Remco Zwetsloot
July 2018
v1

Syllabus: Artificial Intelligence and International Security[1]

Audiences and use. This syllabus covers material located at the intersection between artificial intelligence (AI) and international security. The syllabus can be used in structured self-study (or group-study) for those new to this space, or as a resource for instructors designing class-specific syllabi (which would probably have to be significantly shorter).[2] It is designed to be useful to (a) people new to both AI and international relations (IR); (b) people coming from AI who are interested in an IR angle on the problems; (c) people coming from IR who are interested in working on AI. Depending on which of groups (a)-(c) you fall in, it may be feasible to skip or skim certain sections. For sections that you are particularly interested in, do consider diving into the sources cited in the readings—for most topics what I have assigned just skims the surface and is intended only as a starting point.

Focus. The syllabus grew out of an intensive two-week research bootcamp I organized at Yale University in May 2018. The bootcamp was focused on questions about arms control and arms race dynamics, and these topics are thus the main focus below. Relevant international security-related topics that are somewhat absent are international economic competition, domestic industrial policy, domestic political dynamics more broadly, and long-term international governance questions.[3] The readings are also heavily skewed towards a Western perspective. Future versions (or separate syllabi) will hopefully address these gaps—please contact me if you have suggestions. I intend to update the document every few months.

Organization. Sections 1 and 2 lay the empirical and theoretical foundation for tackling the narrower topics and questions addressed in Section 3. To help people orient themselves, each section and subsection includes some contextual notes and some questions that one can keep in mind while going through the readings. Where it is not obvious, the notes also clarify the relationship between the different sections.

[1] Please send comments to remcozwetsloot@gmail.com. For recommendations and feedback, thanks go to Miles Brundage, Allan Dafoe, Jade Leung, and Matthijs Maas. Special thanks to Will Hunt and Mojmir Stehlik, who participated in the bootcamp and who helped compile the readings.

[2] I do assume a basic familiarity with artificial intelligence. More introductory resources can be found here.

[3] All of these topics are at least briefly discussed in Allan Dafoe's (forthcoming) Research Landscape.

1. Artificial Intelligence and International Security 3
A. AI Trends and Strategies 3
i. Forecasting and Mapping AI Development 3
ii. Country Strategies 4

2. Theoretical Background 6
A. International Relations Frameworks 6
B. Relevant Strategic Concepts 6
i. Bargaining 7
ii. Verification and Enforcement 7
iii. Communication: Signaling and Perception 7
iv. Deterrence and Assurance 8
v. The Offense-Defense Balance 8
vi. Norms, Institutions, and Regimes 9

3. Topics 10
A. Arms Control 10
i. Arms Control and Artificial Intelligence 10
ii. The History of Arms Control 11
iii. The (U.S.) Politics of Arms Control 11
iv. The Role of Ideas, Scientists, and Experts 12
v. The Intellectual History of Arms Control Thinking 12
B. Race Dynamics 13
i. The Idea of an Artificial Intelligence Race 13
ii. International Relations: Arms Races 13
iii. International Relations: The Diffusion of Technology, Strategy, and Arms 14
iv. International Relations: Diffusion, Development, and Conflict 14
v. International Relations: Case Studies 15
vi. Economics: Race and Innovation Models 15
vii. Economics: Case Studies 16
C. Technological Analogies 17
i. Dual-Use Technology 18
ii. General Purpose Technology 18
iii. Nuclear 18
iv. Cyber 19
v. Biotechnology 19
D. Government and Technology 20

1. Artificial Intelligence and International Security

The set of readings in this section present an overview of current thinking on how AI could affect short- to medium-term international security dynamics. They will serve as the best starting point for most investigations of security-related questions; especially relevant parts of these sources will also be referred to in other sections below.

- Allen, G. & Chen, T. (2017), “Artificial Intelligence and National Security,” *Harvard Belfer Center* [PDF]
- Horowitz, M. (2018) “Artificial Intelligence, International Competition, and the Balance of Power,” *Texas National Security Review* [link] [PDF]
- Dafoe, A. (forthcoming), “AI Governance Research Landscape,” especially the section “International Security”
- Brundage, M., Avin, S., et al (2018), “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation” [PDF]
- Danzig, R. (2018), “Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority,” *CNAS Report* [PDF]
- Scharre, P. (2018), *Army of None: Autonomous Weapons and the Future of War*, entire book
- Hoadley, D.S. & Lucas, N. J. (2018), “Artificial Intelligence and National Security,” *Congressional Research Service* [PDF]
- Bostrom, N. (2013), *Superintelligence: Paths, Dangers, Strategies*, chs. 5, 11, 14
- Lieber, K. A. & Press, D. G. (2017), “The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence,” *International Security* 41:4 [link] [PDF]

