anders Sandberg, Andrew Snyder-Beattie, and Max Tegmark. “On the Impossibility of Supersized Machines.” *ArXiv:1703.10987 [Physics]*, March 31, 2017. <http://arxiv.org/abs/1703.10987>.
- Garfinkel, Ben, and Allan Dafoe. “How Does the Offense-Defense Balance Scale?” *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 736–63. <https://doi.org/10.1080/01402390.2019.1631810>.
- Garthoff, Raymond L. “Banning the Bomb in Outer Space.” *International Security* 5, no. 3 (1980): 25–40. <https://doi.org/10.2307/2538418>.
- Gartner. “Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence.” Gartner, 2018. <https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence>.
- Gasser, Urs. “Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy.” *Harvard Law Review Forum, LAW, PRIVACY & TECHNOLOGY COMMENTARY SERIES*, 130 (December 9, 2016): 10.
- Gasser, Urs, and Virgilio A.F. Almeida. “A Layered Model for AI Governance.” *IEEE Internet Computing* 21, no. 6 (November 2017): 58–62. <https://doi.org/10.1109/MIC.2017.4180835>.
- Gehring, Thomas. *Dynamic International Regimes: Institutions for International Environmental Governance*. Frankfurt am Main ; New York: Peter Lang GmbH, Internationaler Verlag der Wissenschaften, 1994.
- Gehring, Thomas, and Benjamin Faude. “The Dynamics of Regime Complexes: Microfoundations and Systemic Effects.” *Global Governance* 19, no. 1 (2013): 119–30.
- Gehring, Thomas, and Sebastian Oberthür. “The Causal Mechanisms of Interaction between International Institutions.” *European Journal of International Relations* 15, no. 1 (March 1, 2009): 125–56. <https://doi.org/10.1177/1354066108100055>.
- Geist, Edward, and Andrew J Lohn. “How Might Artificial Intelligence Affect the Risk of Nuclear War?” RAND, 2018. <https://www.rand.org/pubs/perspectives/PE296.html>.
- Geist, Edward Moore. “It’s Already Too Late to Stop the AI Arms Race—We Must Manage It Instead.” *Bulletin of the Atomic Scientists* 72 (2016): 318–21.
- Gershgorin, Dave. “Here’s How We Prevent The Next Racist Chatbot.” Popular Science, March 24, 2016. <https://www.popsci.com/heres-how-we-prevent-next-racist-chatbot>.
- Gettler, Dan. “The Drone Databook.” The Center for the Study of the Drone at Bard College, 2019. <https://dronecenter.bard.edu/files/2019/10/CSD-Drone-Databook-Web.pdf>.
- Gibor, Lih. “The Brilliant Rocket Scientist to Come Before His Time.” Text. Davidson Institute, November 12, 2016. <http://davidson.weizmann.ac.il/en/online/sciencehistory/brilliant-rocket-scientist-come-his-time>.
- Gibson, James J. “The Ecological Approach to the Visual Perception of Pictures.” *Leonardo* 11, no. 3 (1978): 227. <https://doi.org/10.2307/1574154>.
- Gill, Indermit. “Whoever Leads in Artificial Intelligence in 2030 Will Rule the World until 2100.” *Brookings* (blog), January 17, 2020. <https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/>.
- Gilli, Andrea, and Mauro Gilli. “Why China Has Not Caught Up Yet: Military-Technological Superiority

- and the Limits of Imitation, Reverse Engineering, and Cyber Espionage.” *International Security* 43, no. 3 (February 1, 2019): 141–89. https://doi.org/10.1162/isec_a_00337.
- Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. “Explaining Explanations: An Overview of Interpretability of Machine Learning.” *ArXiv:1806.00069 [Cs, Stat]*, February 3, 2019. <http://arxiv.org/abs/1806.00069>.
- Glennon, Michael J. “The Dark Future of International Cybersecurity Regulation.” *Journal of National Security Law & Policy* 6 (2013): 563–70.
- Global Partnership on Artificial Intelligence. “Joint Statement from Founding Members of the Global Partnership on Artificial Intelligence,” June 15, 2020. <https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding>.
- Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. “Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators.” *Journal of the American Medical Informatics Association: JAMIA* 19, no. 1 (February 2012): 121–27. <https://doi.org/10.1136/amiajnl-2011-000089>.
- Goertzel, Ben. “Artificial General Intelligence: Concept, State of the Art, and Future Prospects.” *Journal of Artificial General Intelligence* 5, no. 1 (December 1, 2014): 1–48. <https://doi.org/10.2478/jagi-2014-0001>.
- Goertzel, Ben, and Cassio Pennachin, eds. *Artificial General Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-68677-4_5.
- Gómez-Mera, Laura. “Regime Complexity and Global Governance: The Case of Trafficking in Persons.” *European Journal of International Relations* 22, no. 3 (September 1, 2016): 566–95. <https://doi.org/10.1177/1354066115600226>.
- Gómez-Mera, Laura, Jean-Frédéric Morin, and Thijs Van De Graaf. “Regime Complexes.” In *Architectures of Earth System Governance: Institutional Complexity and Structural Transformation*, edited by Frank Biermann and Rakhyun E. Kim, 137–57. Cambridge University Press, 2020.
- Good, I.J. “Speculations Concerning the First Ultraintelligent Machine.” In *Advances in Computers*, edited by Franz L. Alt and Moris Rubinoff, 6:31–88. New York: Academic Press, 1964.
- Goode, Lauren. “Google CEO Sundar Pichai Says AI Is More Profound than Electricity or Fire.” The Verge, January 19, 2018. <https://www.theverge.com/2018/1/19/16911354/google-ceo-sundar-pichai-ai-artificial-intelligence-fire-electricity-jobs-cancer>.
- Goodfellow, Ian J., Nicolas Papernot, Sandy Huang, Yan Duan, Pieter Abbeel, and Jack Clark. “Attacking Machine Learning with Adversarial Examples.” *OpenAI* (blog), February 16, 2017. <https://openai.com/blog/adversarial-example-research/>.
- Gould, Joe. “Pentagon Finally Gets Its 2020 Budget from Congress.” Defense News, December 19, 2019. <https://www.defensenews.com/congress/2019/12/19/pentagon-finally-gets-its-2020-budget-from-congress/>.
- Graaf, Thijs Van de. “Fragmentation in Global Energy Governance: Explaining the Creation of IRENA.” *Global Environmental Politics* 13, no. 3 (2013): 14–33.
- Grace, Katja. “Discontinuous Progress in History: An Update.” AI Impacts, April 13, 2020. <https://aiimpacts.org/discontinuous-progress-in-history-an-update/>.
- . “Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation.” Technical Report. Berkeley, CA: Machine Intelligence Research Institute, October 2015. <https://intelligence.org/files/SzilardNuclearWeapons.pdf>.
- Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. “When Will AI Exceed Human Performance? Evidence from AI Experts.” *Journal of Artificial Intelligence Research* 62 (July 31, 2018): 729–54. <https://doi.org/10.1613/jair.1.11222>.
- Gray, Mary L., and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt, 2019.
- Gregg, Aaron, and Paul Sonne. “Air Force Seeks a Radical Shift in How Jets, Missiles and Satellites Are Designed.” *Washington Post*, September 15, 2020. <https://www.washingtonpost.com/business/2020/09/15/air-force-digital-design/>.
- Gregg, Justin. “Is Your Toddler Really Smarter than a Chimpanzee?” *BBC Earth*, October 12, 2014. <http://www.bbc.com/earth/story/20141012-are-toddlers-smarter-than-chimps>.
- Grotto, Andrew. “Genetically Modified Organisms: A Precautionary Tale For AI Governance.” *AI Pulse*, January 24, 2019. <https://aipulse.org/genetically-modified-organisms-a-precautionary-tale-for-ai-governance-2/>.
- Gruetzmacher, Ross, Florian Dorner, Niko Bernaola-Alvarez, Charlie Giattino, and David Manheim. “Forecasting AI Progress: A Research Agenda.” *ArXiv:2008.01848 [Cs]*, August 4, 2020. <http://arxiv.org/abs/2008.01848>.
- Gruetzmacher, Ross, David Paradise, and Kang Bok Lee. “Forecasting Transformative AI: An Expert Survey.” *ArXiv:1901.08579 [Cs]*, January 24, 2019. <http://arxiv.org/abs/1901.08579>.
- Gruetzmacher, Ross, and Jess Whittlestone. “Defining and Unpacking Transformative AI.” *ArXiv:1912.00747 [Cs]*, November 27, 2019. <http://arxiv.org/abs/1912.00747>.
- Guibot, Michael, Anne F. Matthew, and Nicolas Suzor. “Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence.” *Vanderbilt Journal of Entertainment & Technology Law*, July 28, 2017. <https://papers.ssrn.com/abstract=3017004>.
- Gunkel, David J. *Robot Rights*. Cambridge, MA: MIT Press, 2018.
- Gupta, Abhishek, Camille Lanteigne, and Victoria Heath. “Report Prepared by the Montreal AI Ethics Institute (MAIEI) for Publication Norms for Responsible AI by Partnership on AI.” *ArXiv:2009.07262 [Cs]*, September 15, 2020. <http://arxiv.org/abs/2009.07262>.
- Guterres, António. “UN Secretary-General’s Strategy on New Technologies.” United Nations, September 2018. <http://www.un.org/en/newtechnologies/images/pdf/SG-s-Strategy-on-New-Technologies.pdf>.
- Gutierrez Gaviria, Carlos Ignacio. “The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps

- Generated by AI in the U.S." Thesis, Pardee RAND Graduate School, 2020.
https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html.
- Guzman, Andrew T. "Saving Customary International Law." *Michigan Journal of International Law* 27 (2005): 115. <https://doi.org/10.2139/ssrn.708721>.
- Guzman, Andrew T., and Timothy L Meyer. "International Soft Law." *Journal of Legal Analysis* 2, no. 1 (2010): 59.
- Gyűrösí, Miroslav. "The Soviet Fractional Orbital Bombardment System Program." Air Power Australia, January 2, 2010.
<http://www.ausairpower.net/APA-Sov-FOBS-Program.html>.
- Haas, Peter M. "Introduction: Epistemic Communities and International Policy Coordination." *International Organization* 46, no. 1, (1992): 1–35.
- Hachey, Krystal K., Tamir Libel, and Zack Partington. "The Impact of Artificial Intelligence on the Military Profession." In *Rethinking Military Professionalism for the Changing Armed Forces*, edited by Krystal K. Hachey, Tamir Libel, and Waylon H. Dean, 201–11. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-45570-5_13.
- Hadfield-Menell, Dylan, and Gillian Hadfield. "Incomplete Contracting and AI Alignment." *ArXiv:1804.04268 [Cs]*, April 11, 2018. <http://arxiv.org/abs/1804.04268>.
- Hadjeres, Gaëtan, François Pachet, and Frank Nielsen. "DeepBach: A Steerable Model for Bach Chorales Generation." *ArXiv:1612.01010 [Cs]*, December 3, 2016. <http://arxiv.org/abs/1612.01010>.
- Hafner, Gerhard. "Pros and Cons Ensuing from Fragmentation of International Law." *Michigan Journal of International Law* 25, no. 4 (2004): 849–63.
- Hafner-Burton, Emilie M. "The Power Politics of Regime Complexity: Human Rights Trade Conditionality in Europe." *Perspectives on Politics* 7, no. 1 (2009): 33–37.
- Hagemann, Ryan, Jennifer Huddleston, and Adam D. Thierer. "Soft Law for Hard Problems: The Governance of Emerging Technologies in an Uncertain Future." *Colorado Technology Law Journal* 17, no. 1 (2018): 94.
- Hagendorff, Thilo. "The Ethics of AI Ethics -- An Evaluation of Guidelines." *ArXiv:1903.03425 [Cs, Stat]*, February 28, 2019.
<http://arxiv.org/abs/1903.03425>.
- Hagendorff, Thilo, and Katharina Wezel. "15 Challenges for AI: Or What AI (Currently) Can't Do." *AI & SOCIETY*, March 28, 2019.
<https://doi.org/10.1007/s00146-019-00886-y>.
- Hale, Thomas, and David Held. *Beyond Gridlock*. Newark, UNITED KINGDOM: Polity Press, 2017.
- Halina, Marta, and Joseph Martin. "Five Ways AI Is Not Like the Manhattan Project (and One Way It Is)." *3 Quarks Daily* (blog), August 5, 2019.
<https://www.3quarksdaily.com/3quarksdaily/2019/08/five-ways-ai-is-not-like-the-manhattan-project-and-one-way-it-is.html>.
- Hambling, David. "The Pentagon Has a Laser That Can Identify People from a Distance—by Their Heartbeat." *MIT Technology Review*, June 27, 2019.
<https://www.technologyreview.com/s/613891/the-pentagon-has-a-laser-that-can-identify-people-from-a-distance-by-their-heartbeat/>.
- Hancock, P. A. "Some Pitfalls in the Promises of Automated and Autonomous Vehicles." *Ergonomics* 62, no. 4 (April 2019): 479–95.
<https://doi.org/10.1080/00140139.2018.1498136>.
- Haner, Justin, and Denise Garcia. "The Artificial Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development." *Global Policy* 10, no. 3 (2019): 331–37. <https://doi.org/10.1111/1758-5899.12713>.
- Haner, Justin K. "Dark Horses in the Lethal AI Arms Race," 2019. <https://justinkhaner.com/aiarmsrace>.
- Hanson, Robin. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford, New York: Oxford University Press, 2016.
- Hao, Karen. "A Radical New Neural Network Design Could Overcome Big Challenges in AI." *MIT Technology Review*, December 12, 2018.
<https://www.technologyreview.com/s/612561/a-radical-new-neural-network-design-could-overcome-big-challenges-in-ai/>.
- . "An Algorithm That Learns through Rewards May Show How Our Brain Does Too." *MIT Technology Review*, January 15, 2020.
<https://www.technologyreview.com/s/615054/deepmind-ai-reinforcement-learning-reveals-dopamine-neurons-in-brain/>.
- . "Human Rights Activists Want to Use AI to Help Prove War Crimes in Court." *MIT Technology Review*, June 25, 2020.
<https://www.technologyreview.com/2020/06/25/100446/6/ai-could-help-human-rights-activists-prove-war-crimes/>.
- Harding, Verity. "Lessons from History: What Can Past Technological Breakthroughs Teach the AI Community Today," 2020.
<https://www.bennettinstitute.cam.ac.uk/blog/lessons-history-what-can-past-technological-breakthroughs-teach-the-ai-community-today/>.
- Harrington, Jean. "Don't Worry—It Can't Happen." *Scientific American* 162, no. 5 (May 1940): 268–268.
<https://doi.org/10.1038/scientificamerican0540-268>.
- Hart, H. L. A. "Positivism and the Separation of Law and Morals." *Harvard Law Review* 71, no. 4 (February 1958): 593. <https://doi.org/10.2307/1338225>.
- Hartzog, Woodrow, Gregory Conti, John Nelson, and Lisa Shay. "Inefficiently Automated Law Enforcement." *Michigan State Law Review* 2015, no. 5 (January 1, 2016): 1763.
- Harwell, Drew. "Defense Department Pledges Billions toward Artificial Intelligence Research." *Washington Post*, September 7, 2018.
<https://www.washingtonpost.com/technology/2018/09/07/defense-department-pledges-billions-toward-artificial-intelligence-research/>.
- Hathaway, Oona A., and Scott J. Shapiro. *The Internationalists: How a Radical Plan to Outlaw War Remade the World*. Advance Reading Copy edition. New York: Simon & Schuster, 2017.
- Hayward, Keith J., and Matthijs M Maas. "Artificial Intelligence and Crime: A Primer for Criminologists." *Crime, Media, Culture*, June 30, 2020, 1741659020917434.
<https://doi.org/10.1177/1741659020917434>.
- He, He, Nanyun Peng, and Percy Liang. "Pun Generation with Surprise." *ArXiv:1904.06828 [Cs]*, April 14, 2019. <http://arxiv.org/abs/1904.06828>.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing

- Human-Level Performance on ImageNet Classification.” *ArXiv:1502.01852 [Cs]*, February 6, 2015. <http://arxiv.org/abs/1502.01852>.
- Hedberg, Sara Reese. “DART: Revolutionizing Logistics Planning.” *IEEE Intelligent Systems* 17, no. 3 (May 2002): 81–83. <https://doi.org/10.1109/MIS.2002.1005635>.
- Helbing, Dirk, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, and Andrej Zwitter. “Will Democracy Survive Big Data and Artificial Intelligence?” *Scientific American*, 2017. <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/>.
- Helfer, Laurence. “Regime Shifting: The TRIPs Agreement and New Dynamics of International Intellectual Property Lawmaking.” *Yale Journal of International Law* 29 (January 1, 2004): 1–83.
- Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. “Measuring Massive Multitask Language Understanding.” *ArXiv:2009.03300 [Cs]*, September 7, 2020. <http://arxiv.org/abs/2009.03300>.
- Henriksen, Anders. *International Law*. Oxford University Press, n.d.
- Hernandez, Danny, and Tom Brown. “Measuring the Algorithmic Efficiency of Neural Networks.” OpenAI, 2020. https://cdn.openai.com/papers/ai_and_efficiency.pdf.
- Hernandez-Orallo, Jose, Fernando Martinez-Plumed, and Shahar Avin. “Surveying Safety-Relevant AI Characteristics.” In *Proceedings of 1st AAAI’s Workshop on Artificial Intelligence Safety (SafeAI)*, 9. Honolulu, Hawaii, 2019. http://ceur-ws.org/Vol-2301/paper_22.pdf.
- Hernandez-Orallo, Jose, Fernando Martinez-Plumed, Shahar Avin, and Jess Whittlestone. “AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues,” 8. Santiago de Compostela, Spain, 2020. http://ecai2020.eu/papers/1364_paper.pdf.
- Herz, John H. “Idealist Internationalism and the Security Dilemma.” *World Politics* 2, no. 2 (January 1950): 157–80. <https://doi.org/10.2307/2009187>.
- Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostafa Ali Patwary, Yang Yang, and Yanqi Zhou. “Deep Learning Scaling Is Predictable, Empirically.” *ArXiv:1712.00409 [Cs, Stat]*, December 1, 2017. <http://arxiv.org/abs/1712.00409>.
- Heyns, Christof. “Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions.” United Nations General Assembly - Human Rights Council, April 9, 2013. https://doi.org/10.1163/2210-7975_HRD-9970-2016149.
- High-level Panel on Digital Cooperation. “Report of the Secretary-General: Roadmap for Digital Cooperation.” UN Secretary-General’s High-Level Panel on Digital Cooperation, June 2020. https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/Roadmap_for_Digital_Cooperation_EN.pdf.
- Hildebrandt, Mireille. “Law As Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, June 7, 2017. <https://papers.ssrn.com/abstract=2983045>.
- Himmelreich, Johannes. “Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations.” *Ethical Theory and Moral Practice* 21, no. 3 (June 2018): 669–84. <https://doi.org/10.1007/s10677-018-9896-4>.
- Hitaj, Briland, Paolo Gasti, Giuseppe Ateniese, and Fernando Perez-Cruz. “PassGAN: A Deep Learning Approach for Password Guessing.” *ArXiv:1709.00440 [Cs, Stat]*, September 1, 2017. <http://arxiv.org/abs/1709.00440>.
- Hitchens, Theresa. “Kill Chain In The Sky With Data: Army’s Project Convergence.” *Breaking Defense* (blog), September 14, 2020. <https://breakingdefense.com/2020/09/kill-chain-in-the-sky-with-data-armys-project-convergence/>.
- Hoffman, David. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. 1 edition. New York: Anchor, 2010.
- Hoffman, Moshe, Erez Yoeli, N. Aygun Dalkiran, and Martin A. Nowak. “Why Norms Are Categorical.” *PNAS*, February 17, 2018. https://www.tse-fr.eu/sites/default/files/TSE/documents/sem2018/eco_pub/hoffman.pdf.
- Hofmann, Stephanie C. “Overlapping Institutions in the Realm of International Security: The Case of NATO and ESDP.” *Perspectives on Politics* 7, no. 1 (March 2009): 45–52. <https://doi.org/10.1017/S1537592709090070>.
- Hofstadter, Douglas. “The Shallowness of Google Translate.” *The Atlantic*, January 30, 2018. <https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/>.
- Hogarth, Ian. “AI Nationalism.” *Ian Hogarth* (blog), June 13, 2018. <https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism>.
- Hogendorf, Christiaan, and Brett Frischmann. “Infrastructure and General Purpose Technologies: A Technology Flow Framework.” *European Journal of Law and Economics*, February 18, 2020. <https://doi.org/10.1007/s10657-020-09642-w>.
- Holmes, Donald B. *Wilbur’s Story*. Lulu Enterprises, 2008.
- Höne, Katharina E. “Mapping the Challenges and Opportunities of Artificial Intelligence for the Conduct of Diplomacy.” Brussels: DiploFoundation, January 2019. <https://www.diplomacy.edu/sites/default/files/AI-diplo-report.pdf>.
- Horowitz, Michael C. “Artificial Intelligence, International Competition, and the Balance of Power.” *Texas National Security Review*, May 15, 2018. <https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/>.
- . “Do Emerging Military Technologies Matter for International Politics?” *Annual Review of Political Science* 23, no. 1 (2020): 385–400. <https://doi.org/10.1146/annurev-polisci-050718-032725>.
- . “Public Opinion and the Politics of the Killer Robots Debate.” *Research & Politics* 3, no. 1 (February 12, 2016): 2053168015627183. <https://doi.org/10.1177/2053168015627183>.
- . “When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability.” *Journal of*