A. AI Trends and Strategies

This subsection serves to underscore the emerging centrality of artificial intelligence to the international security domain, and to provide some basic background on questions that are relevant for international competition and cooperation. **These readings can safely be skipped initially**, although it may be interesting to return to them once one has gotten more of a feel for how these questions affect our thinking on arms control and race dynamics (Section 3).

i. Forecasting and Mapping AI Development

The likelihood and intensity of AI-driven international competition or cooperation will depend on how fast AI technology advances, who the leading actors are, and so forth. A basic introduction to these questions and relevant references can be found in:

- Dafoe, A. (forthcoming), “AI Governance Research Landscape,” section “Technical Landscape”

Another salient factor when thinking about international competition and cooperation is the distribution of AI-related efforts, as these may shape actors’ interests in or resistance to attempts limit or speed up certain kinds of capabilities. Examples of work on this set of questions includes:

- Baum, S. (2017), “A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy” [PDF]
- Boulanin, V. (2016), “Mapping the Innovation Ecosystem Driving the Advance of Autonomy in Weapon Systems,” *SIPRI Report* [PDF]
- Boulanin, V. & Verbruggen, M. (2017), “Mapping the Development of Autonomy in Weapon Systems,” *SIPRI Report* [PDF]

ii. Country Strategies

Several states have now adopted something akin to a national AI strategy, most of which include military components. For a good overview, see here; the below highlights some of the main geopolitically relevant actors.

The following are good introductions to **China**’s activities:[4]

- China State Council (2017), “A Next Generation Artificial Intelligence Development Plan,” translated by *New America* [PDF]
- Congressional Research Service (2018), “Artificial Intelligence and National Security,” pp. 17-21 [PDF]
- Ding, J. (2018), “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* [PDF]
- Kania, E. (2017), “Battlefield Singularity: Artificial Intelligence, Military Revolution, and China’s Future Military Power,” *CNAS Report* [PDF]

Most of the focus of the readings in the syllabubs focus on the **United States**. Some additional insight can be gained from official government reports (although these date from the Obama administration; it is still unclear what the Trump administration’s approach will be, see e.g. here and here):

- White House NSTC (2016), “The National Artificial Intelligence Research and Development Strategic Plan” [PDF]
- White House NSTC (2016), “Preparing for the Future of Artificial Intelligence” [PDF]

Russia appears to be investing in some AI-related areas (e.g. robotics, cyber security), though not on a scale comparable to the US or China:

[4] Those interested in China might want to subscribe to Jeff Ding’s ChinAI newsletter.

- Congressional Research Service (2018), “Artificial Intelligence and National Security,” pp. 21-22 [PDF]

There have also been several some efforts to lay out strategies in **Europe**, including in France and at the EU level (for more countries’ strategies, see the page linked above):

- Villani (2018), “For a Meaningful Artificial Intelligence: Towards a French and European Strategy” [PDF]
- European Commission High-Level Group on Artificial Intelligence [link]

2. Theoretical Background

A. International Relations Frameworks

There are many different perspectives on international security out there, and it is not feasible to dive into all of them here.[5] For two useful overarching frameworks (which also helpfully discuss the history of security-related debates), see:

- Fearon, J. D. (2018), “Cooperation, Conflict, and the Costs of Anarchy,” *International Organization* [link] [PDF]
- Glaser, C. (2010), *Rational Theory of International Politics: The Logic of Competition and Cooperation* [link to chapters] [PDF]
    - Fearon (2011), “Two States, Two Types, Two Actions,” *Security Studies* 20:3 [link] [PDF], includes a good (brief) summary and discussion of where his framework differs from and overlaps with Glaser’s security dilemma-focused framework.

Both of these perspectives, and many of the readings recommended below, draw—both formally and informally—on game-theoretic ideas. For good introductions to these kinds of ideas, see:

- Kydd, A. (2015), *International Relations Theory: The Game-Theoretic Approach*
- Lake, D. A. & Powell, R. (eds.) (1999), *Strategic Choice and International Relations*

B. Relevant Strategic Concepts

The literature on international security is enormous, but many debates center around a relatively small number of strategic concepts. This subsection lists introductory readings for those concepts that have been most central to the arms control and race dynamics literatures discussed below (Section 3), as well as some readings that apply the strategic concepts to relevant domains (e.g. verification in arms control, offense-defense in cyber). Many of the concepts overlap somewhat, and many will have come up in the readings in Section 2A, but exploring the same concept from different angles is generally helpful for consolidating one’s understanding.

Almost none of these readings mention artificial intelligence, but—at least if one believes that reaping the full benefits and avoiding the catastrophic risks associated with the development of artificial intelligence will require international cooperation—they are highly relevant. While going through the readings, consider occasionally pausing to think about how these strategic

[5] Those looking for broader introductions to international relations and international security could consider looking through some (graduate) course syllabi, many of which are available online.

problems are likely to manifest themselves in the context of (attempted) cooperation or conflict over AI and its applications, drawing on the discussions of AI in Section 1.

i. Bargaining

Bargaining situations are those where actors have something to gain from cooperating (the situation is not “zero-sum”) but where there are also multiple possible outcomes that favor one side more than the other (the situation is not one of “pure coordination”). Bargaining is an important part of the story of most, if not all, instances of significant international cooperation and conflict.

- Fearon, J. D. (1998), “Bargaining, Enforcement, and International Cooperation,” *International Organization* 52:2 [link] [PDF]
- Powell, R. (2002), “Bargaining Theory and International Conflict,” *Annual Review of Political Science* 5 [link] [PDF]
- Powell, R. (2006), “War as a Commitment Problem,” *International Organization* 60:1 [link] [PDF]

For those interested in more in-depth reading on bargaining, Schelling’s seminal book and a more technical (economics-focused) textbook are good starting points:

- Schelling, T. C. (1960), *The Strategy of Conflict* [PDF]
- Muthoo, A. (1999), *Bargaining Theory with Applications* [link to chapters] [PDF]

ii. Verification and Enforcement

Whether cooperation can emerge is partially dependent on how difficult it is to observe and sanction compliance with the cooperative arrangement. If observing compliance is impossible, cooperative arrangements often do not emerge at all. But even when monitoring compliance is technically possible, there are often costs associated with verification that may prevent or erode cooperation. The strategic dynamics around enforcement are thus important to understand.

- Schelling, T. C. & Halperin, M. H. (1962), *Strategy and Arms Control*, chs. 9 (“Inspection and Information”) and 10 (“Regulating an Agreement”)
- Coe, A. & Vaynman, J. (2017), “The Tragedy of Arming,” working paper [PDF]
- Dai, X. (2002), “Information Systems in Treaty Regimes,” *World Politics* 54:5 [link] [PDF]

Some longer, more empirically-driven discussions of verification can be found in:

- Busch, N. E. & Pilat, J. F. (2017), *The Politics of Weapons Inspections: Assessing WMD Monitoring and Verification Regimes*
- Gallagher, N. W. (2003), *The Politics of Verification*

iii. Communication: Signaling and Perception

Prominent among the factors that often hamper international cooperation is the difficulty of communication—or, more precisely, credible communication. Attempts to communicate may be explicit or implicit, public or private, successful or unsuccessful.

Understanding how attempts to communicate (both on the sending, or "signaling," side and on the receiving, or "perceiving," side) play out in the real world has been a central topic in international security.

- Jervis, R. (2002), "Signaling and Perception: Drawing Inferences and Projecting Images," ch. 16 in *Handbook of Political Psychology* [link]
- Trager, R. F. (2016), "The Diplomacy of War and Peace," *Annual Review of Political Science* 19 [link] [PDF]
- O'Neill, B. (2018), "International Negotiation: Some Conceptual Developments," *Annual Review of Political Science* 21 [link] [PDF]

iv. Deterrence and Assurance

In bargaining situations, actors often face competing incentives: on the one hand they want to convince the other side(s) that they will not cede ground (i.e. that they are "resolved"), but on the other hand they also do not want their aims to be perceived as unlimited lest negotiations break down definitively (i.e. they want to "assure"). This tension is an important part of most international cooperation and conflict, and is thus likely to surface in the context of artificial intelligence as well.

- Jervis, R. (1976), "Deterrence, the Spiral Model, and Intentions of the Adversary," ch. 3 in *Perception and Misperception in International Politics* [PDF]
- Kydd, A. & McManus, R. W. (2017), "Threats and Assurances in Crisis Bargaining," *Journal of Conflict Resolution* 61:2 [link] [PDF]

Further related discussion (e.g. of the difference between "deterrence" and "compellence") can be found in:

- Schelling, T. C. (1966), *Arms and Influence*, especially ch. 2 ("The Art of Commitment") [PDF]

v. The Offense-Defense Balance

A prominent idea in international security says that many technologies have properties that tend to make them favorable to either the attacking or the defending side, should conflict break out. Cooperation is generally thought to be more difficult when the offense has the advantage. How artificial intelligence and its applications are likely to shape the offense-defense balance in different domains is thus an important question (Garfinkel & Dafoe discuss this explicitly in the context of cyber).

- Garfinkel, B. & Dafoe, A. (2018), "How Does the Offense-Defense Balance Scale?" [PDF]
- Glaser, C. L. & Kaufmann, C. (1998), "What Is the Offense-Defense Balance and Can We Measure It?" *International Security* 22:4 [link] [PDF] [published responses]
- Jervis, R. (1978), "Cooperation under the Security Dilemma," *World Politics* 30:2 [link] [PDF]
  - See also the Fearon (2018) reference from Section 2A on how the offense-defense balance matters in a different paradigm from Jervis's.