- Strategic Studies* 42, no. 6 (September 19, 2019): 764–88.
<https://doi.org/10.1080/01402390.2019.1621174>.
- . “Who’ll Want Artificially Intelligent Weapons? ISIS, Democracies, or Autocracies?” *Bulletin of the Atomic Scientists*, July 29, 2016.
<http://thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692>.
- Horowitz, Michael C. “Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systems.” *Temp. Int'l & Comp* 30 (2016): 14.
- Horowitz, Michael C., Lauren Kahn, and Casey Mahoney. “The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?” *Orbis*, September 14, 2020.
<https://doi.org/10.1016/j.orbis.2020.08.003>.
- Horowitz, Michael C., Sarah E. Kreps, and Matthew Fuhrmann. “Separating Fact from Fiction in the Debate over Drone Proliferation.” *International Security* 41, no. 2 (October 2016): 7–42.
https://doi.org/10.1162/ISEC_a_00257.
- Horowitz, Michael C., Paul Scharre, and Alexander Velez-Green. “A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence.” *ArXiv:1912.05291 [Cs]*, December 13, 2019.
<http://arxiv.org/abs/1912.05291>.
- Horvitz, Eric, and Mustafa Suleyman. “Introduction from the Founding Co-Chairs.” The Partnership on AI, September 28, 2016.
<https://www.partnershiponai.org/introduction-from-the-founding-co-chairs/>.
- Hotten, Russell. “Volkswagen: The Scandal Explained.” *BBC News*, December 10, 2015, sec. Business.
<https://www.bbc.com/news/business-34324772>.
- Howard, Scarlett R., Aurore Avarguès-Weber, Jair E. Garcia, Andrew D. Greentree, and Adrian G. Dyer. “Numerical Ordering of Zero in Honey Bees.” *Science* 360, no. 6393 (June 8, 2018): 1124–26.
<https://doi.org/10.1126/science.aar4975>.
- Hultman, Nathan, and Jonathan Koomey. “Three Mile Island: The Driver of US Nuclear Power’s Decline?” *Bulletin of the Atomic Scientists* 69, no. 3 (May 1, 2013): 63–70.
<https://doi.org/10.1177/0096340213485949>.
- Human Rights Watch. *Losing Humanity: The Case against Killer Robots*. Amsterdam, Berlin: Human Rights Watch, 2012.
https://www.hrw.org/sites/default/files/reports/arms112_ForUpload.pdf.
- . “Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition: Memorandum to Convention on Conventional Weapons Delegates.” Human Rights Watch, November 8, 2015.
<https://www.hrw.org/news/2015/11/08/precedent-preemption-ban-blinding-lasers-model-killer-robots-prohibition>.
- Hürtgen, Holger, Sebastian Kerkhoff, Jan Lubatschowski, and Manuel Möller. “Rethinking AI Talent Strategy as AutoML Comes of Age.” McKinsey, August 14, 2020. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/rethinking-ai-talent-strategy-as-automated-machine-learning-comes-of-age>.
- Hutchinson, Matthew, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Chansup Byun, Micheal Houle, et al. “Accuracy and Performance Comparison of Video Action Recognition Approaches.” *ArXiv:2008.09037 [Cs]*, August 20, 2020.
<http://arxiv.org/abs/2008.09037>.
- Hutson, Matthew. “AI Researchers Allege That Machine Learning Is Alchemy.” *Science | AAAS*, May 3, 2018.
<https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy>.
- . “Artificial Intelligence Faces Reproducibility Crisis.” *Science* 359, no. 6377 (February 16, 2018): 725–26. <https://doi.org/10.1126/science.359.6377.725>.
- . “Eye-Catching Advances in Some AI Fields Are Not Real.” *Science | AAAS*, May 27, 2020.
<https://www.sciencemag.org/news/2020/05/eye-catching-advances-some-ai-fields-are-not-real>.
- Hwang, Tim. “Computational Power and the Social Impact of Artificial Intelligence,” March 23, 2018.
<https://arxiv.org/abs/1803.08971>.
- IEEE. “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems.” IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, March 2019.
<https://engagestandards.ieee.org/rs/211-FYL-955/images/EAD1e.pdf>.
- Imbrie, Andrew, James Dunham, Rebecca Gelles, and Catherine Aiken. “Mainframes: A Provisional Analysis of Rhetorical Frames in AI.” Center for Security and Emerging Technology, August 2020.
<https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/>.
- Inoue, Sana, and Tetsuro Matsuzawa. “Working Memory of Numerals in Chimpanzees.” *Current Biology* 17, no. 23 (December 4, 2007): R1004–5.
<https://doi.org/10.1016/j.cub.2007.10.027>.
- Insinna, Valerie. “The US Air Force Has Built and Flown a Mysterious Full-Scale Prototype of Its Future Fighter Jet.” Defense News, September 15, 2020.
<https://www.defensenews.com/breaking-news/2020/09/15/the-us-air-force-has-built-and-flown-a-mysterious-full-scale-prototype-of-its-future-fighter-jet/>.
- Interim Agreement between the US and the USSR on Certain Measures with Respect to the Limitation of Strategic Offensive Arms (SALT I Interim Agreement), 944 UNTS 3 § (1972).
- International Convention for the Suppression of Terrorist Bombings, 2149 UNTS 256 § (1997).
- International Court of Justice. Corfu Channel (Merits) (UK v Albania) (International Court of Justice 1949).
- . *Legality of the Threat or Use of Nuclear Weapons: Advisory Opinion of 8 July 1996*. The Hague: International Court of Justice, 1996. <https://www.icj-cij.org/files/case-related/95/095-19960708-ADV-01-00-EN.pdf>.
- International Law Commission. “Draft Articles on Responsibility of States for Internationally Wrongful Acts.” United Nations, 2001.
https://legal.un.org/ilc/texts/instruments/english/commentaries/9_6_2001.pdf.
- . “Draft Articles on State Responsibility, Part 2.” Report of the ILC to the UNGA, UN Doc. A/44/10, 1989.
- International Panel on the Regulation of Autonomous Weapons (iPRAW). “LAWS and Export Control Regimes: Fit for Purpose?,” April 2020.

- <https://www.ipraw.org/working-paper-diffusion-export-control/>.
- IPSOS. "Six in Ten (61%) Respondents Across 26 Countries Oppose the Use of Lethal Autonomous Weapons Systems," January 22, 2019. https://www.ipsos.com/sites/default/files/ct/news/documents/2019-01/human-rights-watch-autonomous-weapons-pr-01-22-2019_0.pdf.
- . "Three in Ten Americans Support Using Autonomous Weapons." Ipsos, February 7, 2017. <https://www.ipsos.com/en-us/news-polls/three-ten-americans-support-using-autonomous-weapons>.
- Irpan, Alex. "My AI Timelines Have Sped Up," August 2020. <http://www.alexirpan.com/2020/08/18/ai-timelines.html>.
- Israel Aerospace Industries. "Harpy Loitering Weapon." Accessed September 11, 2020. <https://www.iai.co.il/p/harpy>.
- Israel, Brian. "Treaty Stasis." *AJIL Unbound* 108 (ed 2014): 63–69. <https://doi.org/10.1017/S2398772300001860>.
- ITU. "AI for Good Global Summit 2019 Insights," 2019. <https://itu.foleon.com/itu/aiforgood2019/home/>.
- . "Artificial Intelligence for Global Good." Geneva: ITU, 2018. https://www.itu.int/en/itunews/Documents/2018/2018-01/2018_ITUNews01-en.pdf.
- . "Measuring Digital Development: Facts and Figures 2019." ITU, 2019. <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2019.pdf>.
- . "United Nations Activities on Artificial Intelligence (AI)." ITU, 2018. https://www.itu.int/dms_pub/itu-s/obp/gen/S-GEN-UNACT-2018-1-PDF-E.pdf.
- . "United Nations Activities on Artificial Intelligence (AI) 2019." ITU, 2019. https://www.itu.int/dms_pub/itu-s/obp/gen/S-GEN-UNACT-2019-1-PDF-E.pdf.
- Jaques, Abby Everett. "Why the Moral Machine Is a Monster," 2019, 10.
- Jassby, Daniel. "Fusion Reactors: Not What They're Cracked up to Be." *Bulletin of the Atomic Scientists* (blog), April 19, 2017. <https://thebulletin.org/2017/04/fusion-reactors-not-what-theyre-cracked-up-to-be/>.
- Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. "Coordinating Committee for the Governance of Artificial Intelligence." G20 - Policy Brief Taskforce 5 (Multilateralism), June 14, 2020. https://www.g20-insights.org/policy_briefs/coordinating-committee-for-the-governance-of-artificial-intelligence/.
- Jenkins, Bonnie. "A Farewell to the Open Skies Treaty, and an Era of Imaginative Thinking." *Brookings* (blog), June 16, 2020. <https://www.brookings.edu/blog/order-from-chaos/2020/06/16/a-farewell-to-the-open-skies-treaty-and-an-era-of-imaginative-thinking/>.
- Jenks, C. Wilfred. "The New Science and the Law of Nations." *The International and Comparative Law Quarterly* 17, no. 2 (1968): 327–45.
- Jensen, Benjamin, Scott Cuomo, and Chris Whyte. "Wargaming with Athena: How to Make Militaries Smarter, Faster, and More Efficient with Artificial Intelligence." *War on the Rocks*, June 5, 2018. <https://warontherocks.com/2018/06/wargaming-with-athena-how-to-make-militaries-smarter-faster-and-more-efficient-with-artificial-intelligence/>.
- Jensen, Benjamin M., Christopher Whyte, and Scott Cuomo. "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence." *International Studies Review* 22, no. 3 (September 1, 2020): 526–50. <https://doi.org/10.1093/isr/viz025>.
- Jensen, Eric Talbot. "The Future of the Law of Armed Conflict: Ostriches, Butterflies, and Nanobots." *Michigan Journal of International Law* 35, no. 2 (2014): 253–317.
- Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence*, September 2, 2019, 1–11. <https://doi.org/10.1038/s42256-019-0088-2>.
- Joerges, Bernward. "Do Politics Have Artefacts?" *Social Studies of Science*, June 29, 2016. <https://doi.org/10.1177/030631299029003004>.
- Joh, Elizabeth E. "Policing the Smart City." *International Journal of Law in Context* 15, no. 2 (June 2019): 177–82. <https://doi.org/10.1017/S1744552319000107>.
- Johnson, James. "Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?" *Journal of Strategic Studies* 0, no. 0 (April 30, 2020): 1–39. <https://doi.org/10.1080/01402390.2020.1759038>.
- Johnson, Keith. "U.S. Effort to Depart WTO Gathers Momentum." *Foreign Policy* (blog), 2020. <https://foreignpolicy.com/2020/05/27/world-trade-organization-united-states-departure-china/>.
- Jones, Benjamin F. "The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder?" *The Review of Economic Studies* 76, no. 1 (January 1, 2009): 283–317. <https://doi.org/10.1111/j.1467-937X.2008.00531.x>.
- Jones, Meg Leta. "Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw." *Journal of Law, Technology & Policy* 2 (2018). <https://doi.org/10.2139/ssrn.2981855>.
- Jordan, Michael I. "Artificial Intelligence—The Revolution Hasn't Happened Yet." *Harvard Data Science Review* 1, no. 1 (June 21, 2019). <https://doi.org/10.1162/99608f92.f06c6e61>.
- Jupe, Louise Marie, and David Adam Keatley. "Airport Artificial Intelligence Can Detect Deception: Or Am I Lying?" *Security Journal*, September 24, 2019. <https://doi.org/10.1057/s41284-019-00204-7>.
- Jupille, Joseph, Walter Mattli, and Duncan Snidal. *Institutional Choice and Global Commerce*. Cambridge: Cambridge University Press, 2013.
- Kacowicz, Arie M. "Global Governance, International Order, and World Order." *The Oxford Handbook of Governance*, March 29, 2012. <https://doi.org/10.1093/oxfordhb/9780199560530.013.0048>.
- Kahneman, Daniel, and Dan Lovallo. "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking." *Management Science* 39, no. 1 (1993): 17–31.
- Kajonius, Petri J., and Therese Björkman. "Individuals with Dark Traits Have the Ability but Not the Disposition to Empathize." *Personality and Individual Differences* 155 (March 1, 2020): 109716. <https://doi.org/10.1016/j.paid.2019.109716>.
- Kallenborn, Zachary. "AI Risks to Nuclear Deterrence Are Real." War on the Rocks, October 10, 2019. <https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/>.
- . "Are Drone Swarms Weapons of Mass Destruction?" The Counterproliferation Papers, Future Warfare Series. U.S. Air Force Center for

- Strategic Deterrence Studies, Air University, May 6, 2020.
<https://media.defense.gov/2020/Jun/29/2002331131/1-1/0/60DRONESWARMS-MONOGRAPH.PDF>.
- Kallenborn, Zachary, and Philipp C. Bleek. "Swarming Destruction: Drone Swarms and Chemical, Biological, Radiological, and Nuclear Weapons." *The Nonproliferation Review* 25, no. 5–6 (September 2, 2018): 523–43.
<https://doi.org/10.1080/10736700.2018.1546902>.
- Kaminski, Margot E. "Authorship, Disrupted: AI Authors in Copyright and First Amendment Law." *UC Davis Law Review* 51, no. 589 (2017): 589–616.
- Kaminski, Margot E. "Authorship, Disrupted: AI Authors in Copyright and First Amendment Law Symposium - Future-Proofing Law: From RDNA to Robots (Part 2)." *U.C. Davis Law Review* 51 (2018 2017): 589–616.
- Kamyshev, Pasha. "Machine Learning In The Judicial System Is Mostly Just Hype." *Palladium Magazine* (blog), March 30, 2019.
<https://palladiummag.com/2019/03/29/machine-learning-in-the-judicial-system-is-mostly-just-hype/>.
- Kanellos, Michael. "Moore's Law to Roll on for Another Decade." *CNET*, February 11, 2003.
<https://www.cnet.com/news/moores-law-to-roll-on-for-another-decade/>.
- Kania, Elsa. "AlphaGo and Beyond: The Chinese Military Looks to Future 'Intelligentized' Warfare." Lawfare, June 5, 2017. <https://www.lawfareblog.com/alphago-and-beyond-chinese-military-looks-future-intelligentized-warfare>.
- . "The Pursuit of AI Is More Than an Arms Race." *Defense One*, April 19, 2018.
<https://www.defenseone.com/ideas/2018/04/pursuit-ai-more-arms-race/147579/>.
- . "数字化 – 网络化 – 智能化: China's Quest for an AI Revolution in Warfare." *The Strategy Bridge* (blog), June 8, 2017. <https://thestrategybridge.org/the-bridge/2017/6/8/chinas-quest-for-an-ai-revolution-in-warfare>.
- Kania, Elsa B. "Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power." Center for a New American Security, November 2017.
<https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235804>.
- Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling Laws for Neural Language Models." *ArXiv:2001.08361 [Cs, Stat]*, January 22, 2020.
<http://arxiv.org/abs/2001.08361>.
- Karnofsky, Holden. "Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity." Open Philanthropy Project, 2016.
<http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity>.
- Katzenstein, Peter, and Rudra Sil. "Eclectic Theorizing in the Study and Practice of International Relations." In *The Oxford Handbook of International Relations*, edited by Christian Reus-Smit and Rudra Sil, 2008.
[https://www.oxfordhandbooks.com/view/10.1093/oxfordhb-9780199219322.001.0001/oxfordhb-9780199219322-e-6](https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199219322.001.0001/oxfordhb-9780199219322-e-6).
- Kaufmann, Mareile, Simon Egbert, and Matthias Leese. "Predictive Policing and the Politics of Patterns." *The British Journal of Criminology* 59, no. 3 (2018): 674–92. <https://doi.org/10.1093/bjc/azy060>.
- Kazakova, Snejha, Allison Dunne, Daan Bijwaard, Julien Gossé, Charles Hoffremon, and Nicolas van Zeebroeck. "European Enterprise Survey on the Use of Technologies Based on Artificial Intelligence." Ipsos Belgium and iCite, July 28, 2020.
<https://ec.europa.eu/digital-single-market/en/news/european-enterprise-survey-use-technologies-based-artificial-intelligence>.
- Keeley, James F. "The Latest Wave: A Critical Review of Regime Literature." In *World Politics: Power, Interdependence and Dependence*, edited by David G. Haglund and Michael K. Hawes, 553–69. Toronto: Harcourt, Brace, Jovanovich Canada, 1990.
- Keeling, Geoff. "Why Trolley Problems Matter for the Ethics of Automated Vehicles." *Science and Engineering Ethics* 26, no. 1 (February 1, 2020): 293–307. <https://doi.org/10.1007/s11948-019-00096-1>.
- Kellow, Aynsley. "Multi-Level and Multi-Arena Governance: The Limits of Integration and the Possibilities of Forum Shopping." *International Environmental Agreements: Politics, Law and Economics* 12, no. 4 (2012): 327–42.
- Kelly, Kevin. "The Myth of a Superhuman AI." Backchannel, April 25, 2017.
<https://backchannel.com/the-myth-of-a-superhuman-ai-59282b686c62>.
- Kemp, Luke. "Critical Mass Governance: Addressing US Participation in Environmental Multilateralism." Australian National University, 2016.
- . "Fragmented, Incidental and Inadequate: Mapping the Archipelago of AGI Governance," Forthcoming.
- . "US-Proving the Paris Climate Agreement." *Climate Policy* 17, no. 1 (January 2, 2017): 86–101.
<https://doi.org/10.1080/14693062.2016.1176007>.
- Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigearthaigh. "UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance." UN High-Level Panel on Digital Cooperation, February 26, 2019.
https://digitalcooperation.org/wp-content/uploads/2019/02/Luke_Kemp_Submission-to-the-UN-High-Level-Panel-on-Digital-Cooperation-2019-Kemp-et-al.pdf.
- Kennan, George F. "To Prevent a World Wasteland: A Proposal." *Foreign Affairs*, 1970.
- Keohane, Robert O., and David G. Victor. "The Regime Complex for Climate Change." *Perspectives on Politics* 9, no. 1 (2011): 7–23.
<https://doi.org/10.1017/s1537592710004068>.
- Kerr, Orin S. "A Theory of Law." *GREEN BAG* 16 (2012): 111.
- Kerr, Paul K. "U.S.-Proposed Missile Technology Control Regime Changes." Congressional Research Service, July 27, 2020.
<https://fas.org/sgp/crs/nuke/IF11069.pdf>.
- Kertysova, Katarina. "Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation Is Produced, Disseminated, and Can Be Counteracted." *Security and Human Rights* 29, no. 1–4 (December 12, 2018): 55–81.
<https://doi.org/10.1163/18750230-02901005>.

- Khan, Hassan N., David A. Hounshell, and Erica R. H. Fuchs. "Science and Research Policy at the End of Moore's Law." *Nature Electronics* 1, no. 1 (January 2018): 14. <https://doi.org/10.1038/s41928-017-0005-9>.
- Khan, Lina M. "Amazon's Antitrust Paradox." *The Yale Law Journal* 126 (2017): 710–805.
- Khan, Saif M., and Alexander Mann. "AI Chips: What They Are and Why They Matter." Center for Security and Emerging Technology, April 2020. <https://cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/>.
- Kim, Rakhyun E. "Is Global Governance Fragmented, Polycentric, or Complex? The State of the Art of the Network Approach." *International Studies Review*, 2019. <https://doi.org/10.1093/isr/viz052>.
- Kim, Younghwan, Minki Kim, and Wonjoon Kim. "Effect of the Fukushima Nuclear Disaster on Global Public Acceptance of Nuclear Energy." *Energy Policy* 61 (October 1, 2013): 822–28. <https://doi.org/10.1016/j.enpol.2013.06.107>.
- Kimball, Daryl G. "U.S. Aims to Expand Drone Sales." Arms Control Association, August 2020. <https://www.armscontrol.org/act/2020-07/news/us-aims-expand-drone-sales>.
- King, Thomas C., Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions." *Science and Engineering Ethics*, February 14, 2019. <https://doi.org/10.1007/s11948-018-00081-0>.
- Kirchner, Lauren, Julia Angwin, Jeff Larson, and Surya Mattu. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." ProPublica, May 23, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Kirilenko, Andrei A., Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market." *SSRN Electronic Journal*, 2014. <https://doi.org/10.2139/ssrn.1686004>.
- Klare, Michael T. "'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation." *Arms Control Today; Washington* 50, no. 3 (April 2020): 10–15.
- Kleeman, Alexandra. "Cooking with Chef Watson, I.B.M.'s Artificial-Intelligence App." *The New Yorker*, 2016. <http://www.newyorker.com/magazine/2016/11/28/cooking-with-chef-watson-ibms-artificial-intelligence-app>.
- Klein, Natalie. "Maritime Autonomous Vehicles within the International Law Framework to Enhance Maritime Security." *International Law Studies* 95 (2019): 29.
- Klijn, Hugo, and Maaike Okano-Heijmans. "Managing Robotic and Autonomous Systems (RAS)." The Hague: The Hague Centre for Strategic Studies, March 17, 2020. <https://www.clingendael.org/publication/managing-robotic-and-autonomous-systems-ras>.
- Klinger, Joel, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. "A Narrowing of AI Research?" *ArXiv:2009.10385 [Cs]*, September 23, 2020. <http://arxiv.org/abs/2009.10385>.
- Knight, Will. "A Dogfight Renews Concerns About AI's Lethal Potential." *Wired*, August 25, 2020. <https://www.wired.com/story/dogfight-renews-concerns-ai-lethal-potential/>.
- . "AI Is All the Rage. So Why Aren't More Businesses Using It?" *Wired*, July 30, 2020. <https://www.wired.com/story/ai-why-not-more-businesses-use/>.
- . "Alpha Zero's 'Alien' Chess Shows the Power, and the Peculiarity, of AI." *MIT Technology Review*, December 8, 2017. <https://www.technologyreview.com/s/609736/alpha-zeros-alien-chess-shows-the-power-and-the-peculiarity-of-ai/>.
- . "Anduril's New Drone Offers to Inject More AI Into Warfare." *Wired*, September 10, 2020. <https://www.wired.com/story/anduril-new-drone-inject-ai-warfare/>.
- . "Baidu Breaks Off an AI Alliance Amid Strained US-China Ties." *Wired*, June 18, 2020. <https://www.wired.com/story/baidu-breaks-ai-alliance-strained-us-china-ties/>.
- . "Facebook's Head of AI Says the Field Will Soon 'Hit the Wall.'" *Wired*, 2019. <https://www.wired.com/story/facebook-ai-says-field-hit-wall/>.
- Koblentz, Gregory C., and Paul F. Walker. "Can Bill Gates Rescue the Bioweapons Convention?" *Bulletin of the Atomic Scientists* (blog), April 3, 2017. <https://thebulletin.org/2017/04/can-bill-gates-rescue-the-bioweapons-convention/>.
- Koenig, Alexa. "'Half the Truth Is Often a Great Lie': Deep Fakes, Open Source Information, and International Criminal Law." *AJIL Unbound* 113 (2019): 250–55. <https://doi.org/10.1017/ajlu.2019.47>.
- Kohda, Masanori, Hatta Takashi, Tmohiro Takeyama, Satoshi Awata, Hirokazu Tanaka, Jun-ya Asai, and Alex Jordan. "Cleaner Wrasse Pass the Mark Test. What Are the Implications for Consciousness and Self-Awareness Testing in Animals?" *BioRxiv*, August 21, 2018, 397067. <https://doi.org/10.1101/397067>.
- Koplow, David A. "Back to the Future and up to the Sky: Legal Implications of Open Skies Inspection for Arms Control." *California Law Review* 79 (1991): 421–96.
- Koppelman, Ben. "How Would Future Autonomous Weapon Systems Challenge Current Governance Norms?" *The RUSI Journal* 164, no. 5–6 (September 19, 2019): 98–109. <https://doi.org/10.1080/03071847.2019.1694261>.
- Koremenos, Barbara, Charles Lipson, and Duncan Snidal. "Rational Design: Looking Back to Move Forward." *International Organization* 55, no. 4 (ed 2001): 1051–82. <https://doi.org/10.1162/002081801317193691>.
- . "The Rational Design of International Institutions." *International Organization* 55, no. 4 (ed 2001): 761–99. <https://doi.org/10.1162/002081801317193592>.
- Koskeniemi, Martti. "Law, Teleology and International Relations: An Essay in Counterdisciplinarity." *International Relations* 26, no. 1 (March 1, 2012): 3–34. <https://doi.org/10.1177/0047117811433080>.
- Koskeniemi, Martti, and Study Group of the International Law Commission. "Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law." United Nations - General Assembly, April 13, 2006. http://legal.un.org/ilc/documentation/english/a_cn4_1682.pdf.
- Kott, Alexander. "Challenges and Characteristics of Intelligent Autonomy for Internet of Battle Things in Highly Adversarial Environments." US Army

- Research Laboratory, 2018.
<https://arxiv.org/ftp/arxiv/papers/1803/1803.11256.pdf>
- Kott, Alexander, and Philip Perconti. "Long-Term Forecasts of Military Technologies for a 20-30 Year Horizon: An Empirical Assessment of Accuracy." *ArXiv:1807.08339 [Cs]*, July 22, 2018.
<http://arxiv.org/abs/1807.08339>.
- Koubi, Vally. "International Tensions and Arms Control Agreements." *American Journal of Political Science* 37, no. 1 (1993): 148–64.
<https://doi.org/10.2307/2111527>.
- Krafft, P. M., Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. "Defining AI in Policy versus Practice." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. New York NY USA: ACM, 2020.
<https://doi.org/10.1145/3375627.3375835>.
- Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. "Specification gaming: the flip side of AI ingenuity." *Deepmind* (blog), April 21, 2020.
<https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity>.
- Krasner, Stephen D. *International Regimes*. Cornell University Press, 1983.
- . "Structural Causes and Regime Consequences: Regimes as Intervening Variables." *International Organization* 36, no. 2 (1982): 185–205.
<https://doi.org/10.1017/S0020818300018920>.
- Krass, Allan S. "Arms Control Treaty Verification." In *Encyclopedia of Arms Control and Disarmament*, edited by Richard Dean Burns, 1:297–316. Charles Scribner's Sons, 1993.
- Kratsios, Michael. "Artificial Intelligence Can Serve Democracy." *Wall Street Journal*, May 27, 2020, sec. Opinion. <https://www.wsj.com/articles/artificial-intelligence-can-serve-democracy-11590618319>.
- Kreil, Michael. "The Army That Never Existed: The Failure of Social Bots Research." Presented at the OpenFest Conference, Sofia, Bulgaria, November 2, 2019. <https://michaelkreil.github.io/openbots/>.
- Kreuder-Sonnen, Christian, and Michael Zürn. "After Fragmentation: Norm Collisions, Interface Conflicts, and Conflict Management." *Global Constitutionalism* 9, no. 2 (July 2020): 241–67.
<https://doi.org/10.1017/S2045381719000315>.
- Krisch, Nico. "The Decay of Consent: International Law in an Age of Global Public Goods." *American Journal of International Law* 108, no. 1 (January 2014): 1–40.
<https://doi.org/10.5305/amerintlaw.108.1.0001>.
- Krishna, Arvind. "IBM CEO's Letter to Congress on Racial Justice Reform." THINKPolicy Blog, June 8, 2020. <https://www.ibm.com/blogs/policy/facial-recognition-susset-racial-justice-reforms/>.
- Krishnan, Armin. *Killer Robots: Legality and Ethicality of Autonomous Weapons*. 1 edition. Farnham, England ; Burlington, VT: Routledge, 2009.
- Kristensen, Hans M., Matthew McKinzie, and Theodore A. Postol. "How US Nuclear Force Modernization Is Undermining Strategic Stability: The Burst-Height Compensating Super-Fuze." *Bulletin of the Atomic Scientists* (blog), March 1, 2017.
<https://thebulletin.org/how-us-nuclear-force-modernization-undermining-strategic-stability-burst-height-compensating-super10578>.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105. Curran Associates, Inc., 2012. <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- Kroenig, Matthew, and Bharath Gopalaswamy. "Will Disruptive Technology Cause Nuclear War?" *Bulletin of the Atomic Scientists* (blog), November 12, 2018.
<https://thebulletin.org/2018/11/will-disruptive-technology-cause-nuclear-war/>.
- Kroll, Joshua A. "The Fallacy of Inscrutability." *Phil. Trans. R. Soc. A* 376, no. 2133 (November 28, 2018): 20180084. <https://doi.org/10.1098/rsta.2018.0084>.
- Kumar, Ram Shankar Siva, David O. Brien, Kendra Albert, Salomé Viljöen, and Jeffrey Snover. "Failure Modes in Machine Learning Systems." *ArXiv:1911.11034 [Cs, Stat]*, November 25, 2019.
<http://arxiv.org/abs/1911.11034>.
- Kunz, Martina. "Global Artificial Intelligence Governance: Rationale, Architecture and Challenges," 10. Beijing, 2019.
- Kunz, Martina, and Seán Ó hÉigeartaigh. "Artificial Intelligence and Robotization." In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2020.
<https://papers.ssrn.com/abstract=3310421>.
- Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial Machine Learning at Scale." *ArXiv:1611.01236 [Cs, Stat]*, February 10, 2017.
<http://arxiv.org/abs/1611.01236>.
- Kurzweil, Ray. *The Singularity Is Near*. Viking Press, 2005.
- Lachow, Irving. "The Upside and Downside of Swarming Drones." *Bulletin of the Atomic Scientists* 73, no. 2 (March 4, 2017): 96–101.
<https://doi.org/10.1080/00963402.2017.1290879>.
- Lanier, Jaron, and Glen Weyl. "AI Is An Ideology, Not A Technology." *Wired*, March 15, 2020.
<https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/>.
- Lasry, Brigitte, Hael Kobayashi, Unesco, and NetExplor. *Human Decisions: Thoughts on AI*. UNESCO, United Nations Educational, Scientific and Cultural Organization/International Bureau of Education, 2018.
<http://unesdoc.unesco.org/images/0026/002615/261563e.pdf>.
- Laurie, Graeme, Shawn H. E. Harmon, and Fabiana Arzuaga. "Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty." *Law, Innovation & Technology* 4, no. 1 (July 2012): 1–33.
- LawGeex. "Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts." LawGeex, February 2018.
<https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf>.
- Leal, Natalie. "Council of Europe Starts Work on Legally-Binding AI Treaty." *Global Government Forum* (blog), July 22, 2020.

- <https://www.globalgovernmentforum.com/council-of-europe-starts-work-on-legally-binding-ai-treaty/>.
- Lee, Kai-Fu. *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston: Houghton Mifflin Harcourt, 2018.
- Leenes, Ronald, Erica Palmerini, Bert-Jaap Koops, Andrea Bertolini, Pericle Salvini, and Federica Lucivero. "Regulatory Challenges of Robotics: Some Guidelines for Addressing Legal and Ethical Issues." *Law, Innovation and Technology* 9, no. 1 (January 2, 2017): 1–44.
<https://doi.org/10.1080/17579961.2017.1304921>.
- Legg, Shane, and Marcus Hutter. "A Collection of Definitions of Intelligence." *ArXiv:0706.3639 [Cs]*, June 25, 2007. <http://arxiv.org/abs/0706.3639>.
- . "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17, no. 4 (December 1, 2007): 391–444.
<https://doi.org/10.1007/s11023-007-9079-x>.
- Lehman, Joel, Jeff Clune, Dusan Misevic, Christoph Adami, Julie Beaulieu, Peter J. Bentley, Samuel Bernard, et al. "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities." *ArXiv:1803.03453 [Cs]*, March 9, 2018. <http://arxiv.org/abs/1803.03453>.
- Lehr, David, and Paul Ohm. "Playing with the Data: What Legal Scholars Should Learn About Machine Learning." *UC Davis Law Review* 51 (2017): 653.
- Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. "AI Safety Gridworlds." *ArXiv:1711.09883 [Cs]*, November 27, 2017.
<http://arxiv.org/abs/1711.09883>.
- Leiserson, Charles E., Neil C. Thompson, Joel S. Emer, Bradley C. Kuszmahl, Butler W. Lampson, Daniel Sanchez, and Tao B. Schardl. "There's Plenty of Room at the Top: What Will Drive Computer Performance after Moore's Law?" *Science* 368, no. 6495 (June 5, 2020). <https://doi.org/10.1126/science.aam9744>.
- Lempel, Howie, Robert Wiblin, and Keiran Harris. "Ben Garfinkel on Scrutinising Classic AI Risk Arguments." 80,000 Hours Podcast. Accessed September 26, 2020.
<https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/>.
- Lessig, Lawrence. *Code: And Other Laws of Cyberspace, Version 2.0*. 2nd Revised ed. edition. New York: Basic Books, 2006. <http://codev2.cc/download+remix/Lessig-Codev2.pdf>.
- . "The Law of the Horse: What Cyberlaw Might Teach." *Harvard Law Review* 113, no. 2 (December 1999): 501. <https://doi.org/10.2307/1342331>.
- . "The New Chicago School." *The Journal of Legal Studies* 27, no. S2 (June 1, 1998): 661–91.
<https://doi.org/10.1086/468039>.
- "Letter to Google C.E.O.," 2018.
<https://static01.nyt.com/files/2018/technology/googleletter.pdf>.
- Leung, Jade. "Who Will Govern Artificial Intelligence? Learning from the History of Strategic Politics in Emerging Technologies." University of Oxford, 2019.
<https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665>.
- Leung, Jade, Sophie-Charlotte Fischer, and Allan Dafoe. "Export Controls in the Age of AI." War on the Rocks, August 28, 2019.
<https://warontherocks.com/2019/08/export-controls-in-the-age-of-ai/>.
- Levin, John-Clark, and Matthijs M. Maas. "Roadmap to a Roadmap: How Could We Tell When AGI Is a 'Manhattan Project' Away?" Santiago de Compostela, Spain, 2020.
http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf.
- Lewis, Dustin A., Gabriella Blum, and Naz K. Modirzadeh. "War-Algorithm Accountability." Research Briefing. HARVARD LAW SCHOOL PROGRAM ON INTERNATIONAL LAW AND ARMED CONFLICT, August 31, 2016.
<https://blogs.harvard.edu/pilac/files/2016/09/War-Algorithm-Accountability-Without-Appendices-August-2016.pdf>.
- Lewis-kraus, Gideon. "The Great A.I. Awakening." *The New York Times*, December 14, 2016.
<https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>.
- Leys, Nathan. "Autonomous Weapon Systems, International Crises, and Anticipatory Self-Defense." *The Yale Journal of International Law* 45, no. 2 (2020): 377–411.
- Lieber, Keir A. *War and the Engineers: The Primacy of Politics over Technology*. 1 edition. Ithaca; London: Cornell University Press, 2008.
- Lieber, Keir A., and Daryl G. Press. "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence." *International Security* 41, no. 4 (April 1, 2017): 9–49.
https://doi.org/10.1162/ISEC_a_00273.
- Lim, Daniel. "Killer Robots and Human Dignity," 6, 2019. Limitation and Reduction of Naval Armament (London Naval Treaty), 112 LNTS 65 § (1930).
- Limitation of Naval Armament (Second London Naval Treaty), 197 LNTS 387 § (1937).
- Lin, Herbert. "The Existential Threat from Cyber-Enabled Information Warfare." *Bulletin of the Atomic Scientists* 75, no. 4 (July 4, 2019): 187–96.
<https://doi.org/10.1080/00963402.2019.1629574>.
- Lin, Patrick, Keith Abney, and George A. Bekey. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, 2011.
- Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford, New York: Oxford University Press, 2017.
- Lin, Zhiyuan "Jerry," Jongbin Jung, Sharad Goel, and Jennifer Skeem. "The Limits of Human Predictions of Recidivism." *Science Advances* 6, no. 7 (February 1, 2020): eaaz0652.
<https://doi.org/10.1126/sciadv.aaz0652>.
- Linn, Allison. "Microsoft Researchers Win ImageNet Computer Vision Challenge." *Next at Microsoft* (blog), December 10, 2015.
<https://blogs.microsoft.com/next/2015/12/10/microsoft-researchers-win-imagenet-computer-vision-challenge/>.
- Liu, Han-Wei, and Ching-Fu Lin. "Artificial Intelligence and Global Trade Governance: A Pluralist Agenda." *Harvard International Law Journal* 61, no. 2 (2020). <https://papers.ssrn.com/abstract=3675505>.
- Liu, Hin-Yan. "Categorization and Legality of Autonomous and Remote Weapons Systems." *International Review of the Red Cross* 94, no. 886