More in-depth discussions and two recent applications to cyber security can be found in:

- Brown, M. E. et al (eds.) (2004), *Offense, Defense, and War* [link] [ToC]
- Buchanan, B. (2017), *The Cybersecurity Dilemma: Hacking, Trust, and Fear Between Nations*, Introduction and ch. 5
- Slayton, R. (2017), "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment," *International Security* 41:3 [link] [PDF]

vi. Norms, Institutions, and Regimes

Cooperation can take many forms. Informal cooperation often functions through norms, while formal cooperation can involve single or multiple interlocking treaties and institutions that, beyond some level of complexity, tend to be referred to as "regimes." Each form of cooperation has upsides and downsides, and a significant body of work investigates when and how these different forms emerge and which circumstances call for which form.

- Barrett, S. (2007), *Why Cooperate? The Incentive to Supply Global Public Goods* [link to chapters]
- Koremenos, B., Lipson, C. & Snidal, D. (2001), "The Rational Design of International Institutions," *International Organization* 55:4 [link] [PDF]
    - Further discussion and empirical applications of this theory can be found in the eponymous book [link to chapters]
- Morrow, J. D. (2002), "The Laws of War, Common Conjectures, and Legal Systems in International Politics," *Journal of Legal Studies* 31:S1 [link] [PDF]
    - A detailed account of the theory that situates it within IR theory more explicitly can be found in *Order Within Anarchy* (2014) [link to chapters]

An in-depth review of the social science literature on norms, along with an application to cyber security, is in:

- Finnemore, M. & Hollis, D. B. (2016), "Constructing Norms for Global Cybersecurity," *American Journal of International Law* [link]

3. Topics

You should now have an understanding of the most security-relevant properties of AI (Section 1) and a basic handle on some central theories and concepts in international security (Section 2). With this as background, we can now zoom in on set of narrower topics and questions that are likely to be relevant in most future discussions of AI and international security. The sections below are arranged somewhat arbitrarily—they are obviously related, but each can also be read in isolation, and one need not follow any particular order.

A. Arms Control

Although it is somewhat old, Schelling & Halperin still is the best and most readable introduction to many of the strategic dimensions of arms control:

- Schelling, T. C. & Halperin, M. H. (1962), *Strategy and Arms Control*

Note that, as many of the readings below emphasize, arms control is defined differently by different people. Some emphasize reducing the number of weapons and the scope of capabilities, while others (like Schelling & Halperin) consider any measure -- including arms purchases -- that is likely to induce mutual constraint an instance of "arms control" (the goal is often said to be the achievement of "strategic stability").

As emphasized in Section 2B, most of the readings do not mention AI, and while often the (in)applicability of past work on arms control to possible future AI dynamics will be obvious, sometimes one will have to exercise a bit of imagination. It may be worth going back and forth between the readings in this section and Section 3C ("Technological Analogies"), since questions of whether and how we can draw lessons from historical episodes is central to both.

i. Arms Control and Artificial Intelligence

Not much thinking has been done at the intersection of arms control and artificial intelligence yet, but some good work is currently emerging. One angle from which people work on this intersection is to think about whether arms control on near-term AI applications (primarily lethal autonomous weapons [LAWs]) is desirable and feasible:

- Scharre (2018), *Army of None: Autonomous Weapons and the Future of War*, Part VI, especially ch. 20 ("The Pope and the Crossbow")
- Crootof, R. (2015), "The Killer Robots Are Here: Legal and Policy Implications," *Cardozo Law Review* 36, mainly Parts III and IV [PDF]

A second angle is to consider how developments in AI are affecting the prospects for arms control in other domains (primarily nuclear, at least thus far):

- Geist, E. G. & Lohn, A. J. (2018), "How Might Artificial Intelligence Affect the Risk of Nuclear War?" *RAND Corporation* [link] [PDF]
- Lieber, K. A. & Press, D. G. (2018), "The End of Nuclear Arms Control" [PDF]

ii. The History of Arms Control

Strategic thinking on arms control was often in flux during the Cold War, and the post-Cold War period has similarly seen many changes. Part of this is due to the waxing and waning of political power of various domestic coalitions (see also subsection iii.), while other changes are the result of shifts in the international system (e.g. the number and type of relevant actors) and the accessibility of various kinds of potentially harmful technologies (with a broad trend toward the proliferation of capabilities). The following readings provide good introductory accounts of how and why arms control policy changed since WWII:

- Miller, S. E. (2003), "Skepticism Triumphant: The Bush Administration and the Waning of Arms Control," address to the International Pugwash Movement [PDF]
    - While focused on the early Bush administration, most of the points are still a fairly accurate description of attitudes in a significant part of the DC establishment today.
- Schelling, T. C. (1985), "What Went Wrong with Arms Control," *Foreign Affairs* [link]
    - Focuses (like Trachtenberg) on the theory underlying successful Cold War arms control initiatives, and how it came to be abandoned.
- Trachtenberg, M. (1991), "The Past and Future of Arms Control," *Daedelus* 120:1 [link] [PDF]

Decent book-length overviews, each with a somewhat different lens, are:

- Chevrier, M. I. (2012), *Arms Control Policy: A Guide to the Issues*
- Colby, E. A. & Gerson, M. S. (eds.) (2013), *Strategic Stability: Contending Interpretations* [PDF]
- Croft, S. (1996), *Strategies of Arms Control: A History and Typology*
- Kearn, D. W. (2015), *Great Power Security Cooperation: Arms Control and the Challenge of Technological Change*

Encyclopedia-style overviews of historical arms control agreements can be found in:

- Burns (2009), *The Evolution of Arms Control: From Antiquity to the Nuclear Age*
- Goldblat, J. (2002), *Arms Control: The New Guide to Negotiations and Agreements*, 2nd ed.

iii. The (U.S.) Politics of Arms Control

Domestic politics affects all aspects of arms control. Two actors that both support or oppose the same arms control agreement can do so for very different reasons, many of which are not directly related to the international consequences of the agreement. Political factors also often affect how well-positioned certain actors are to achieve arms control. The following readings touch on these and other domestic dynamics:

- Maurer, J. (2018), "The Purposes of Arms Control" [non-public] [blog version]
- Miller, S. E. (1984), "Politics over Promise: Domestic Impediments to Arms Control," *International Security* 8:4 [link] [PDF]

- Gallagher, N. W. (2015), "Re-thinking the Unthinkable: Arms Control in the Twenty-first Century," *The Nonproliferation Review* 22:3, mainly pp. 269-284 [PDF]
- Kreps, S. E., Saunders, E. N. & Schultz, K. A. (2017), "The Ratification Premium: Hawks, Doves, and Arms Control," *World Politics* [PDF]

iv. The Role of Ideas, Scientists, and Experts

Most work on arms control emphasizes the importance of technological factors (e.g. verification methods) or structural variables (e.g. the distribution of power). Another strand of readings, however, focuses on the role played by individuals and groups in affecting arms control dynamics. Given the central role that scientists and industry are likely to play in efforts at mutual restraint in AI, this set of readings is likely to have some relevance for the AI domain.

- Adler, E. (1992), "The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Arms Control," *International Organization* [link] [PDF]
- Barth, K.-H. (2003), "The Politics of Seismology: Nuclear Testing, Arms Control, and the Transformation of a Discipline," *Social Studies of Science* [link] [PDF]
- Greene, B. P. (2015), "'Captive of a Scientific-Technological Elite': Eisenhower and the Nuclear Test Ban," *Presidential Studies Quarterly* 45:1 [link] [PDF]
- Grace, K. (2015), "Leo Szilard and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation," *MIRI Technical Report* [PDF]

Book-length treatments and case histories of this topic are in:

- Evangelista, M. (1999), *Unarmed Forces: The Transnational Movement to End the Cold War*
- Hymans, J. E. C. (2012), *Achieving Nuclear Ambitions: Scientists, Politicians, and Proliferation* [PDF]
- Ouagrham-Gormley, S. B. (2014), *Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development* [PDF]
    - There is also a shorter paper (2012), "Barriers to Bioweapons: Intangible Obstacles to Proliferation," *International Security* 36:4 [link] [PDF]
- Bridger, S. (2015), *Scientists at War: The Ethics of Cold War Weapons Research* [PDF]

v. The Intellectual History of Arms Control Thinking

It can take a long time for strategic and societal thinking about a technology to coalesce into coherent frameworks, and the same is likely to be true for AI. Most work looking at how this process played out historically has focused on nuclear technology. Good examples of this kind of work are:

- Miller, S. E. (2017), "Cyber Threats, Nuclear Analogies? Divergent Trajectories in Adapting to New Dual-Use Technologies," in Perkovich & Levite, *Understanding Cyber Conflict: 14 Analogies* [PDF]

- Sims, J. (1990), *Icarus Restrained: An Intellectual History Of Nuclear Arms Control, 1945-1960*

B. Race Dynamics

News articles about AI often use the term “race” to characterize the large amounts of investment that companies and governments are making into AI. There is not a great deal of research on racing in AI directly (see subsection i.), but the general competitive dynamic that people generally refer to as “racing” has received attention in multiple literatures. Below, we focus on several aspects of the literatures on racing in IR (subsections ii.-v.) and in economics (subsections vi.-vii.), which could plausibly inform our thinking on racing in AI. At the same time, there are several clear disanalogies between racing in these two domains and a possible race in AI, which are highlighted in the different subsections below.

i. The Idea of an Artificial Intelligence Race

The idea of an “arms race” in AI is so commonplace that it even has its very own Wikipedia page. Despite this, there is not (yet) a great deal of research on what the causes and consequences of a race -- of the “arms” or “non-arms” variety -- might be, or if one is actually taking place. The following papers present early work in this direction, though none go very deep:

- ** Armstrong, S., Bostrom, N. & Shulman, C. (2016), “Racing to the Precipice: A Model of Artificial Intelligence Development,” *AI & Society* 31:2 [link] [PDF]
- * Cave, S. & ÓhÉigeartaigh, S. S. (2017), “An AI Race for Strategic Advantage: Rhetoric and Risks” [PDF]
- * Geist, E. M. (2016), “It’s Already too Late to Stop the AI Arms Race—We Must Manage It Instead,” *Bulletin of the Atomic Scientist* [link] [PDF]

Many of the introductory readings in Section 1 also discuss racing, at least in passing. After one has read some of the sections below, it may also be useful to return to the take-off scenario pieces in Section 1A and think about how a take-off scenario could influence pre-take-off race dynamics.

ii. International Relations: Arms Races

Arms races have long been seen as an important part of international relations, although less attention has been paid to the concept since the end of the Cold War. The central strategic tension that drives most thinking about arms races is that, on the one hand, arming is necessary for security, but on the other, arming is expensive and takes away resources from other areas people and governments want to invest in. Plausibly, this tension need not exist in AI (for at least some applications), given that investments can yield commercial returns and thereby grow the amount of money available for spending on other things instead of shrinking it. With this caveat in mind, however, there are still likely to be aspects of IR race thinking that are applicable to AI. Good starting points for the IR literature are:

- Fearon, J. D. (2011), "Arming and Arms Races" [PDF]
- Glaser, C. (2000), "The Causes and Consequences of Arms Races," *Annual Review of Political Science* 3 [PDF]
- Koubi, V. (1999), "Military Technology Races," *International Organization* 55:3 [link] [PDF]

A recent brief discussion (with limited empirics) in the context of cyber can be found in:

- Craig, A. & Valeriano, B. (2016), "Conceptualizing Cyber Arms Races," *ICCC* [PDF]

iii. International Relations: The Diffusion of Technology, Strategy, and Arms

One factor that usually affects race dynamics is how durable a technological lead or the advantage one gains from an innovation is likely to be. The strategic considerations are complicated. If innovations diffuse quickly, for example, this might decrease the incentive to invest in new innovations, but it may also make a race more competitive by preventing any side from gaining a large lead. Technological questions, moreover, are only part of the picture. For instance, the US Defense Innovation Board recently concluded that the DoD "does not have an innovation problem; it has an innovation adoption problem," pointing to bureaucracy rather than technical limitations as an obstacle to the integration of emerging technologies. Good starting points in the literature on diffusion are:

- Horowitz, M. (2010), *The Diffusion of Military Power: Causes and Consequences for International Politics* [link to chapters]
- Goldman, E. O. & Eliason, L. C. (eds.) (2003), *The Diffusion of Military Technology and Ideas*, especially Part IV
  - A forthcoming book that is likely to be relevant as well is Lindsay, J., *Shifting the Fog of War: Information Technology and Military Power* [link]

A good summary of the literature on nuclear (non)proliferation specifically is:

- Debs, A. & Monteiro, N. P. (2017), "Conflict and Cooperation on Nuclear Nonproliferation," *Annual Review of Political Science* 20 [link] [PDF]

For those interested in this strategic angle, a deeper dive into a specific case is:

- Gormley, D. M. (2008), *Missile Contagion: Cruise Missile Proliferation and the Threat to International Security*

iv. International Relations: Diffusion, Development, and Conflict

One possible source of AI risk (if probably a distant one) is that the prospect of technological diffusion and development lead to a preventive strike—even if it is not clear whether or when an actor will obtain a particular capability, other actors can decide that intervening today to eliminate the possibility of a power shift is worth the cost of conflict.

The following papers discuss some of the important strategic dynamics in such scenarios:

- Buchanan, B. (2017), *The Cybersecurity Dilemma,* ch. 6 ("Information Distribution and the Status Quo")
- Coe, A. J. (2018), "Containing Rogues: A Theory of Asymmetric Arming," *Journal of Politics* [PDF]
- Debs, A. & Monteiro, N. P. (2014), "Known Unknowns: Power Shifts, Uncertainty, and War," *International Organization* 68:1 [link] [PDF]

An in-depth account of a relevant case can be found in:

- Burr, W. & Richelson, J. T. (2000), "Whether to 'Strangle the Baby in the Cradle': The United States and the Chinese Nuclear Program, 1960-64," *International Security* 25:3 [link] [PDF]

v. International Relations: Case Studies

A few good (sets of) qualitative case studies on race dynamics can be found here:

- Mahnken, T., Maiolo, J. & Stevenson, D. (2016), *Arms Races in International Politics: From the Nineteenth to the Twenty-first Century* [link to chapters]
- Hammond, G. T. (1993), *Plowshares into Swords: Arms Races in International Politics, 1840-1991*
- York, H. F. (1970), *Race to Oblivion: A Participant's View of the Arms Race*
- Evangelista, M. (1988), *Innovation and the Arms Race: How the United States and the Soviet Union Develop New Military Technologies*

vi. Economics: Contests and Races in Industry

There is a very large body of relevant work in economics that addresses, among other questions, whether effort is more intense in close or distant races, why this is the case, and so forth. Early models of "patent races" focused on one-time competitions with a fixed endpoint (a single technological discovery). Later, these races were embedded in models of "sequences of innovations," in which firms compete not only to be first in one-time races but rather to dominate the market in general, often in the hope of racing competitors out of business entirely. A parallel theoretical literature on "contests" draws in insights from dynamics including but not limited to industrial competition.[6] Good introductions to these three literatures are:

- Budd, C., Harris, C. & Vickers, J. (1993), "A Model of the Evolution of Duopoly: Does the Asymmetry between Firms Tend to Increase or Decrease?", *Review of Economic Studies* 60 [link] [PDF]

[6] Other related concepts in economics include auctions (especially the "all-pay" kind) and attrition-based bargaining. These are likely to be slightly less applicable to AI than contest and race models, but may nonetheless contain relevant insights. If this area is of interest, see Bulow, J. & Klemperer, P. (1999), "The Generalized War of Attrition," *American Economic Review* 89:1 [link] for a relevant discussion and further references (see e.g. footnote discussions of all-pay auctions).

- Konrad, K. A. (2012), “Dynamic Contests and the Discouragement Effect,” *Revue d’Economie Politique* [link] [PDF]
- Harris, C. & Vickers, J. (1987), “Racing with Uncertainty,” *Review of Economic Studies* 54:1 [link] [PDF]

In economics, work on inter-firm competition is situated within the subfield of Industrial Organization (IO). For those who are interested in exploring this area further, a commonly used introductory textbook is:

- Belleflamme, P. & Peitz, M. (2010), *Industrial Organization: Markets and Strategies* [PDF]

The classic reference is:

- Tirole, J. (1988), *The Theory of Industrial Organization* [link]

vi. Economics: Case Studies

The literature on racing in economics is somewhat lacking in interesting qualitative detail, but some commonly cited empirical studies are:

- Cockburn, I. & Henderson, R. (1995), “Racing to Invest? The Dynamics of Competition in Ethical Drug Discovery,” *Journal of Economic & Management Strategy* 3:3 [link] [PDF]
- Lerner, J. (1997), “An Empirical Exploration of a Technology Race,” *RAND Journal of Economics* 28:2 [link] [PDF]

For a more recent paper, see (also check the references for more studies):

- Wang, I. K., Qian, L. & Lehrer, M. (2017), “From Technology Race to Technology Marathon: A Behavioral Explanation of Technology Advancement,” *European Management Journal* 35 [link]

C. Technological Analogies

AI has been analogized to a large range of other technologies. Sometimes comparisons mainly serve an argumentative or political purpose,[7] but they are also often used for the purpose of furthering understanding and research. There are downsides as well as upsides to this approach, especially when one analogizes at a very abstract level (“AI is like electricity”) rather than situating a comparison in a strategic context (“the verification problems with this AI application are similar to those in biotechnology”). When done carefully, however, comparing AI (applications) to other technologies can be productive.

The subsections below present an introductory set of readings on various technological categories. Those interested in using analogies in their research may also benefit from engaging with some theoretical and empirical work on the uses, advantages, and drawbacks of analogical thinking:

[7] See for example this use of the space race, a relatively popular comparison.

- On recent emerging technologies, see Crootof, R. (2018), "Autonomous Weapon Systems and the Limits of Analogy," *Harvard National Security Journal* 9 [PDF] and Pauwels, E. (2013), "Mind the Metaphor," *Nature* 500 [PDF]
- On national security, see also Khong, Y. F. (1992), *Analogies at War*; and Neustadt, R. E. & May, E. R. (1986), *Thinking in Time: The Uses of History for Decision Makers*
- More generally, see Hofstadter, D. & Sander, E. (2013), *Surfaces and Essences: Analogy as the Fuel and Fire of our Thinking*

The first two subsections focus on two aspects of technologies that are commonly said to characterize AI: its "dual-use" and "general purpose" nature (somewhat confusingly, the latter is also sometimes referred to as "omni-use").

i. Dual-Use Technology

Dual-use technologies are typically defined either as technologies that can be used for both civilian and military purposes, or, more broadly, as technologies that can be used for both positive and nefarious purposes. By either definition, AI is a dual-use technology.[8] Moreover, like some (but not all) other dual-use technologies, it has large commercial as well as (potential) military value. For (governance-focused) introductions to dual-use technologies, see:

- Harris, E. D. (ed.) (2016), *Governance of Dual-Use Technologies: Theory and Practice*, esp. the conclusion [link] [PDF]
- Resnik, D. B. (2013), "Scientific Control Over Dual-Use Research: Prospects for Self-Regulation," in Rappert & Selgelid, *On the Dual Uses of Science and Ethics: Principles, Practices, and Prospects* [book link] [chapter PDF]

ii. General Purpose Technology

AI is also often thought of as a "general purpose technology" (GPT), akin to the steam engine and electricity. A large literature in economics discusses the emergence, characteristics, and implications of GPTs. For an introduction, see:

- Bresnahan, T. (2010), "General Purpose Technologies," ch. 10 in Hall & Rosenberg, *Handbook of the Economics of Innovation*, volume 2 [book link] [chapter link] [PDF]
- Bekar, C., Carlaw, K. & Lipsey, R. (2017), "General Purpose Technologies in Theory, Application and Controversy: A Review," *Journal of Evolutionary Economics* [link]
- Korzinov, V. & Savin, I. (2018), "General Purpose Technologies as an Emergent Property," *Technological Forecasting and Social Change* [link]

[8] Illustrating the policy relevance of this issue, one of the questions flagged by the UN GGE on LAWS was: "Does the transformative character of AI and its possible ubiquity limit the [lethal autonomous weapon systems] discussion in any manner, or is AI like other dual-use technologies in the past?"

The next three subsections focus on three technology domains that are often compared to AI, whether in general or on some particular strategic dimensions. I have tried to select readings that, in addition to discussing important features of these technologies, illustrate the benefits and limitations of analogizing.

iii. Nuclear

The AI-nuclear comparison is often motivated with reference to nuclear technology's general transformative impact. While this is an interesting angle to take (see the Miller reading), there are also many ways in which nuclear technology is very different from artificial intelligence, including, notably, the available mechanisms for agreement verification (see the Acton reading and Harris's conclusion in the same volume).

- Acton, J. M. (2016), "On the Regulation of Dual-Use Nuclear Technology," in Harris, E. D., *Governance of Dual-Use Technologies: Theory and Practice* [link] [PDF]
- Miller, S. E. (2017), "Cyber Threats, Nuclear Analogies? Divergent Trajectories in Adapting to New Dual-Use Technologies," in Perkovich & Levite, *Understanding Cyber Conflict: 14 Analogies* [PDF]
- See many of the readings in Section 3A for the history of strategic thought on nuclear questions.

iv. Cyber

There are obvious similarities between AI and cyber in terms of the digital fundamentals. It is also likely that both the actors involved in and the strategic challenges to successful AI governance are going to be similar to those that we've seen in action in cyber. The literature on cyber is still relatively nascent, but the following provide a good (governance-focused) introduction:

- Lin, H. (2016), "Governance of Information Technology and Cyber Weapons," in Harris, E. D., *Governance of Dual-Use Technologies: Theory and Practice* [link] [PDF]
- Buchanan, B. (2016), *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*
- Perkovich, G. & Levite, A. E. (2017), "Conclusions," in Perkovich, G. & Levite, A. E., *Understanding Cyber Conflict: 14 Analogies* [link] [PDF]

There have also been some instructive attempts to understand cyber through analogies:

- Perkovich, G. & Levite, A. E. (eds.) (2017), *Understanding Cyber Conflict: 14 Analogies* [link] [PDF]
- Goldman & Arquilla (2014), *Cyber Analogies* [link]

For those interested in reading more about cyber, a great general reference is Max Smeets's Cyber References Project.

v. Biotechnology

An increasingly common comparison for AI is biotechnology. One reason for this is that there are, generally speaking, relatively few barriers to the development and usage of both technologies. A possible disanalogy is that biotechnology is at least superficially related to the pre-existing regime covering biological risks and weapons more broadly, both domestically and internationally, and governance appears somewhat less challenging.

- Harris, E. D. (2016), "Dual-Use Threats: The Case of Biological Technology," in Harris, E. D., *Governance of Dual-Use Technologies: Theory and Practice* [link] [PDF]
- Carus, W. S. (2017), "A Century of Biological-Weapons Programs," *The Nonproliferation Review* 24:1-2 [link] [PDF]
- Koblentz, G. D. & Mazanec, B. M. (2013), "Viral Warfare: The Security Implications of Cyber and Biological Weapons," *Comparative Strategy* 32:5 [link] [PDF]

An interesting effort at the intersection of biotechnology and national security that has potential applicability to AI is discussed in:

- Zhang, L. & Gronvall, G. K. (2018), "Red Teaming the Biological Sciences for Deliberate Threats," *Terrorism and Political Violence* [link] [PDF]

D. Government and Technology

Zooming out, there may be relevant insights in bigger-picture efforts to understand past attempts by governments to harness technologies to increase their power and improve society. Some interesting work in this general area is:

- McNeill, W. H. (1982), *The Pursuit of Power: Technology, Armed Force, and Society since A.D. 1000*
- Taylor, M. Z. (2016), *The Politics of Innovation: Why Some Countries Are Better than Others at Science and Technology* [link to chapters]
- Ruttan, V. W. (2006), "Is War Necessary for Economic Growth?", Clemons Lecture [PDF]
  - For more details, see also his two books *Technology, Growth, and Development: An Induced Innovation Perspective* (2001) and *Is War Necessary for Economic Growth? Military Procurement and Technology Development* (2006) [link to chapters]