- (June 2012): 627–52.
<https://doi.org/10.1017/S181638311300012X>.
- . “From the Autonomy Framework towards Networks and Systems Approaches for ‘Autonomous’ Weapons Systems.” *Journal of International Humanitarian Legal Studies* 10, no. 1 (June 9, 2019): 89–110. <https://doi.org/10.1163/18781527-01001010>.
- . “The Digital Disruption of Human Rights Foundations.” In *Human Rights, Digital Society and the Law: A Research Companion*. London: Routledge, 2019.
- . “The Power Structure of Artificial Intelligence.” *Law, Innovation and Technology* 10, no. 2 (July 3, 2018): 197–229.
<https://doi.org/10.1080/17579961.2018.1527480>.
- . “Three Types of Structural Discrimination Introduced by Autonomous Vehicles.” *UC Davis Law Review* 51 (2018): 32.
- Liu, Hin-Yan, Kristian Cedervall Lauta, and Matthijs M. Maas. “Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research.” *Futures* 102 (2018): 6–19.
<https://doi.org/10.1016/j.futures.2018.04.009>.
- Liu, Hin-Yan, Matthijs Maas, John Danaher, Luisa Scarella, Michaela Lexer, and Leonard Van Rompaey. “Artificial Intelligence and Legal Disruption: A New Model for Analysis.” *Law, Innovation and Technology* 0, no. 0 (September 16, 2020): 1–54.
<https://doi.org/10.1080/17579961.2020.1815402>.
- Liu, Hin-Yan, and Matthijs M. Maas. “Solving for X? Towards a Problem-Finding Framework That Grounds Long-Term Governance Strategies for Artificial Intelligence.” *Futures*, Forthcoming 2020, 35.
- Liu, Hin-Yan, and Andrew Mazibrada. “Artificial Intelligence Affordances: Deep-Fakes as Exemplars of AI Challenges to Criminal Justice Systems.” *UNICRI Special Collection on Artificial Intelligence*, July 2020. <http://unicri.it/towards-responsible-artificial-intelligence-innovation>.
- Liu, Hin-Yan, Léonard Van Rompaey, and Matthijs M. Maas. “Editorial: Beyond Killer Robots: Networked Artificial Intelligence Systems Disrupting the Battlefield?” *Journal of International Humanitarian Legal Studies* 10, no. 1 (June 9, 2019): 77–88.
<https://doi.org/10.1163/18781527-01001014>.
- Liu, Hin-Yan, and Karolina Zawieska. “A New Human Rights Regime to Address Robotics and Artificial Intelligence.” In *2017 Proceedings of the 20th International Legal Informatics Symposium*, 179–84. Oesterreichische Computer Gesellschaft,*, 2017.
- Livingston, Steven, and Sushma Raman. “Human Rights Documentation in Limited Access Areas: The Use of Technology in War Crimes and Human Rights Abuse Investigations.” Cambridge Mass.: Carr Center for Human Rights Policy, May 2018.
https://carrcenter.hks.harvard.edu/files/cchr/files/ccdp_2018_003_humanrightsdocumentation.pdf.
- Livingston, Steven, and Mathias Risse. “The Future Impact of Artificial Intelligence on Humans and Human Rights.” *Ethics & International Affairs* 33, no. 2 (ed 2019): 141–58.
<https://doi.org/10.1017/S089267941900011X>.
- Loeff, Agnes Schim van der, Iggy Bassi, Sachin Kapila, and Jevgenij Gamper. “AI Ethics for Systemic Issues: A Structural Approach.” Vancouver, Canada, 2019.
<http://arxiv.org/abs/1911.03216>.
- LoPucki, Lynn M. “Algorithmic Entities.” *UCLA School of Law, Law-Econ Research Paper*, April 17, 2017.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2954173.
- Lorenz, Philippe. “AI Governance through Political Fora and Standards Developing Organizations.” Stiftung Neue Verantwortung, September 2020.
<https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations>.
- Loss, Rafael, and Joseph Johnson. “Will Artificial Intelligence Imperil Nuclear Deterrence?” War on the Rocks, September 19, 2019.
<https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/>.
- Lowther, Adam, and Curtis McGiffin. “America Needs a ‘Dead Hand.’” War on the Rocks, August 16, 2019.
<https://warontherocks.com/2019/08/america-needs-a-dead-hand/>.
- Lucioni, Alexandra, and Yoshua Bengio. “On the Morality of Artificial Intelligence.” *ArXiv:1912.11945 [Cs]*, December 26, 2019.
<http://arxiv.org/abs/1912.11945>.
- Lutz, Roman. “Learning from the Past to Create Responsible AI: A Collection of Controversial, and Often Unethical AI Use Cases.” Roman Lutz. Accessed June 22, 2020.
<https://romanlutz.github.io/ResponsibleAI/>.
- Lynch, Shana. “Andrew Ng: Why AI Is the New Electricity.” *Stanford Graduate School of Business*, March 11, 2017.
<https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity>.
- Maas, Matthijs M. “How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons.” *Contemporary Security Policy* 40, no. 3 (February 6, 2019): 285–311.
<https://doi.org/10.1080/13523260.2019.1576464>.
- . “Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot.” *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57.
<https://doi.org/10.1163/18781527-01001006>.
- . “International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order.” *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56.
- . “Regulating for ‘Normal AI Accidents’: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment.” In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 223–228. AIES ’18. New York, NY, USA: Association for Computing Machinery, 2018.
<https://doi.org/10.1145/3278721.3278766>.
- Mackenzie, Donald. “Uninventing the Bomb?” *Medicine and War* 12, no. 3 (July 1, 1996): 202–11.
<https://doi.org/10.1080/13623699608409285>.
- MacKenzie, Donald, and Graham Spinardi. “Tacit Knowledge, Weapons Design, and the Uninvention of Nuclear Weapons.” *American Journal of Sociology* 101, no. 1 (1995): 44–99.
- Madsen, Mikael Rask, Pola Cebulak, and Micha Wiebusch. “Backlash against International Courts: Explaining the Forms and Patterns of Resistance to International Courts.” *International Journal of Law*

- in Context* 14, no. 2 (June 2018): 197–220.
<https://doi.org/10.1017/S1744552318000034>.
- Mandel, Gregory N. “Legal Evolution in Response to Technological Change.” *The Oxford Handbook of Law, Regulation and Technology*, July 20, 2017.
<https://doi.org/10.1093/oxfordhb/9780199680832.013.45>.
- Manheim, David. “Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence.” *Big Data and Cognitive Computing* 3, no. 2 (June 2019): 21.
<https://doi.org/10.3390/bdcc3020021>.
- . “Overoptimization Failures and Specification Gaming in Multi-Agent Systems.” *ArXiv:1810.10862 [Cs]*, October 16, 2018.
<http://arxiv.org/abs/1810.10862>.
- . “The Fragile World Hypothesis: Complexity, Fragility, and Systemic Existential Risk.” *Futures*, May 25, 2020.
<https://doi.org/10.1016/j.futures.2020.102570>.
- Manheim, David, and Scott Garrabrant. “Categorizing Variants of Goodhart’s Law.” *ArXiv:1803.04585 [Cs, q-Fin, Stat]*, March 12, 2018.
<http://arxiv.org/abs/1803.04585>.
- Manivasagam, Sivabalan, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. “LiDARsim: Realistic LiDAR Simulation by Leveraging the Real World.” *ArXiv:2006.09348 [Cs]*, June 16, 2020.
<http://arxiv.org/abs/2006.09348>.
- Manson, Janet. “International Law, German Submarines and American Policy.” Portland State University. Department of History, 1977.
<https://doi.org/10.15760/etd.2489>.
- Maras, Marie-Helen, and Alex Alexandrou. “Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos.” *The International Journal of Evidence & Proof*, October 28, 2018, 1365712718807226.
<https://doi.org/10.1177/1365712718807226>.
- Marchant, Gary. “‘Soft Law’ Governance Of Artificial Intelligence.” *AI Pulse* (blog), January 25, 2019.
<https://aipulse.org/soft-law-governance-of-artificial-intelligence/>.
- Marchant, Gary E. “The Growing Gap Between Emerging Technologies and the Law.” In *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*, edited by Gary E. Marchant, Braden R. Allenby, and Joseph R. Herkert, 19–33. The International Library of Ethics, Law and Technology. Dordrecht: Springer Netherlands, 2011.
https://doi.org/10.1007/978-94-007-1356-7_2.
- Marchant, Gary E., and Wendell Wallach. “Coordinating Technology Governance.” *Issues in Science & Technology* 31, no. 4 (Summer 2015): 43–50.
- Marcin, Szczepanski. “Economic Impacts of Artificial Intelligence (AI).” EPRI | European Parliamentary Research Service, July 2019.
[https://europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRI\(2019\)637967_EN.pdf](https://europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRI(2019)637967_EN.pdf).
- Marcus, Gary. “Deep Learning: A Critical Appraisal.” *ArXiv:1801.00631 [Cs, Stat]*, January 2, 2018.
<http://arxiv.org/abs/1801.00631>.
- . “GPT-2 and the Nature of Intelligence.” The Gradient, January 25, 2020.
<https://thegradient.pub/gpt2-and-the-nature-of-intelligence/>.
- Marcus, Gary, and Ernest Davis. “GPT-3, Bloviator: OpenAI’s Language Generator Has No Idea What It’s Talking about.” *MIT Technology Review*, August 22, 2020.
<https://www.technologyreview.com/2020/08/22/100753/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>.
- . *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon, 2019.
- Margulis, Matias E. “The Regime Complex for Food Security: Implications for the Global Hunger Challenge.” *Global Governance* 19, no. 1 (2013): 53–67.
- Marin, Milena, Freddie Kalaitzis, and Buffy Price. “Using Artificial Intelligence to Scale up Human Rights Research: A Case Study on Darfur.” Amnesty International & Citizen Evidence Lab, July 6, 2020.
<https://citizenevidence.org/2020/07/06/using-artificial-intelligence-to-scale-up-human-rights-research-a-case-study-on-darfur/>.
- Markoff, John. “As Artificial Intelligence Evolves, So Does Its Criminal Potential.” *The New York Times*, October 23, 2016, sec. Technology.
<https://www.nytimes.com/2016/10/24/technology/artificial-intelligence-evolves-with-its-criminal-potential.html>.
- Markou, Christopher, and Simon Deakin. “Is Law Computable? From Rule of Law to Legal Singularity.” In *Is Law Computable? Critical Perspectives on Law + Artificial Intelligence*, edited by Christopher Markou and Simon Deakin. Hart, 2020.
<https://papers.ssrn.com/abstract=3589184>.
- Marshall, Alex. “From Jingles to Pop Hits, A.I. Is Music to Some Ears.” *The New York Times*, January 22, 2017.
<https://www.nytimes.com/2017/01/22/arts/music/juke-deck-artificial-intelligence-songwriting.html>.
- Martens, Jens. *Corporate Influence on the G20: The Case of the B20 and Transnational Business Networks*. First edition. Bonn New York, NY: Global Policy Forum and Heinrich-Böll-Stiftung, 2017.
https://www.boell.de/sites/default/files/corporate_influence_on_the_g20.pdf.
- Martin, Christopher Flynn, Rahul Bhui, Peter Bossaerts, Tetsuro Matsuzawa, and Colin Camerer. “Chimpanzee Choice Rates in Competitive Games Match Equilibrium Game Theory Predictions.” *Scientific Reports* 4, no. 1 (June 5, 2014): 5182.
<https://doi.org/10.1038/srep05182>.
- Martin, Will, and Patrick Messerlin. “Why Is It so Difficult? Trade Liberalization under the Doha Agenda.” *Oxford Review of Economic Policy* 23, no. 3 (October 1, 2007): 347–66.
<https://doi.org/10.1093/oxrep/grm026>.
- Marx, Leo. “Technology’: The Emergence of a Hazardous Concept.” *Social Research; Camden, N. J.* 64, no. 3 (January 1, 1997): 965–988.
- Mascarenhas, Hyacinth. “Associated Press to Expand Its Sports Coverage by Using AI to Write Minor League Baseball Articles.” International Business Times UK, July 5, 2016.
<http://www.ibtimes.co.uk/associated-press-expand-its-sports-coverage-by-using-ai-write-minor-league-baseball-articles-1568804>.
- Matz-Lück, Nele. “Framework Conventions as a Regulatory Tool.” *Goettingen Journal of International*

- Law* 3 (2009): 439–58. <https://doi.org/10.3249/1868-1581-1-3-MATZ-LUECK>.
- McAllister, Amanda. “Stranger than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture.” *MINNESOTA LAW REVIEW*, 2018, 47.
- McBain, Graham. “Abolishing Some Obsolete Common Law Crimes.” *King’s Law Journal* 20, no. 1 (February 2009): 89–114. <https://doi.org/10.1080/09615768.2009.11427722>.
- McCarthy, J., M.L. Minsky, N. Rochester, and C.E. Shannon. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,” August 31, 1955. <http://robotics.cs.tamu.edu/dshell/cs625/2014-09-02.pdf>.
- McCarthy, John. “From Here to Human-Level AI.” In *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning*, 640–646. KR’96. Cambridge, Massachusetts, USA: Morgan Kaufmann Publishers Inc., 1996.
- . “Programs with Common Sense,” 15, 1959. <http://www-formal.stanford.edu/jmc/mcc59.html>.
- McConnell, Steve. *Code Complete: A Practical Handbook of Software Construction, Second Edition*. 2nd edition. Redmond, Wash: Microsoft Press, 2004.
- McCorduck, Pamela. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 25th anniversary update. Natick, Mass: A.K. Peters, 2004.
- McElheran, Kristina, Nathan Goldschlag, Zachary Kroff, David Beede, Lucia Foster, Erik Brynjolfsson, Catherine Buffington, and Emin Dinlersoz. “Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey,” July 2020, 72.
- McGinnis, John O. “Accelerating AI.” *Northwestern University Law Review* 104 (2010). https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1193&context=nulr_online.
- McGrath, Rita Gunther. “The Pace of Technology Adoption Is Speeding Up.” *Harvard Business Review*, November 25, 2013. <https://hbr.org/2013/11/the-pace-of-technology-adoption-is-speeding-up>.
- McGregor, Lorna. “Are New Technologies an Aid to Reputation as a Disciplinarian?” *AJIL Unbound* 113 (ed 2019): 238–41. <https://doi.org/10.1017/aju.2019.54>.
- McGregor, Lorna, Daragh Murray, and Vivian Ng. “International Human Rights Law as a Framework for Algorithmic Accountability.” *International & Comparative Law Quarterly* 68, no. 2 (April 2019): 309–43. <https://doi.org/10.1017/S0020589319000046>.
- McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafiyan, Trevor Back, et al. “International Evaluation of an AI System for Breast Cancer Screening.” *Nature* 577, no. 7788 (January 2020): 89–94. <https://doi.org/10.1038/s41586-019-1799-6>.
- McKinsey & Company. “Global AI Survey: AI Proves Its Worth, but Few Scale Impact,” November 2019. <https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact>.
- McLaughlin, Robert. “Unmanned Naval Vehicles at Sea: USVs, UUVs, and the Adequacy of the Law.” *Journal of Law, Information and Science*, 2011. <https://doi.org/10.5778/JLIS.2011.21.McLaughlin.1>.
- McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill. “Does ACM’s Code of Ethics Change Ethical Decision Making in Software Development?” In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018*, 729–33. Lake Buena Vista, FL, USA: ACM Press, 2018. <https://doi.org/10.1145/3236024.3264833>.
- McNamara, Robert. “The Diffusion of Nuclear Weapons with and without a Test Ban Agreement,” February 12, 1963.
- McPherson, Ella, Isabel Guenette Thornton, and Matt Mahmoudi. “Open Source Investigations and the Technology-Driven Knowledge Controversy in Human Rights Fact-Finding.” In *Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation and Accountability*, edited by A Koenig, S Duberley, and D Murray, 21. Oxford: Oxford University Press, 2019.
- Medvedeva, Masha, Michel Vols, and Martijn Wieling. “Using Machine Learning to Predict Decisions of the European Court of Human Rights.” *Artificial Intelligence and Law* 28, no. 2 (June 1, 2020): 237–66. <https://doi.org/10.1007/s10506-019-09255-y>.
- Meer, Sico van der. “Forgoing the Nuclear Option: States That Could Build Nuclear Weapons but Chose Not to Do So.” *Medicine, Conflict and Survival* 30, no. S1 (2014): s27–34. <https://doi.org/10.1080/13623699.2014.930238>.
- Megiddo, Tamar. “Knowledge Production, Big Data and Data-Driven Customary International Law.” Rochester, NY: Social Science Research Network, 2019. <https://papers.ssrn.com/abstract=3497477>.
- Metz, Luke, Niru Maheswaranathan, C. Daniel Freeman, Ben Poole, and Jascha Sohl-Dickstein. “Tasks, Stability, Architecture, and Compute: Training More Effective Learned Optimizers, and Using Them to Train Themselves.” *ArXiv:2009.11243 [Cs, Stat]*, September 23, 2020. <http://arxiv.org/abs/2009.11243>.
- Michel, Arthur Holland. *Eyes in the Sky: The Secret Rise of Gorgon Stare and How It Will Watch Us All*. HMH Books, 2019. <https://www.hmhbooks.com/shop/books/Eyes-in-the-Sky/9780544972001>.
- . “The Killer Algorithms Nobody’s Talking About.” *Foreign Policy* (blog), January 20, 2020. <https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/>.
- Ministère de l’Europe et des Affaires étrangères. “Launch of the Global Partnership on Artificial Intelligence by 15 Founding Members.” France Diplomacy - Ministry for Europe and Foreign Affairs, June 15, 2020. <https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding>.
- Minsky, Marvin. Consciousness is a Big Suitcase. Interview by John Brockman, February 26, 1998. https://www.edge.org/conversation/marvin_minsky-consciousness-is-a-big-suitcase.

- Miranda, Enrique Martinez, Peter McBurney, and Matthew J. W. Howard. "Learning Unfair Trading: A Market Manipulation Analysis from the Reinforcement Learning Perspective." In *Proceedings of the 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2016*, 103–9. Institute of Electrical and Electronics Engineers Inc., 2016. <https://doi.org/10.1109/EAIS.2016.7502499>.
- Mirhoseini, Azalia, Anna Goldie, Mustafa Yazgan, Joe Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, et al. "Chip Placement with Deep Reinforcement Learning." *ArXiv:2004.10746 [Cs]*, April 22, 2020. <http://arxiv.org/abs/2004.10746>.
- Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1, no. 11 (November 2019): 501–7. <https://doi.org/10.1038/s42256-019-0114-4>.
- MMC Ventures. "The State of AI 2019: Divergence." MMC Ventures, 2019. <https://www.mmcventures.com/wp-content/uploads/2019/02/The-State-of-AI-2019-Divergence.pdf>.
- Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. "Playing Atari with Deep Reinforcement Learning." *ArXiv:1312.5602 [Cs]*, December 19, 2013, 9.
- Mola, Roger. "Super Helmet." Air & Space Magazine, September 2017. <https://www.airspacemag.com/military-aviation/super-helmet-180964342/>.
- Molnar, Petra. "Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective." *Cambridge International Law Journal* 8, no. 2 (December 1, 2019): 305–30. <https://doi.org/10.4337/cilj.2019.02.07>.
- Molnar, Petra, and Lex Gill. "Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System." Toronto: International Human Rights Program (Faculty of Law, University of Toronto), and the Citizen Lab (Munk School of Global Affairs and Public Policy, University of Toronto), 2018. <https://citizenlab.ca/wp-content/uploads/2018/09/IHRP-Automated-Systems-Report-Web-V2.pdf>.
- Monett, Dagmar, Colin W. P. Lewis, Kristinn R. Thórisson, Joscha Bach, Gianluca Baldassarre, Giovanni Granato, Istvan S. N. Berkeley, et al. "Special Issue 'On Defining Artificial Intelligence'—Commentaries and Author's Response." *Journal of Artificial General Intelligence* 11, no. 2 (February 1, 2020): 1–100. <https://doi.org/10.2478/jagi-2020-0003>.
- Moore, Gordon E. "Cramming More Components onto Integrated Circuits." *Electronics* 38, no. 8 (1965): 82–85.
- . "Progress in Digital Integrated Electronics." *Technical Digest*, 1975.
- Moore, Jared. "AI for Not Bad." *Frontiers in Big Data* 2 (September 11, 2019): 32. <https://doi.org/10.3389/fdata.2019.00032>.
- Morgan, Forrest E., Benjamin Boudreaux, Andrew J Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World." RAND Corporation, 2020. https://www.rand.org/content/dam/rand/pubs/research_reports/RR3100/RR3139-1/RAND_RR3139-1.pdf.
- Morgan, M. Granger. "Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy." *Proceedings of the National Academy of Sciences* 111, no. 20 (May 20, 2014): 7176–84. <https://doi.org/10.1073/pnas.1319946111>.
- Morin, Jean-Frédéric, Hugo Dobson, Claire Peacock, Miriam Prys-Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietzsch, et al. "How Informality Can Address Emerging Issues: Making the Most of the G7." *Global Policy* 10, no. 2 (May 2019): 267–73. <https://doi.org/10.1111/1758-5899.12668>.
- Morin, Jean-Frédéric, and Amandine Orsini. "Regime Complexity and Policy Coherency: Introducing a Co-Adjustments Model." *Global Governance* 19, no. 1 (2013): 41–51.
- Morris, Ian, Richard Seaford, Jonathan D. Spence, Christine M. Korsgaard, and Margaret Atwood. *Foragers, Farmers, and Fossil Fuels: How Human Values Evolve*. Edited by Stephen Macedo. Updated ed. edition. Princeton: Princeton University Press, 2015.
- Morse, Julia C., and Robert O. Keohane. "Contested Multilateralism." *The Review of International Organizations* 9, no. 4 (December 1, 2014): 385–412. <https://doi.org/10.1007/s11558-014-9188-2>.
- Mozur, Paul. "Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras." *The New York Times*, October 15, 2018, sec. Business. <https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html>.
- Mozur, Paul, Jonah M. Kessel, and Melissa Chan. "Made in China, Exported to the World: The Surveillance State." *The New York Times*, April 24, 2019, sec. Technology. <https://www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-police-government.html>.
- Mozur, Paul, and Cade Metz. "A U.S. Secret Weapon in A.I.: Chinese Talent." *The New York Times*, June 9, 2020, sec. Technology. <https://www.nytimes.com/2020/06/09/technology/china-a-ai-research-education.html>.
- Muller, Catelijne. "The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law." Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), 2020. <https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da>.
- Müller, Harald, and Carmen Wunderlich. *Norm Dynamics in Multilateral Arms Control: Interests, Conflicts, and Justice*. University of Georgia Press, 2013.
- Müller, Vincent C. "Ethics of AI and Robotics." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Palo Alto: CSLI, Stanford University, 2020. <https://plato.stanford.edu/entries/ethics-ai/#Sing>.
- Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, edited by Müller, Vincent C. Berlin: Synthese Library, 2016. <http://www.nickbostrom.com/papers/survey.pdf>.
- Murdick, Dewey, James Dunham, and Jennifer Melot. "AI Definitions Affect Policymaking." CSET Issue Brief. Center for Security and Emerging Technology, June 2020. <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Definitions-Affect-Policymaking.pdf>.

- Musgrave, Kevin, Serge Belongie, and Ser-Nam Lim. "A Metric Learning Reality Check." *ArXiv:2003.08505 [Cs]*, March 18, 2020. <http://arxiv.org/abs/2003.08505>.
- Musthaler, Linda. "How to Use Deep Learning AI to Detect and Prevent Malware and APTs in Real-Time." Network World, March 11, 2016. <http://www.networkworld.com/article/3043202/security/how-to-use-deep-learning-ai-to-detect-and-prevent-malware-and-apts-in-real-time.html>.
- Narayanan, Arvind. "How to Recognize AI Snake Oil." 2019. <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>.
- National Human Genome Research Institute. "Human Genome Project FAQ." Genome.gov, February 24, 2020. <https://www.genome.gov/human-genome-project/Completion-FAQ>.
- Nature Editors. "International AI Ethics Panel Must Be Independent." *Nature* 572 (August 21, 2019): 415–415. <https://doi.org/10.1038/d41586-019-02491-x>.
- Nelson, Amy J. "Innovation Acceleration, Digitization, and the Arms Control Imperative." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 26, 2019. <https://papers.ssrn.com/abstract=3382956>.
- . "The Death of the INF Treaty Has Lessons for Arms Control." *Bulletin of the Atomic Scientists* (blog), November 4, 2019. <https://thebulletin.org/2019/11/the-death-of-the-inf-treaty-has-lessons-for-arms-control/>.
- Nelson, Amy J. "The Impact of Emerging Technologies on Arms Control Regimes." Presented at the ISODARCO, 2018. <http://www.isodarco.it/courses/andalo18/paper/iso18-AmyNelson.pdf>.
- Nelson, Jacob L., and Harsh Taneja. "The Small, Disloyal Fake News Audience: The Role of Audience Availability in Fake News Consumption." *New Media & Society* 20, no. 10 (October 1, 2018): 3720–37. <https://doi.org/10.1177/1461444818758715>.
- Nemitz, Paul. "Constitutional Democracy and Technology in the Age of Artificial Intelligence." *Phil. Trans. R. Soc. A* 376, no. 2133 (November 28, 2018): 20180089. <https://doi.org/10.1098/rsta.2018.0089>.
- Nesta. "Documenting Mass Human Rights Violations through Collective Intelligence." *nesta*, 2020. <https://www.nesta.org.uk/feature/collective-intelligence-grants/documenting-mass-human-rights-violations-through-collective-intelligence/>.
- Newman, Jessica Cussins. "Decision Points in AI Governance." Berkeley, CA: Center for Long-Term Cybersecurity, May 5, 2020. https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf.
- Newman, LilHay. "AI Can Recognize Your Face Even If You're Pixelated." WIRED, 2016. <https://www.wired.com/2016/09/machine-learning-can-identify-pixelated-faces-researchers-show/>.
- Newman, Lily Hay. "How Leaked NSA Spy Tool 'EternalBlue' Became a Hacker Favorite." *Wired*, July 3, 2018. <https://www.wired.com/story/eternalblue-leaked-nsa-spy-tool-hacked-world/>.
- Ngo, Richard. "Disentangling Arguments for the Importance of AI Safety." *Thinking Complete* (blog), January 21, 2019. <https://thinkingcomplete.blogspot.com/2019/01/disentangling-arguments-for-importance.html>.
- Nichols, Michelle. "U.S. Withdrawal from WHO over Claims of China Influence to Take Effect July 2021: U.N." *Reuters*, July 8, 2020. <https://www.reuters.com/article/us-health-coronavirus-trump-who-idUSKBN2482YZ>.
- Nieder, Andreas. "Honey Bees Zero in on the Empty Set." *Science* 360, no. 6393 (June 8, 2018): 1069–70. <https://doi.org/10.1126/science.aat8958>.
- Niiler, Eric. "An AI Epidemiologist Sent the First Warnings of the Wuhan Virus." *Wired*, January 25, 2020. <https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/>.
- Nilsson, Nils J. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge ; New York: Cambridge University Press, 2010.
- Nindler, Reinmar. "The United Nation's Capability to Manage Existential Risks with a Focus on Artificial Intelligence." *International Community Law Review* 21, no. 1 (March 11, 2019): 5–34. <https://doi.org/10.1163/18719732-12341388>.
- Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism*. 1 edition. New York: NYU Press, 2018.
- Noorman, Merel. "Computing and Moral Responsibility." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2020. Metaphysics Research Lab, Stanford University, 2018. <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/>.
- Nordhaus, William D. "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth." Working Paper. NATIONAL BUREAU OF ECONOMIC RESEARCH, 2015. <https://www.nber.org/papers/w21547.pdf>.
- Norman, Donald A. *The Design of Everyday Things*. Revised and Expanded edition. New York, New York: Basic Books, 2013.
- Nuclear Energy Agency, and OECD. *Impacts of the Fukushima Daiichi Accident on Nuclear Development Policies*. Nuclear Development. OECD Publishing, 2017. <https://doi.org/10.1787/9789264276192-en>.
- Núñez, Ignacio Cartagena. "Managing Complexity: Three Possible Models for a Future Non-Proliferation and Arms Control Regime." *UNIDIR* (blog), November 7, 2019. <https://www.unidir.org/commentary/managing-complexity-three-possible-models-future-non-proliferation-and-arms-control>.
- Nye, Joseph S. "Nuclear Learning and U.S.-Soviet Security Regimes." *International Organization* 41, no. 3 (1987): 371–402.
- . "The Regime Complex for Managing Global Cyber Activities." Global Commission on Internet Governance, 2014. <https://dash.harvard.edu/bitstream/handle/1/12308565/Nye-GlobalCommission.pdf>.
- . "The World Needs an Arms-Control Treaty for Cybersecurity." *The Washington Post*, October 1, 2015. <https://www.belfercenter.org/publication/world-needs-arms-control-treaty-cybersecurity>.
- Nyhan, Brendan. "Fake News and Bots May Be Worrisome, but Their Political Power Is Overblown." *The New York Times*, February 16, 2018, sec. The Upshot. <https://www.nytimes.com/2018/02/13/upshot/fake-news-political-power.html>.

- news-and-bots-may-be-worrisome-but-their-political-power-is-overblown.html.
- Nyholm, Sven, and Jilles Smids. "The Ethics of Accident Algorithms for Self-Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Moral Practice* 19, no. 5 (November 1, 2016): 1275–89.
<https://doi.org/10.1007/s10677-016-9745-2>.
- Oberthür, Sebastian. "Interplay Management: Enhancing Environmental Policy Integration among International Institutions." *International Environmental Agreements: Politics, Law and Economics* 9, no. 4 (August 13, 2009): 371.
<https://doi.org/10.1007/s10784-009-9109-7>.
- Oberthür, Sebastian, and Olav Schram Stokke, eds. *Managing Institutional Complexity: Regime Interplay and Global Environmental Change*. The MIT Press, 2011.
<https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262015912.001.0001/upso-9780262015912>.
- OECD. "OECD AI Policy Observatory: A Platform for AI Information, Evidence, and Policy Options." OECD, September 2019. <https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policy-observatory.pdf>.
- . "OECD Supporting G20 Policy Priorities at Osaka Summit," June 30, 2019.
https://www.oecd.org/g20/summits/osaka/publications_anddocuments/oecd-supporting-g20-policy-priorities-at-osaka-summit.htm.
- . "OECD to Host Secretariat of New Global Partnership on Artificial Intelligence," June 15, 2020.
<https://www.oecd.org/going-digital/ai/OECD-to-host-Secretariat-of-new-Global-Partnership-on-Artificial-Intelligence.htm>.
- . "Recommendation of the Council on Artificial Intelligence." *OECD Legal Instruments - OECD/LEGAL/0449*, May 22, 2019.
<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL/0449>.
- . "The OECD Artificial Intelligence Policy Observatory." Accessed September 17, 2020.
<https://www.oecd.ai/>.
- Office of Science and Technology Policy. "American Artificial Intelligence Initiative: Year One Annual Report." The White House, 2020.
<https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf>.
- . "The National Artificial Intelligence Research and Development Strategic Plan." National Science and Technology Council, 2016.
https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf.
- Office of the Deputy Under Secretary of Defense for Research and Engineering (Strategic and Space Systems). "ICBM Basing Options: A Summary of Major Studies to Define A Survivable Basing Concept for ICBMs." Department of Defense, December 1980.
<http://www.dtic.mil/cgi/tr/fulltext/u2/a956443.pdf>.
- ÓhÉigearthaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. "Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance." *Philosophy & Technology*, May 15, 2020.
<https://doi.org/10.1007/s13347-020-00402-x>.
- O'Keefe, Cullen, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. "The Windfall Clause: Distributing the Benefits of AI for the Common Good." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 327–31. New York NY USA: ACM, 2020.
<https://doi.org/10.1145/3375627.3375842>.
- Olson, Parmy. "Nearly Half Of All 'AI Startups' Are Cashing In On Hype." *Forbes*, March 4, 2019.
<https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/>.
- Omohundro, Stephen M. "The Basic AI Drives." *Frontiers in Artificial Intelligence and Applications* 171 (January 2008): 483–92.
- Open Roboethics Initiative. "The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll." Open Roboethics Initiative, 2015. http://www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf.
- OpenAI. "OpenAI Five." OpenAI, June 25, 2018.
<https://openai.com/blog/openai-five/>.
- OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, et al. "Learning Dexterous In-Hand Manipulation." *ArXiv:1808.00177 [Cs, Stat]*, August 1, 2018. <http://arxiv.org/abs/1808.00177>.
- Ord, Toby. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette Books, 2020.
- Orseau, Laurent, and Stuart Armstrong. "Safely Interruptible Agents," 10, 2016.
- Orsini, Amandine. "The Negotiation Burden of Institutional Interactions: Non-State Organizations and the International Negotiations on Forests." *Cambridge Review of International Affairs* 29, no. 4 (October 1, 2016): 1421–40.
<https://doi.org/10.1080/09557571.2017.1293610>.
- Orsini, Amandine, Jean-Frédéric Morin, and Oran Young. "Regime Complexes: A Buzz, a Boom, or a Boost for Global Governance?" *Global Governance: A Review of Multilateralism and International Organizations* 19, no. 1 (August 12, 2013): 27–39.
<https://doi.org/10.1163/19426720-01901003>.
- Osborn, M., R. Day, P. Komesaroff, and A. Mant. "Do Ethical Guidelines Make a Difference to Decision-Making?" *Internal Medicine Journal* 39, no. 12 (2009): 800–805. <https://doi.org/10.1111/j.1445-5994.2009.01954.x>.
- Osoba, Osonde A. "Technocultural Pluralism: A 'Clash of Civilizations' in Technology?" In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 132–37. New York NY USA: ACM, 2020.
<https://doi.org/10.1145/3375627.3375834>.
- Osoba, Osonde, and William Welser. *The Risks of Artificial Intelligence to Security and the Future of Work*. RAND Corporation, 2017.
<https://doi.org/10.7249/PE237>.
- Ostrom, Elinor. "Polycentric Systems for Coping with Collective Action and Global Environmental Change." *Global Environmental Change* 20, no. 4 (October 2010): 550–57.
<https://doi.org/10.1016/j.gloenvcha.2010.07.004>.
- Overdorf, Rebekah, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. "POTs: Protective Optimization Technologies." *ArXiv:1806.02711 [Cs]*, June 7, 2018.
<http://arxiv.org/abs/1806.02711>.
- Owen, Richard, Jack Stilgoe, Phil Macnaghten, Mike Gorman, Erik Fisher, and Dave Guston. "A Framework for Responsible Innovation." In

- Responsible Innovation*, 27–50. John Wiley & Sons, Ltd, 2013.
<https://doi.org/10.1002/9781118551424.ch2>.
- Oye, Kenneth A. *Cooperation under Anarchy*. Princeton University Press, 1986.
<https://press.princeton.edu/books/paperback/9780691022406/cooperation-under-anarchy>.
- Paoli, Giacomo Persi, Kerstin Vignard, David Danks, and Paul Meyer. “Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision-Making.” Geneva, Switzerland: UNIDIR, August 30, 2020. <https://unidir.org/publication/modernizing-arms-control>.
- Papernot, Nicolas. “A Marauder’s Map of Security and Privacy in Machine Learning.” *ArXiv:1811.01134 [Cs]*, November 2, 2018.
<http://arxiv.org/abs/1811.01134>.
- Paris, Britt, and Joan Donovan. “DeepFakes and Cheap Fakes: The Manipulation of Audio & Visual Evidence.” *Data & Society*, September 18, 2019.
<https://datasociety.net/output/deepfakes-and-cheap-fakes/>.
- Parker, Jack, and David Danks. “How Technological Advances Can Reveal Rights,” 7, 2019.
http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_129.pdf.
- Parson, Edward, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant, and Nick Novelli. “Could AI Drive Transformative Social Progress? What Would This Require?” *AI Pulse*, September 26, 2019. <https://aipulse.org/could-ai-drive-transformative-social-progress-what-would-this-require/>.
- Parson, Edward, Richard Re, Alicia Solow-Niederman, and Alana Zeide. “Artificial Intelligence in Strategic Context: An Introduction.” *AI Pulse* (blog), February 8, 2019. <https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/>.
- Pasquale, Frank. “The Second Wave of Algorithmic Accountability.” LPE Project, November 25, 2019. <https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/>.
- Pasquale, Frank A. “A Rule of Persons, Not Machines: The Limits of Legal Automation.” *George Washington Law Review* 87, no. 1 (2019): 1–55.
- Pasquale, Frank A., and Glyn Cashwell. “Prediction, Persuasion, and the Jurisprudence of Behaviorism.” *University of Maryland Francis King Carey School of Law Legal Studies Research Paper*, November 8, 2017. <https://papers.ssrn.com/abstract=3067737>.
- Pasquale, Frank, and Glyn Cashwell. “Four Futures of Legal Automation.” *UCLA Law Review Discourse* 26 (2015): 23.
- Patrick, Jeremy. “Not Dead, Just Sleeping: Canada’s Prohibition on Blasphemous Libel as a Case Study in Obsolete Legislation.” *U.B.C. Law Review* 41, no. 2 (2008): 193–248.
- Patrick, Stewart. “The Unruled World: The Case for Good Enough Global Governance.” *Foreign Affairs* 93, no. 1 (2014): 58–73.
- Pauwelyn, J., R. A. Wessel, and J. Wouters. “When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking.” *European Journal of International Law* 25, no. 3 (August 1, 2014): 733–63. <https://doi.org/10.1093/ejil/chu051>.
- Pauwelyn, Joost. “Fragmentation of International Law.” *Max Planck Encyclopedia of Public International Law*, 2006, 13.
- Pavlus, John. “Common Sense Comes to Computers.” *Quanta Magazine*, April 30, 2020.
<https://www.quantamagazine.org/common-sense-comes-to-computers-20200430/>.
- Payne, Kenneth. “Artificial Intelligence: A Revolution in Strategic Affairs?” *Survival* 60, no. 5 (September 3, 2018): 7–32.
<https://doi.org/10.1080/00396338.2018.1518374>.
- Pearson, Gavin, Phil Jolley, and Geraint Evans. “A Systems Approach to Achieving the Benefits of Artificial Intelligence in UK Defence,” 6. Arlington, Virginia, USA, 2018.
<https://arxiv.org/abs/1809.11089>.
- Pekkanen, Saadia M., Mireya Solís, and Saori N. Katada. “Trading Gains for Control: International Trade Forums and Japanese Economic Diplomacy.” *International Studies Quarterly* 51, no. 4 (2007): 945–70.
- Perrault, Raymond, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grossz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. “The AI Index 2019 Annual Report.” Stanford, CA: AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, December 2019.
https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf.
- Perrrow, Charles. “Normal Accidents: Living with High Risk Technologies,” 1984.
<http://press.princeton.edu/titles/6596.html>.
- Peters, Jay. “IBM Will No Longer Offer, Develop, or Research Facial Recognition Technology.” *The Verge*, June 8, 2020.
<https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-recognition-analysis-software>.
- Peterson, Hayley. “Here’s What It Costs to Open a McDonald’s Restaurant.” *Business Insider*, May 6, 2019. <https://www.businessinsider.com/what-it-costs-to-open-a-mcdonalds-2014-11>.
- Petit, Nicolas. “Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 9, 2017.
<https://papers.ssrn.com/abstract=2931339>.
- Pettey, Christy, and Rob van der Meulen. “Gartner Says Global Artificial Intelligence Business Value to Reach \$1.2 Trillion in 2018.” Gartner, April 25, 2018.
<https://www.gartner.com/en/newsroom/press-releases/2018-04-25-gartner-says-global-artificial-intelligence-business-value-to-reach-1-point-2-trillion-in-2018>.
- Picker, Colin B. “A View from 40,000 Feet: International Law and the Invisible Hand of Technology.” *Cardozo Law Review* 23 (2001): 151–219.
- Pidgeon, Nick. “In Retrospect: Normal Accidents.” *Nature* 477, no. 7365 (September 2011): 404–5.
<https://doi.org/10.1038/477404a>.
- Plebe, Alessio, and Pietro Perconti. “The Slowdown Hypothesis.” In *Singularity Hypotheses*, edited by Amnon H. Eden, James H. Moor, Johnny H. Søraker, and Eric Steinhart, 349–65. The Frontiers Collection. Springer Berlin Heidelberg, 2012.
https://doi.org/10.1007/978-3-642-32560-1_17.

- Plesch, Dan, and Thomas G. Weiss. "1945's Lesson: 'Good Enough' Global Governance Ain't Good Enough." *Global Governance* 21, no. 2 (2015): 203.
- Polyakova, Alina. "Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare." *Brookings* (blog), November 15, 2018. <https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/>.
- Prado, Mariana Mota, and Steven J. Hoffman. "The Promises and Perils of International Institutional Bypasses: Defining a New Concept and Its Policy Implications for Global Governance." *Transnational Legal Theory* 10, no. 3–4 (October 2, 2019): 275–94. <https://doi.org/10.1080/20414005.2019.1686866>.
- Pratt, Tyler. "Deference and Hierarchy in International Regime Complexes." *International Organization* 72, no. 3 (ed 2018): 561–90. <https://doi.org/10.1017/S0020818318000164>.
- Preston, Elizabeth. "A 'Self-Aware' Fish Raises Doubts About a Cognitive Test." *Quanta Magazine*, December 12, 2018. <https://www.quantamagazine.org/a-self-aware-fish-raises-doubts-about-a-cognitive-test-20181212/>.
- Price, W. Nicholson. "Black-Box Medicine." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, September 22, 2014. <https://papers.ssrn.com/abstract=2499885>.
- Prosser, Tony. *The Regulatory Enterprise: Government, Regulation, and Legitimacy*. Oxford University Press, 2010. <https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199579839.001.0001/acprof-9780199579839>.
- Protocol Supplementary to the Convention for the Suppression of Unlawful Seizure of Aircraft, 50 ILM 153 § (2011).
- Prunkl, Carina, and Jess Whittlestone. "Beyond Near-and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 138–43. New York NY USA: ACM, 2020. <https://doi.org/10.1145/3375627.3375803>.
- PwC. "The Macroeconomic Impact of Artificial Intelligence." PwC, 2018. <https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf>.
- Radford, Alec, Jeff Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. "Better Language Models and Their Implications." OpenAI Blog, February 14, 2019. <https://blog.openai.com/better-language-models/>.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models Are Unsupervised Multitask Learners," 2019, 24.
- Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." *ArXiv:1906.09208 [Cs]*, December 6, 2019. <https://doi.org/10.1145/3351095.3372828>.
- Raghu, Maithra, and Eric Schmidt. "A Survey of Deep Learning for Scientific Discovery." *ArXiv:2003.11755 [Cs, Stat]*, March 26, 2020. <http://arxiv.org/abs/2003.11755>.
- Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, et al. "Machine Behaviour."
- Nature* 568, no. 7753 (April 2019): 477. <https://doi.org/10.1038/s41586-019-1138-y>.
- Raji, Inioluwa Deborah, and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," 7, 2019.
- Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020. <https://doi.org/10.1145/3351095.3372873>.
- Ram, Natalie. "One Shot Learning In AI Innovation." *AI Pulse* (blog), January 25, 2019. <https://aipulse.org/one-shot-learning-in-ai-innovation/>.
- Ranganathan, Surabhi. *Strategically Created Treaty Conflicts and the Politics of International Law*. Cambridge Studies in International and Comparative Law. Cambridge: Cambridge University Press, 2014. <https://doi.org/10.1017/CBO9781107338005>.
- Rapaport, William J. "What Is Artificial Intelligence?" *Journal of Artificial General Intelligence*, Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response, 11, no. 2 (February 1, 2020): 52–56. <https://doi.org/10.2478/jagi-2020-0003>.
- Raso, Filippo, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. "Artificial Intelligence & Human Rights: Opportunities & Risks." Berkman Klein Center for Internet & Society at Harvard University, September 25, 2018. https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf?
- Ratcliffe, Susan, ed. "Roy Amara 1925–2007: American Futurologist." In *Oxford Essential Quotations*. Oxford University Press, 2016. <http://www.oxfordreference.com/view/10.1093/acref/9780191826719.001.0001/acref-9780191826719>.
- Raustitala, Kal. "Institutional Proliferation and the International Legal Order." In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, edited by Jeffrey L. Dunoff and Mark A. Editors Pollack, 293–320. Cambridge University Press, 2012. <https://doi.org/10.1017/CBO9781139107310.015>.
- Raustitala, Kal, and David G. Victor. "The Regime Complex for Plant Genetic Resources." *International Organization* 58, no. 2 (April 2004): 277–309. <https://doi.org/10.1017/S0020818304582036>.
- Rawls, John. "The Idea of an Overlapping Consensus." *Oxford Journal of Legal Studies* 7, no. 1 (1987): 1–25.
- Rayfuse, Rosemary. "Public International Law and the Regulation of Emerging Technologies." In *The Oxford Handbook of Law, Regulation and Technology*, 2017. <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22>.
- Re, Richard M., and Alicia Solow-Niederman. "Developing Artificially Intelligent Justice." *STANFORD TECHNOLOGY LAW REVIEW* 22, no. 2 (2019): 48.
- Real, Esteban, Chen Liang, David R. So, and Quoc V. Le. "AutoML-Zero: Evolving Machine Learning

- Algorithms From Scratch.” *ArXiv:2003.03384 [Cs, Stat]*, March 6, 2020. <http://arxiv.org/abs/2003.03384>.
- “Regulation of Artificial Intelligence in Selected Jurisdictions.” The Law Library of Congress, Global Legal Research Directorate, January 2019. <https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf>.
- Requarth, Tim. “The Big Problem With ‘Big Science’ Ventures—Like the Human Brain Project.” *Nautlius*, April 22, 2015. <http://nautil.us/blog/the-big-problem-with-big-science-ventureslike-the-human-brain-project>.
- Rességuier, Anaïs, and Rowena Rodrigues. “AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics.” *Big Data & Society* 7, no. 2 (July 1, 2020): 2053951720942541. <https://doi.org/10.1177/2053951720942541>.
- Reynolds, Matt. “Even a Mask Won’t Hide You from the Latest Face Recognition Tech.” *New Scientist*, September 7, 2017. <https://www.newscientist.com/article/2146703-even-a-mask-wont-hide-you-from-the-latest-face-recognition-tech/>.
- Rhodes, Richard. *The Making of the Atomic Bomb*. New York, NY: Simon & Schuster, 1986.
- Richards, Neil M., and William D. Smart. “How Should the Law Think about Robots?” In *Robot Law*, edited by Ryan Calo, A. Froomkin, and Ian Kerr. Edward Elgar Publishing, 2016. <https://doi.org/10.4337/9781783476732>.
- Richardson, Rashida, Jason Schultz, and Kate Crawford. “Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice.” *94 NYU Law Review* 192 (February 13, 2019). <https://papers.ssrn.com/abstract=3333423>.
- Rid, Thomas. *Cyber War Will Not Take Place*. 1 edition. Oxford ; New York: Oxford University Press, 2013.
- Rini, Regina. “Deepfakes and the Epistemic Backstop,” June 21, 2019. <https://philpapers.org/archive/RINDAT.pdf>.
- Risse, Mathias. “Human Rights and Artificial Intelligence: An Urgently Needed Agenda.” *Human Rights Quarterly* 41, no. 1 (February 7, 2019): 1–16. <https://doi.org/10.1353/hrq.2019.0000>.
- Roberts, A., and R. Guelff. *Documents on the Laws of War*. 3rd ed. Oxford University Press, 2000.
- Roberts, Siobhan. “Who’s a Bot? Who’s Not?” *The New York Times*, June 16, 2020, sec. Science. <https://www.nytimes.com/2020/06/16/science/social-media-bots-kazemi.html>.
- Robinson, Mark. “The CERN Community: A Mechanism for Effective Global Collaboration?” *Global Policy*, November 18, 2018. <https://doi.org/10.1111/1758-5899.12608>.
- Rock, Allan. “The Human Security Network, Fifteen Years On.” *Centre for International Policy Studies* (blog), May 21, 2013. <https://www.cips-cepi.ca/2013/05/21/the-human-security-network-fifteen-years-on/>.
- Roff, Heather M. “Advancing Human Security Through Artificial Intelligence.” In *Artificial Intelligence and International Affairs: Disruption Anticipated*, by Mary L. Cummings, Heather M. Roff, Kenneth Cukier, Jacob Parakilas, and Hannah Bryce. Chatham House, 2018. <https://www.chathamhouse.org/sites/default/files/publ>
- ications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.
- . “The Frame Problem: The AI ‘Arms Race’ Isn’t One.” *Bulletin of the Atomic Scientists* 0, no. 0 (April 26, 2019): 1–4. <https://doi.org/10.1080/00963402.2019.1604836>.
- . “What Do People Around the World Think About Killer Robots?” *Slate*, February 8, 2017. http://www.slate.com/articles/technology/future_tense/2017/02/what_do_people_around_the_world_think_a_bout_killer_robots.html.
- Rogers, James. “The Dark Side of Our Drone Future.” *Bulletin of the Atomic Scientists* (blog), October 4, 2019. <https://thebulletin.org/2019/10/the-dark-side-of-our-drone-future/>.
- Roland, Alex, and Philip Shiman. *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*. History of Computing. Cambridge, Mass: MIT Press, 2002.
- Rolnick, David, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, et al. “Tackling Climate Change with Machine Learning.” *ArXiv:1906.05433 [Cs, Stat]*, June 10, 2019. <http://arxiv.org/abs/1906.05433>.
- Romano, Cesare P.R. “Proliferation of International Judicial Bodies: The Pieces of the Puzzle.” *New York University Journal of International Law and Politics* 31 (1999 1998): 709.
- Romera, Beatriz Martinez. *Regime Interaction and Climate Change : The Case of International Aviation and Maritime Transport*. Routledge, 2017. <https://doi.org/10.4324/9781315451817>.
- Rompaey, Léonard van. “Discretionary Robots: Conceptual Challenges in the Legal Regulation of Machine Behaviour.” University of Copenhagen, 2020.
- Rompaey, Léonard Van. “Shifting from Autonomous Weapons to Military Networks.” *Journal of International Humanitarian Legal Studies* 10, no. 1 (June 9, 2019): 111–28. <https://doi.org/10.1163/18781527-01001011>.
- Roque, Ashley. “Project Convergence 2020: US Army Hosting Kill Chain Demonstration.” Janes.com, September 16, 2020. <https://www.janes.com/defence-news/news-detail/project-convergence-2020-us-army-hosting-kill-chain-demonstration>.
- Rosert, Elvira. “Norm Emergence as Agenda Diffusion: Failure and Success in the Regulation of Cluster Munitions.” *European Journal of International Relations* 25, no. 4 (December 1, 2019): 1103–31. <https://doi.org/10.1177/1354066119842644>.
- Rosert, Elvira, and Frank Sauer. “How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies.” *Contemporary Security Policy* 0, no. 0 (May 30, 2020): 1–26. <https://doi.org/10.1080/13523260.2020.1771508>.
- . “Prohibiting Autonomous Weapons: Put Human Dignity First.” *Global Policy* 10, no. 3 (2019): 370–75. <https://doi.org/10.1111/1758-5899.12691>.
- Ruggie, John. “International Regimes, Transactions and Change: Embedded Liberalism in the Postwar Economic Order.” In *International Regimes*, edited by Stephen Krasner, 195–232. Ithaca: Cornell University Press, 1983.
- Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

- <https://www.amazon.com/Human-Compatible-Artificial-Intelligence-Problem-ebook/dp/B07N5J5FTS>.
- Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River: Pearson, 2016.
- Sagan, Scott D. "Why Do States Build Nuclear Weapons?: Three Models in Search of a Bomb." *International Security* 21, no. 3 (1996): 54–86.
- Salles, Arleen, Kathinka Evers, and Michele Farisco. "Anthropomorphism in AI." *AJOB Neuroscience* 11, no. 2 (April 2, 2020): 88–95. <https://doi.org/10.1080/21507740.2020.1740350>.
- Sandler, Todd. "Strategic Aspects of Difficult Global Challenges." *Global Policy* 7, no. S1 (2016): 33–44. <https://doi.org/10.1111/1758-5899.12266>.
- Sayler, Kelley M., and Daniel S. Hoadley. "Artificial Intelligence and National Security." Congressional Research Service, August 26, 2020.
- Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. 1 edition. New York: W. W. Norton & Company, 2018.
- . "Autonomous Weapons and Operational Risk." Ethical Autonomy Project. 20YY Future of Warfare Initiative. Center for a New American Security, 2016. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf.
- . "Autonomous Weapons and Stability." King's College, 2020.
- . "Flash War - Autonomous Weapons and Strategic Stability." Presented at the Understanding Different Types of Risk, Geneva, April 11, 2016. <http://www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf>.
- . "How Swarming Will Change Warfare." *Bulletin of the Atomic Scientists* 74, no. 6 (November 2, 2018): 385–89. <https://doi.org/10.1080/00963402.2018.1533209>.
- . "Killer Apps: The Real Danger of an AI Arms Race." *Foreign Affairs*, April 16, 2019. <https://www.foreignaffairs.com/articles/2019-04-16/killer-apps>.
- Scharre, Paul, and Michael C Horowitz. "Artificial Intelligence: What Every Policymaker Needs to Know." Center for a New American Security, 2018. <https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know>.
- Schatzberg, Eric. "'Technik' Comes to America: Changing Meanings of 'Technology' before 1930." *Technology and Culture* 47, no. 3 (2006): 486–512.
- Schelling, Thomas C. *Arms and Influence*. Yale University Press, 1966.
- Scherer, Matthew U. "Is AI Personhood Already Possible under U.S. LLC Laws? (Part One: New York)." *Law and AI* (blog), May 14, 2017. <http://www.lawandai.com/2017/05/14/is-ai-personhood-already-possible-under-current-u-s-laws-dont-count-on-it-part-one/>.
- . "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology*, no. 2 (Spring 2016). <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf>.
- Schick, Timo, and Hinrich Schütze. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners." *ArXiv:2009.07118 [Cs]*, September 15, 2020. <http://arxiv.org/abs/2009.07118>.
- Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." New York, NY, USA: ACM, 2020. <https://econpapers.repec.org/paper/osfsocarx/8jaz4.htm>.
- Schlag, Pierre. "No Vehicles in the Park." *Seattle University Law Review* 23 (1999): 381–89.
- . "Spam Jurisprudence, Air Law, and the Rank Anxiety of Nothing Happening (A Report on the State of the Art)." *The Georgetown Law Journal*, no. 97 (January 1, 2009): 803.
- Schneider, Jacquelyn. "Digitally-Enabled Warfare: The Capability-Vulnerability Paradox." Center for a New American Security, August 2016. <https://www.cnas.org/publications/reports/digitally-enabled-warfare-the-capability-vulnerability-paradox>.
- . "The Capability/Vulnerability Paradox and Military Revolutions: Implications for Computing, Cyber, and the Onset of War." *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 841–63. <https://doi.org/10.1080/01402390.2019.1627209>.
- Schneier, Bruce. "Bots Are Destroying Political Discourse As We Know It." *The Atlantic*, January 7, 2020. <https://www.theatlantic.com/technology/archive/2020/01/future-politics-bots-drowning-out-humans/604489/>.
- . "Who Are the Shadow Brokers?" *The Atlantic*, May 23, 2017. <https://www.theatlantic.com/technology/archive/2017/05/shadow-brokers/527778/>.
- Scholl, Keller, and Robin Hanson. "Testing the Automation Revolution Hypothesis." *Economics Letters* 193 (August 1, 2020): 109287. <https://doi.org/10.1016/j.econlet.2020.109287>.
- Scholl, Zackary. "Turing Test: Passed, Using Computer-Generated Poetry." *Raspberry PI AI* (blog), January 24, 2015. <https://rpiai.wordpress.com/2015/01/24/turing-test-passed-using-computer-generated-poetry/>.
- Schön, Donald A. *Technology and Change: The New Heraclitus*. Delacorte Press, 1967.
- Schuett, Jonas. "A Legal Definition of AI." *ArXiv:1909.01095 [Cs]*, August 26, 2019. <http://arxiv.org/abs/1909.01095>.
- Schwab, Klaus. *The Fourth Industrial Revolution*. Currency, 2017.
- Schwaller, Philippe, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction." *ACS Central Science* 5, no. 9 (September 25, 2019): 1572–83. <https://doi.org/10.1021/acscentsci.9b00576>.
- Schwartz, Baron. "Heavier-Than-Air Flight Is Impossible," June 4, 2017. <https://www.xaprb.com/blog/flight-is-impossible/>.
- Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI." *ArXiv:1907.10597 [Cs, Stat]*, July 22, 2019. <http://arxiv.org/abs/1907.10597>.
- Schwarz, Elke. "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems." ICRC Humanitarian Law & Policy Blog, August 29, 2018. <https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/>.

- Scott, James Brown, ed. *The Hague Conventions and Declarations of 1899 and 1907, Accompanied by Tables of Signatures, Ratifications and Adhesions of the Various Powers, and Texts of Reservations*. New York: Oxford University Press - American Branch, 1915.
<http://archive.org/details/hagueconventions00inteuoft>
- Scott, J.C. *Seeing Like A State - How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press, 1998.
- Scott, Karen. "International Law in the Anthropocene: Responding to the Geoengineering Challenge." *Michigan Journal of International Law* 34, no. 2 (January 1, 2013): 309–58.
- Scott, Mark. "In 2019, the 'Techlash' Will Go from Strength to Strength." POLITICO, December 30, 2018. <https://www.politico.eu/article/tech-predictions-2019-facebook-techclash-europe-united-states-data-misinformation-fake-news/>.
- Sculley, D., Jasper Snoek, Alex Wiltschko, and Ali Rahimi. "Winner's Curse? On Pace, Progress, and Empirical Rigor," 2018.
<https://openreview.net/forum?id=rJWF0Fywf>.
- Sears, Nathan Alexander. "Existential Security: Towards a Security Framework for the Survival of Humanity." *Global Policy* 11, no. 2 (2020): 255–66.
<https://doi.org/10.1111/1758-5899.12800>.
- . "International Politics in the Age of Existential Threats." *Journal of Global Security Studies*, June 18, 2020, 1–23. <https://doi.org/10.1093/jogss/ogaa027>.
- Severance, C. "Bruce Schneier: The Security Mindset." *Computer* 49, no. 2 (February 2016): 7–8.
<https://doi.org/10.1109/MC.2016.38>.
- Shaffer, Gregory C., and Mark A Pollack. "Hard vs. Soft Law: Alternatives, Complements, and Antagonists in International Governance." *Minnesota Law Review* 94 (2010): 706–99.
- Shaffer, Gregory, and Mark A. Pollack. "Hard and Soft Law." In *Interdisciplinary Perspectives on International Law and International Relations*, edited by Jeffrey L. Dunoff and Mark A. Pollack, 197–222. Cambridge: Cambridge University Press, 2012.
<https://doi.org/10.1017/CBO9781139107310.011>.
- Shanahan, Murray. "Beyond Humans, What Other Kinds of Minds Might Be out There?" Aeon, 2016.
<https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there>.
- . *The Technological Singularity*. 1 edition. Cambridge, Massachusetts: The MIT Press, 2015.
- Shane, Scott. "Malware Case Is Major Blow for the N.S.A." *The New York Times*, May 16, 2017, sec. U.S. <https://www.nytimes.com/2017/05/16/us/nsa-malware-case-shadow-brokers.html>.
- Sharikov, Pavel. "Artificial Intelligence, Cyberattack, and Nuclear Weapons—A Dangerous Combination." *Bulletin of the Atomic Scientists* 74, no. 6 (November 2, 2018): 368–73.
<https://doi.org/10.1080/00963402.2018.1533185>.
- Sharkey, Noel, Marc Goodman, and Nick Ros. "The Coming Robot Crime Wave." *IEEE Computer Magazine* 43, no. 8 (August 2010): 116–115.
<https://doi.org/10.1109/MC.2010.242>.
- Shead, Sam. "Are We on the Cusp of an 'AI Winter'?" *BBC News*, January 12, 2020, sec. Technology.
<https://www.bbc.com/news/technology-51064369>.
- Shelton, Dinah L, ed. *Commitment and Compliance: The Role of Non-Binding Norms in the International Legal System*. Oxford University Press, 2003.
- Sheppard, Brian. "Warming up to Inscrutability: How Technology Could Challenge Our Concept of Law." *University of Toronto Law Journal* 68, no. supplement 1 (January 2018): 36–62.
<https://doi.org/10.3138/utlj.2017-0053>.
- Shermeyer, Jacob, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. "RarePlanes: Synthetic Data Takes Flight." *ArXiv:2006.02963 [Cs]*, June 4, 2020.
<http://arxiv.org/abs/2006.02963>.
- Shevlane, Toby, and Allan Dafoe. "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. New York: ACM, 2020.
<http://arxiv.org/abs/2001.00463>.
- Shoham, Yoav, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz, and Zoe Bauer. "AI Index 2018 Annual Report." Stanford, CA: AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, December 2018.
<http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf>.
- Siliezar, Juan. "African Grey Parrot Outperforms Children and College Students." *Harvard Gazette* (blog), July 2, 2020.
<https://news.harvard.edu/gazette/story/2020/07/africa-n-grey-parrot-outperforms-children-and-college-students/>.
- Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529, no. 7587 (January 27, 2016): 484–89.
<https://doi.org/10.1038/nature16961>.
- Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play." *Science* 362, no. 6419 (December 7, 2018): 1140–44.
<https://doi.org/10.1126/science.aar6404>.
- . "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." *ArXiv:1712.01815 [Cs]*, December 5, 2017.
<http://arxiv.org/abs/1712.01815>.
- Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. "Mastering the Game of Go without Human Knowledge." *Nature* 550, no. 7676 (October 18, 2017): nature24270.
<https://doi.org/10.1038/nature24270>.
- Sim, Joo Yong, Hyung Wook Noh, Woonhoe Goo, Namkeun Kim, Seung-Hoon Chae, and Chang-Geun Ahn. "Identity Recognition Based on Bioacoustics of Human Body." *IEEE Transactions on Cybernetics*, 2019, 1–12.
<https://doi.org/10.1109/TCYB.2019.2941281>.
- Simonite, Tom. "The World Has a Plan to Rein in AI—but the US Doesn't Like It." *Wired*, January 6, 2020.
<https://www.wired.com/story/world-plan-rein-ai-us-doesnt-like/>.

- Simpson, John. "The Nuclear Non-Proliferation Regime: Back to the Future?" *Disarmament Forum*, no. 1 (2004): 12.
- Singer, P. W. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. New York: Penguin Press, 2009.
- Singh, Amarjot, Devendra Patil, G. Meghana Reddy, and S. N. Omkar. "Disguised Face Identification (DFI) with Facial KeyPoints Using Spatial Fusion Convolutional Network." *ArXiv:1708.09317 [Cs]*, August 30, 2017. <http://arxiv.org/abs/1708.09317>.
- Sirleaf, Matiangai. "The African Justice Cascade and the Malabo Protocol." *International Journal of Transitional Justice* 11, no. 1 (March 2017): 71–91. <https://doi.org/10.1093/ijtj/ijx002>.
- Slaughter, Anne-Marie. "International Law and International Relations Theory: Twenty Years Later." In *Interdisciplinary Perspectives on International Law and International Relations*, edited by Jeffrey L. Dunoff and Mark A. Pollack, 611–25. Cambridge: Cambridge University Press, 2012. <https://doi.org/10.1017/CBO9781139107310.030>.
- . "Remarks, The Big Picture: Beyond Hot Spots & Crises in Our Interconnected World." *Penn State Journal of Law & International Affairs* 1, no. 2 (2012): 286–302.
- Slee, Tom. "The Incompatible Incentives of Private Sector AI." In *The Oxford Handbook of AI Ethics*, edited by M Dubber and F. Pasquale, 34. Oxford University Press, 2019.
- Slijper, Frank, Alice Beck, Daan Kayser, and Maaike Beenes. "Don't Be Evil? A Survey of the Tech Sector's Stance on Autonomous Weapons." Pax, 2019. <https://www.paxforpeace.nl/publications/all-publications/dont-be-evil>.
- Smith. "The Need for a Digital Geneva Convention." *Microsoft on the Issues* (blog), February 14, 2017. <https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/>.
- Smith, Bryant Walker. "New Technologies and Old Treaties." *AJIL Unbound* 114 (ed 2020): 152–57. <https://doi.org/10.1017/ajlu.2020.28>.
- Smithrosser, Elizabeth. "How to Get Good Horses in Medieval China." *Medievalists.Net* (blog), June 9, 2020. <https://www.medievalists.net/2020/06/horses-medieval-china/>.
- Smuha, Nathalie A. "Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea." *Philosophy & Technology*, May 24, 2020. <https://doi.org/10.1007/s13347-020-00403-w>.
- Snyder, Jack, and Keir A. Lieber. "Defensive Realism and the 'New' History of World War I." *International Security* 33, no. 1 (June 26, 2008): 174–94. <https://doi.org/10.1162/isec.2008.33.1.174>.
- Sohn, Louis. "The Impact of Technological Changes on International Law." *Washington and Lee Law Review* 30, no. 1 (March 1, 1973): 1.
- Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, et al. "Release Strategies and the Social Impacts of Language Models." *ArXiv:1908.09203 [Cs]*, November 12, 2019. <http://arxiv.org/abs/1908.09203>.
- Solaiman, S. M. "Legal Personality of Robots, Corporations, Idols and Chimpanzees: A Quest for Legitimacy." *Artificial Intelligence and Law* 25, no. 2 (June 1, 2017): 155–79. <https://doi.org/10.1007/s10506-016-9192-3>.
- Solingen, Etel. "The Political Economy of Nuclear Restraint." *International Security*, 1994, 126–59. <https://doi.org/10.2307/2539198>.
- Sotala, Kaj, and Roman V. Yampolskiy. "Responses to Catastrophic AGI Risk: A Survey." Technical Report. Berkeley, CA: Machine Intelligence Research Institute, 2013. <https://intelligence.org/files/ResponsesAGIRisk.pdf>.
- Sparrow, Robert. "Predators or Plowshares? Arms Control of Robotic Weapons." *IEEE Technology and Society Magazine* 28, no. 1 (Spring 2009): 25–29. <https://doi.org/10.1109/MTS.2009.931862>.
- St Petersburg Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight, LXVI UKPP (1869) 659 § (1868).
- Stanley, Jay. "The Dawn of Robot Surveillance: AI, Video Analytics, and Privacy." American Civil Liberties Union, 2019. https://www.aclu.org/sites/default/files/field_document/061119-robot_surveillance.pdf.
- Stark, Luke. "Facial Recognition Is the Plutonium of AI." *XRDS* 25, no. 3 (April 2019): 50–55. <https://doi.org/10.1145/3313129>.
- Stark, Luke, and Anna Lauren Hoffmann. "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture." *Journal of Cultural Analytics* 1, no. 1 (May 1, 2019): 11052. <https://doi.org/10.22148/16.036>.
- Stein, Arthur A. "Coordination and Collaboration: Regimes in an Anarchic World." *International Organization* 36, no. 2 (1982): 299–324.
- Stein, Peter, and Peter Feaver. *Assuring Control of Nuclear Weapons: The Evolution of Permissive Action Links*. Cambridge, Mass.: Lanham, MD: Univ Pr of Amer, 1989.
- Stilgoe, Jack, Richard Owen, and Phil Macnaghten. "Developing a Framework for Responsible Innovation." *Research Policy* 42, no. 9 (November 1, 2013): 1568–80. <https://doi.org/10.1016/j.respol.2013.05.008>.
- Stone, John. "Cyber War Will Take Place!" *Journal of Strategic Studies* 36, no. 1 (February 1, 2013): 101–8. <https://doi.org/10/cp6d>.
- Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, et al. "Artificial Intelligence and Life in 2030." One Hundred Yearly on Artificial Intelligence. Stanford, CA: Stanford University, September 2016. <http://ai100.stanford.edu/2016-report>.
- Stoner, Ian. "In Defense of Hyperlinks: A Response to Dreyfus." *Techné: Research in Philosophy and Technology* 7, no. 3 (2004). <https://scholar.lib.vt.edu/ejournals/SPT/v7n3/stoner.hmtl>.
- Strange, Susan. "Cave! Hic Dragones: A Critique of Regime Analysis." *International Organization* 36, no. 2, (1982): 479–96.
- Struett, Michael J., Mark T. Nance, and Diane Armstrong. "Navigating the Maritime Piracy Regime Complex." *Global Governance* 19, no. 1 (2013): 93–104.
- Sunstein, Cass. "Holberg Prize 2018, Acceptance Speech." Text presented at the Holberg Prize 2018, Bergen, Norway, June 7, 2018. <https://www.holbergprisen.uib.no/en/cass-sunsteins-acceptance-speech>.

- Sunstein, Cass R. "Incompletely Theorized Agreements." *Harvard Law Review* 108, no. 7 (1995): 1733–72.
<https://doi.org/10.2307/1341816>.
- Susskind, Jamie. *Future Politics: Living Together in a World Transformed by Tech*. Oxford, United Kingdom ; New York, NY: Oxford University Press, 2018.
- Sutton, Richard. "The Bitter Lesson," March 13, 2019.
<http://www.incompleteideas.net/IncIdeas/BitterLesson.html>.
- Svyatkovskiy, Alexey, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. "IntelliCode Compose: Code Generation Using Transformer." *ArXiv:2005.08025 [Cs]*, May 16, 2020. <http://arxiv.org/abs/2005.08025>.
- Swedberg, Richard. "Does Speculation Belong in Social Science Research?" *Sociological Methods & Research*, April 24, 2018, 0049124118769092.
<https://doi.org/10.1177/0049124118769092>.
- Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing Properties of Neural Networks." *ArXiv:1312.6199 [Cs]*, February 19, 2014. <http://arxiv.org/abs/1312.6199>.
- Taddeo, Mariarosaria, and Luciano Floridi. "How AI Can Be a Force for Good." *Science* 361, no. 6404 (August 24, 2018): 751–52.
<https://doi.org/10.1126/science.aat5991>.
- Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. 2nd ed. Cambridge: Cambridge University Press, 2017.
<https://doi.org/10.1017/9781316822524>.
- Tan, Joshua Z., and Jeffrey Ding. "AI Governance through AI Markets," 2019, 7.
- Tang, Shiping. "The Security Dilemma: A Conceptual Analysis." *Security Studies* 18, no. 3 (September 18, 2009): 587–623.
<https://doi.org/10.1080/09636410903133050>.
- Tasioulas, John. "First Steps Towards an Ethics of Robots and Artificial Intelligence." *Journal of Practical Ethics* 7, no. 1 (June 2019): 61–95.
- Taylor, Astra. "The Automation Charade." *Logic Magazine*, August 1, 2018.
<https://logicmag.io/failure/the-automation-charade/>.
- Taylor, Charles. "Conditions of an Unforced Consensus on Human Rights." Bangkok, 1996.
<https://www.iilj.org/wp-content/uploads/2016/08/Taylor-Conditions-of-an-Unforced-Consensus-on-Human-Rights-1996.pdf>.
- Taylor, Trevor. "Artificial Intelligence in Defence: When AI Meets Defence Acquisition Processes and Behaviours." *The RUSI Journal* 164, no. 5–6 (September 19, 2019): 72–81.
<https://doi.org/10.1080/03071847.2019.1694229>.
- Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.
- Teller, Edward. *The Legacy of Hiroshima*. Doubleday, 1962.
- Tencent Keen Security Lab. "Experimental Security Research of Tesla Autopilot." Tencent, March 2019.
https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf.
- Teo, Sue Anne. "Artificial Intelligence and Corporate Human Rights Self-Regulation Initiatives: The Dangers of Letting Business Go on as Usual." In (*Working Paper*). Maastricht, 2020.
- Teoh, Eric R., and David G. Kidd. "Rage against the Machine? Google's Self-Driving Cars versus Human Drivers." *Journal of Safety Research* 63 (2017): 57–60. <https://doi.org/10.1016/j.jsr.2017.08.008>.
- Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Reprint edition. Broadway Books, 2016.
- The Economist. "Are Data More like Oil or Sunlight?" *The Economist*, February 20, 2020.
<https://www.economist.com/special-report/2020/02/20/are-data-more-like-oil-or-sunlight>.
- . "The Incredible Shrinking Machine - A New Material Helps Transistors Become Vanishingly Small." *The Economist*, July 18, 2020.
<https://www.economist.com/science-and-technology/2020/07/18/a-new-material-helps-transistors-become-vanishingly-small>.
- Thelen, Kathleen. "Historical Institutionalism in Comparative Politics." *Annual Review of Political Science* 2, no. 1 (1999): 369–404.
<https://doi.org/10.1146/annurev.polisci.2.1.369>.
- Third Geneva Convention Relative to the Treatment of Prisoners of War (1949).
- Thomas, Rachel, and David Uminsky. "The Problem with Metrics Is a Fundamental Problem for AI." *ArXiv:2002.08512 [Cs]*, February 19, 2020.
<http://arxiv.org/abs/2002.08512>.
- Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. "The Computational Limits of Deep Learning." *ArXiv:2007.05558 [Cs, Stat]*, July 10, 2020. <http://arxiv.org/abs/2007.05558>.
- Thompson, Neil, and Svenja Spanuth. "The Decline of Computers As a General Purpose Technology: Why Deep Learning and the End of Moore's Law Are Fragmenting Computing." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, November 20, 2018.
<https://papers.ssrn.com/abstract=3287769>.
- Thompson, Nicholas. "Inside the Apocalyptic Soviet Doomsday Machine." *Wired*, September 21, 2009.
<https://www.wired.com/2009/09/mf-deadhand/>.
- Thompson, Nicholas, and Ian Bremmer. "The AI Cold War That Threatens Us All." *Wired*, October 23, 2018.
<https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/>.
- Tikk, Eneken, and Mika Kerttunen. "The Alleged Demise of the UN GGE: An Autopsy and Eulogy." Cyber Policy Institute, 2017. <https://cpi.ee/wp-content/uploads/2017/12/2017-Tikk-Kerttunen-Demise-of-the-UN-GGE-2017-12-17-ET.pdf>.
- Tilghman, Paul. "Adaptive Radar Countermeasures (ARC)." DARPA. Accessed March 12, 2018.
<https://www.darpa.mil/program/adaptive-radar-countermeasures>.
- . "Behavioral Learning for Adaptive Electronic Warfare (BLADE)." DARPA. Accessed March 12, 2018. <https://www.darpa.mil/program/behavioral-learning-for-adaptive-electronic-warfare>.
- Tobey, Danny. "Software Malpractice in the Age of AI: A Guide for the Wary Tech Company." In *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society 2018*, 6. New Orleans, 2018. http://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_43.pdf.
- Torres, Phil. *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Durham, North Carolina: Pitchstone Publishing, 2017.
- Trachtman, Joel P. "The Growing Obsolescence of Customary International Law." In *Custom's Future*:

- International Law in a Changing World*, edited by Curtis A. Bradley, 172–204. Cambridge University Press, 2016.
<https://doi.org/10.1017/CBO9781316014264.008>.
- Trajtenberg, Manuel. “AI as the next GPT: A Political-Economy Perspective.” Working Paper, National Bureau of Economic Research, January 2018.
<https://doi.org/10.3386/w24245>.
- Trask, Andrew. “Safe Crime Prediction: Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance,” June 5, 2017.
<https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/>.
- Treaty on Open Skies, CTS No 3 § (1992).
<https://www.osce.org/files/f/documents/1/5/14127.pdf>.
- Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies, 610 UNTS 205 § (1967).
- Treaty on the Non-Proliferation of Nuclear Weapons, 729 UNTS 161 § (1968).
- Trevithick, Joseph. “Turkey Now Has Swarming Suicide Drones It Could Export.” *The Drive*, June 18, 2020.
<https://www.thedrive.com/the-war-zone/34204/turkey-now-has-a-swarming-quadcopter-suicide-drone-that-it-could-export>.
- Truby, Jon. “Governing Artificial Intelligence to Benefit the UN Sustainable Development Goals.” *Sustainable Development* n/a, no. n/a (2020).
<https://doi.org/10.1002/sd.2048>.
- Trudeau, Justin. “Mandate for the International Panel on Artificial Intelligence.” Prime Minister of Canada, December 6, 2018.
<https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence>.
- Tshitoyan, Vahe, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. “Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature.” *Nature* 571, no. 7763 (July 2019): 95–98.
<https://doi.org/10.1038/s41586-019-1335-8>.
- Tucker, Aaron D., Markus Anderljung, and Allan Dafoe. “Social and Governance Implications of Improved Data Efficiency.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 378–84. New York NY USA: ACM, 2020.
<https://doi.org/10.1145/3375627.3375863>.
- Tucker, Patrick. “SecDef: China Is Exporting Killer Robots to the Mideast.” Defense One, November 5, 2019.
<https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/>.
- Turner, Alexander Matt. “Optimal Farsighted Agents Tend to Seek Power.” *ArXiv:1912.01683 [Cs]*, December 3, 2019. <http://arxiv.org/abs/1912.01683>.
- Turner, Jacob. *Robot Rules: Regulating Artificial Intelligence*. New York, NY: Springer Berlin Heidelberg, 2018.
- UN. Statute of the International Court of Justice, 33 UNTS 993 § (1946).
<https://www.refworld.org/docid/3deb4b9c0.html>.
- UN General Assembly. *International Covenant on Civil and Political Rights*. Vol. 999. Treaty Series. United Nations, 1966.
- . Universal Declaration of Human Rights, 217 A(III) § (1948).
- UN SG HLPDC. “The Age of Digital Interdependence: Report of the UN Secretary-General’s High-Level Panel on Digital Cooperation.” UN Secretary-General’s High-Level Panel on Digital Cooperation, 2019. <https://digitalcooperation.org/wp-content/uploads/2019/06/DigitalCooperation-report-for-web.pdf>.
- Union of International Associations. “Yearbook of International Organizations: Volume 5: Statistics, 2016.” Union of International Associations, 2016.
<http://brill.uia.org/content/4384>.
- United Nations. “Guiding Principles on Business and Human Rights: Implementing the United Nations ‘Protect, Respect and Remedy’ Framework.” United Nations, 2011.
https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.
- . “Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.,” 1977, 126.
- . “United Nations Conference to Negotiate a Legally Binding Instrument to Prohibit Nuclear Weapons, Leading Towards Their Total Elimination,” July 7, 2017.
<https://www.un.org/disarmament/publications/library/ptnw/>.
- . Vienna Convention on the Law of Treaties, 1155 UNTS 331 § (1969).
https://treaties.un.org/pages/ViewDetailsIII.aspx?src=TREATY&mtdsg_no=XXIII-1&chapter=23&Temp=mtdsg3&clang=_en.
- . Vienna Convention on the Law of Treaties Between States and International Organizations or Between International Organizations, 25 ILM 543 § (1989).
- United Nations Convention against Illicit Traffic in Narcotic Drugs and Psychotropic Substances (1988).
- United Nations Convention on Prohibition or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 1342 UNTS 137 § (1980).
- United Nations General Assembly. “Road Map for Digital Cooperation: Implementation of the Recommendations of the High-Level Panel on Digital Cooperation.” United Nations General Assembly, May 29, 2020.
<https://undocs.org/pdf?symbol=en/A/74/821>.
- United Nations Security Council. Resolution 1373 (2001).
https://www.unodec.org/pdf/crime/terrorism/res_1373_english.pdf.
- . Resolution 1540 (2004).
- United States, and USSR. Treaty on the Limitation of Anti-Ballistic Missile Systems, 944 UNTS 13 § (1972).
- UNOG. “Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.” Geneva: United Nations Office at Geneva, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, September 25, 2019.
[https://www.unog.ch/80256EDD006B8954/\(httpAsset](https://www.unog.ch/80256EDD006B8954/(httpAsset)

- s)/5497DF9B01E5D9CFC125845E00308E44/\$file/CC_W_GGE.1_2019_CRP.1_Rev2.pdf.
- Urpelainen, Johannes, and Thijs Van de Graaf. "Your Place or Mine? Institutional Capture and the Creation of Overlapping International Institutions." *British Journal of Political Science* 45, no. 4 (October 2015): 799–827.
<https://doi.org/10.1017/S0007123413000537>.
- Van Evera, Stephen. "The Cult of the Offensive and the Origins of the First World War." *International Security* 9, no. 1 (1984): 58–107.
<https://doi.org/10.2307/2538636>.
- Vardi, Moshe Y. "Artificial Intelligence: Past and Future." *Communications of the ACM* 55, no. 1 (January 2012): 5. <https://doi.org/10.1145/2063176.2063177>.
- Velez-Green, Alexander. "The Foreign Policy Essay: The South Korean Sentry—A 'Killer Robot' to Prevent War." Lawfare, March 1, 2015.
<https://www.lawfareblog.com/foreign-policy-essay-south-korean-sentry%E2%80%94-killer-robot-prevent-war>.
- Verbruggen, Maaike. "AI & Military Procurement: What Computers Still Can't Do." War on the Rocks, May 5, 2020. <https://warontherocks.com/2020/05/ai-military-procurement-what-computers-still-cant-do/>.
- . "In Defense of Technological Determinism." Brussels, Belgium, 2020.
- . "The Question of Swarms Control: Challenges to Ensuring Human Control over Military Swarms." Non-Proliferation and Disarmament Papers. EU Non-Proliferation and Disarmament Consortium, December 2019. https://www.nonproliferation.eu/wp-content/uploads/2019/12/EUNPDC_no-65_031219.pdf.
- . "The Role of Civilian Innovation in the Development of Lethal Autonomous Weapon Systems." *Global Policy* 10, no. 3 (2019): 338–42.
<https://doi.org/10.1111/1758-5899.12663>.
- Vinci, Anthony. "The Coming Revolution in Intelligence Affairs." *Foreign Affairs*, August 31, 2020.
<https://www.foreignaffairs.com/articles/north-america/2020-08-31/coming-revolution-intelligence-affairs>.
- Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Langhans, Max Tegmark, and Francesco Fuso Nerini. "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals." *Nature Communications* 11 (January 13, 2020).
<https://www.nature.com/articles/s41467-019-14108-y>.
- Vogt, RJ. "DoNotPay Founder Opens Up On 'Robot Lawyers.'" Law360, February 9, 2020.
<https://www.law360.com/articles/1241251/donotpay-founder-opens-up-on-robot-lawyers>.
- Vöneky, Silja. "How Should We Regulate AI? Current Rules and Principles as Basis for 'Responsible Artificial Intelligence,'" May 19, 2020.
<https://papers.ssrn.com/abstract=3605440>.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The Spread of True and False News Online." *Science* 359, no. 6380 (March 9, 2018): 1146–51.
<https://doi.org/10.1126/science.aap9559>.
- Wallach, Wendell, and Colin Allen. "Framing Robot Arms Control." *Ethics and Information Technology; Dordrecht* 15, no. 2 (June 2013): 125–35.
<http://dx.doi.org.ep.fjernadgang.kb.dk/10.1007/s10676-012-9303-0>.
- Wallach, Wendell, and Gary Marchant. "Toward the Agile and Comprehensive International Governance of AI and Robotics." *Proceedings of the IEEE* 107, no. 3 (March 2019): 505–8.
<https://doi.org/10.1109/JPROC.2019.2899422>.
- Wallach, Wendell, and Gary E Marchant. "An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics," 2018, 7.
- Walsh, James Igoe, and Marcus Schulzke. "The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War?" U.S. Army War College & Strategic Studies Institute, 2015.
<https://ssi.armywarcollege.edu/pdffiles/PUB1289.pdf>.
- Walz, Axel, and Kay Firth-Butterfield. "Implementing Ethics into Artificial Intelligence: A Contribution, From a Legal Perspective, To The Development of an AI Governance Regime." *Duke Law & Technology Review* 18, no. 1 (December 2019): 166–231.
- Watts, Sean. "Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War." *International Law Studies* 91 (2015): 83.
- Weaver, John Frank. "Autonomous Weapons and International Law: We Need These Three International Treaties to Govern 'Killer Robots.'" *Slate Magazine*, December 1, 2014.
<https://slate.com/technology/2014/12/autonomous-weapons-and-international-law-we-need-these-three-treaties-to-govern-killer-robots.html>.
- Webb, Greg. "Trump Officials Consider Nuclear Testing." Arms Control Association, June 2020.
<https://www.armscontrol.org/act/2020-06/news/trump-officials-consider-nuclear-testing>.
- Weinberg, Justin. "Philosophers On GPT-3 (Updated with Replies by GPT-3)." Daily Nous, July 30, 2020.
<http://dailynous.com/2020/07/30/philosophers-gpt-3/>.
- Weinberger, Sharon. "Hollywood and Hyper-Surveillance: The Incredible Story of Gorgon Stare." *Nature* 570 (June 11, 2019): 162–63.
<https://doi.org/10.1038/d41586-019-01792-5>.
- . *The Imagineers of War: The Untold Story of DARPA, the Pentagon Agency That Changed the World*. Random House LLC, 2017.
- Weiss, Edith Brown. "International Responses to Weather Modification." *International Organization* 29, no. 3 (ed 1975): 805–26.
<https://doi.org/10.1017/S0020818300031775>.
- Weng, Y., and T. Izumo. "Natural Law and Its Implications for AI Governance." *Delphi - Interdisciplinary Review of Emerging Technologies* 2, no. 3 (2019): 122–28.
<https://doi.org/10.21552/delphi/2019/3/5>.
- Wenger, Andreas, Ursula Jasper, and Myriam Dunn Cavelty, eds. *The Politics and Science of Prevision: Governing and Probing the Future*. CSS Studies in Security and International Relations. Abingdon, Oxon ; New York: Routledge, 2020.
- West, Darrell M. "Brookings Survey Finds Divided Views on Artificial Intelligence for Warfare, but Support Rises If Adversaries Are Developing It." *Brookings* (blog), August 29, 2018.
<https://www.brookings.edu/blog/techtank/2018/08/29/brookings-survey-finds-divided-views-on-artificial-intelligence-for-warfare-but-support-rises-if-adversaries-are-developing-it/>.

- Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions." In *Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics and Society 2019*, 7, 2019. http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_188.pdf.
- Whittlestone, Jess, and Aviv Ovadya. "The Tension between Openness and Prudence in AI Research." *ArXiv:1910.01170 [Cs]*, January 13, 2020. <http://arxiv.org/abs/1910.01170>.
- Wiblin, Robert, Keiran Harris, and Allan Dafoe. The academics preparing for the possibility that AI will destabilise global politics. The 80,000 Hours Podcast, March 18, 2018. <https://80000hours.org/podcast/episodes/allan-dafoe-politics-of-ai>.
- Wiener, Jonathan B. "The Tragedy of the Uncommons: On the Politics of Apocalypse." *Global Policy* 7, no. S1 (2016): 67–80. <https://doi.org/10.1111/1758-5899.12319>.
- Wiggers, Kyle. "OpenAI's Massive GPT-3 Model Is Impressive, but Size Isn't Everything." *VentureBeat* (blog), June 1, 2020. <https://venturebeat.com/2020/06/01/ai-machine-learning-openai-gpt-3-size-isnt-everything/>.
- Williams, James. *Stand out of Our Light: Freedom and Resistance in the Attention Economy*. Reprint edition. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2018.
- Williamson, Richard. "Hard Law, Soft Law, and Non-Law in Multilateral Arms Control: Some Compliance Hypotheses." *Chicago Journal of International Law* 4, no. 1 (April 1, 2003). <https://chicagounbound.uchicago.edu/cjil/vol4/iss1/7>.
- Wilson, Grant. "Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law." *Va. Envtl. LJ* 31 (2013): 307.
- Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121–36.
- Wischemeyer, Thomas, and Timo Rademacher, eds. *Regulating Artificial Intelligence*. Springer International Publishing, 2020. <https://doi.org/10.1007/978-3-030-32361-5>.
- Wolpe, Paul Root. "We Have Met AI, and It Is Not Us." *AJOB Neuroscience* 11, no. 2 (April 2, 2020): 75–76. <https://doi.org/10.1080/21507740.2020.1739876>.
- Wong, Kenneth H. "Framework for Guiding Artificial Intelligence Research in Combat Casualty Care." In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, 10954:109540Q. International Society for Optics and Photonics, 2019. <https://doi.org/10.1117/12.2512686>.
- Wong, Yuna, John Yurchak, Robert Button, Aaron Frank, Burgess Laird, Osonde Osoba, Randall Steeb, Benjamin Harris, and Sebastian Bae. *Deterrence in the Age of Thinking Machines*. RAND Corporation, 2020. <https://doi.org/10.7249/RR2797>.
- Woolf, Amy F., Paul K Kerr, and Mary Beth D Nikitin. "Arms Control and Nonproliferation: A Catalog of Treaties and Agreements," 2018, 73.
- Woolgar, Steve, and Geoff Cooper. "Do Artefacts Have Ambivalence? Moses' Bridges, Winner's Bridges and Other Urban Legends in S&TS." *Social Studies of Science* 29, no. 3 (1999): 433–49.
- WTO. Agreement on Technical Barriers to Trade, 1868 U.N.T.S. 120 § (1994). https://www.wto.org/english/docs_e/legal_e/17-tbt_e.htm.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *ArXiv:1609.08144 [Cs]*, September 26, 2016. <http://arxiv.org/abs/1609.08144>.
- Wyatt, Sally. "Technological Determinism Is Dead; Long Live Technological Determinism." In *The Handbook of Science and Technology Studies*, edited by Edward J. Hackett, Olga Amsteramska, Judy Wajcman, Michael Lynch, Anthony Giddens, and Judy Wajcman, 165–80. MIT Press, 2008.
- Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. "Achieving Human Parity in Conversational Speech Recognition." *ArXiv Preprint ArXiv:1610.05256*, 2016. <https://arxiv.org/abs/1610.05256>.
- Yampolskiy, R.V., and K. Sotala. "Responses to Catastrophic AGI Risk: A Survey." *Physica Scripta* 90, no. 1 (2015). <https://doi.org/10.1088/0031-8949/90/1/018001>.
- Yang, Wei, Kuang Lu, Peilin Yang, and Jimmy Lin. "Critically Examining the 'Neural Hype': Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models." In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1129–1132. SIGIR'19. Paris, France: Association for Computing Machinery, 2019. <https://doi.org/10.1145/3331184.3331340>.
- Yeung, Karen. "Hypernudge: Big Data as a Mode of Regulation by Design." *Information, Communication & Society* 20, no. 1 (January 2, 2017): 118–36. <https://doi.org/10.1080/1369118X.2016.1186713>.
- Yeung, Karen, Andrew Howes, and Ganna Pogrebna. "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing." In *The Oxford Handbook of AI Ethics*, edited by M. Dubber and F. Pasquale, 78–106. Oxford University Press, 2020. <https://papers.ssrn.com/abstract=3435011>.
- Yong, Ed. "The Human Brain Project Hasn't Lived Up to Its Promise." The Atlantic, July 22, 2019. <https://www.theatlantic.com/science/archive/2019/07/en-years-human-brain-project-simulation-markram-ed-talk/594493/>.
- Young, Margaret A. "Regime Interaction in Creating, Implementing and Enforcing International Law." In *Regime Interaction in International Law: Facing Fragmentation*, edited by Margaret A. Young, 85–110. Cambridge: Cambridge University Press, 2012. <https://doi.org/10.1017/CBO9780511862403.005>.
- Young, Oran R. "Institutional Linkages in International Society: Polar Perspectives." *Global Governance* 2, no. 1 (1996): 1–23.
- Yu, Hanzhi, and Lan Xue. "Shaping the Evolution of Regime Complex: The Case of Multiactor Punctuated Equilibrium in Governing Human Genetic Data." *Global Governance: A Review of Multilateralism and International Organizations* 25, no. 4 (December 10, 2019): 645–69. <https://doi.org/10.1163/19426720-02504005>.

- Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic, 308–45. New York: Oxford University Press, 2008.
- . "Cognitive Biases Potentially Affecting Judgment of Global Risks." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Cirkovic. New York: Oxford University Press, 2011. <https://intelligence.org/files/CognitiveBiases.pdf>.
- . "There's No Fire Alarm for Artificial General Intelligence." *Machine Intelligence Research Institute* (blog), October 13, 2017. <https://intelligence.org/2017/10/13/fire-alarm/>.
- Yusuf, Moeed. "Predicting Proliferation: The History of the Future of Nuclear Weapon." Brookings Institute, 2009.
- Zelikman, Eric. "OpenAI Shouldn't Release Their Full Language Model." *The Gradient*, March 3, 2019. <https://thegradient.pub/openai-shouldnt-release-their-full-language-model/>.
- Zellers, Rowan. "Grover: A State-of-the-Art Defense against Neural Fake News," 2019. <https://rowanzellers.com/grover/?domain=nytimes.com&date=May+29%2C+2019&authors=&title=Answer+s+to+Where+to+Watch+the+Pope%E2%80%99s+U.S.+Visit&article=>.
- Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. "Defending Against Neural Fake News." *ArXiv:1905.12616 [Cs]*, May 29, 2019. <http://arxiv.org/abs/1905.12616>.
- Zelli, Fariborz. "The Fragmentation of the Global Climate Governance Architecture." *Wiley Interdisciplinary Reviews: Climate Change* 2, no. 2 (2011): 255–70. <https://doi.org/10.1002/wcc.104>.
- Zelli, Fariborz, and Harro van Asselt. "Introduction: The Institutional Fragmentation of Global Environmental Governance: Causes, Consequences, and Responses." *Global Environmental Politics* 13, no. 3 (July 22, 2013): 1–13. https://doi.org/10.1162/GLEP_a_00180.
- Zeng, Yi, Enmeng Lu, and Cunqing Huangfu. "Linking Artificial Intelligence Principles." *ArXiv:1812.04814 [Cs]*, December 12, 2018. <http://arxiv.org/abs/1812.04814>.
- Zhang, Baobao, and Allan Dafoe. "Artificial Intelligence: American Attitudes and Trends." Center for the Governance of AI, Future of Humanity Institute, University of Oxford, January 2019. <https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/>.
- Zhang, Hugh. "Dear OpenAI: Please Open Source Your Language Model." *The Gradient*, February 19, 2019. <https://thegradient.pub/openai-please-open-source-your-language-model/>.
- Zimmerman, David. "Neither Catapults nor Atomic Bombs: Technological Determinism and Military History from a Post-Industrial Revolution Perspective." *Vulcan* 7, no. 1 (December 5, 2019): 45–61. <https://doi.org/10.1163/22134603-00701005>.
- Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1 edition. New York: PublicAffairs, 2019.
- Zürn, Michael. "Contested Global Governance." *Global Policy* 9, no. 1 (February 1, 2018): 138–45. <https://doi.org/10.1111/1758-5899.12521>.
- Zürn, Michael, and Benjamin Faude. "Commentary: On Fragmentation, Differentiation, and Coordination." *Global Environmental Politics* 13, no. 3 (August 2013): 119–30. https://doi.org/10.1162/GLEP_a_00186.
- Zwetsloot, Remco, and Allan Dafoe. "Thinking About Risks From AI: Accidents, Misuse and Structure." Lawfare, February 11, 2019. <https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure>.
- Zwetsloot, Remco, Helen Toner, and Jeffrey Ding. "Beyond the AI Arms Race: America, China, and the Dangers of Zero-Sum Thinking." *Foreign Affairs*, November 16, 2018. <https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race>.

Appendix: Papers [I – IV]

(in original layout)

Paper [I]:

Maas, Matthijs M. "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons." *Contemporary Security Policy* 40, no. 3 (February 6, 2019): 285–311. <https://doi.org/10.1080/13523260.2019.1576464>.

Paper [II]:

Maas, Matthijs M. "Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot." *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019): 129–57. <https://doi.org/10.1163/18781527-01001006>.

Paper [III]:

Maas, Matthijs M. "International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order." *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56.

Paper [IV]:

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History." In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, 228-34. New York, NY, USA: ACM, 2020. <https://doi.org/10.1145/3375627.3375857>.