A Theory of International AI Coordination:

Strategic implications of perceived benefits, harms, capacities, and distribution in AI development[*]

Ehrik L. Aldana[†]

Advisor: Professor Allan Dafoe

December 8, 2017

Undergraduate Senior Essay

[*] I thank my advisor, Professor Allan Dafoe, for his time, support, and introduction to this paper's subject matter in his *Global Politics of AI* seminar. Additionally, the feedback, discussion, resource recommendations, and inspiring work of friends, colleagues, and mentors in several time zones – especially Amy Fan, Carrick Flynn, Chelsea Guo, Will Hunt, Jade Leung, Matthijs Maas, Peter McIntyre, Professor Nuno Monteiro, Gabe Rissman, Thomas Weng, Baobao Zhang, and Remco Zwetsloot – were vital to this paper and are profoundly appreciated. It truly takes a village, to whom this paper is dedicated.

[†] BA Candidate, Department of Political Science, Yale University, ehrik.aldana@yale.edu

Table of Contents

1. Introduction

> *"The first technology revolution caused World War I. The second technology revolution caused World War II. This is the third technology revolution."*
>
> —Jack Ma (June 2017)[1]

> *"Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world."*
>
> —Vladimir Putin (September 2017)[2]

> *"China, Russia, soon all countries w strong computer science. Competition for AI superiority at national level most likely cause of WW3 imo."*
>
> —@elonmusk (September 2017)[3]

We have recently seen an increase in media acknowledgement of the benefits of artificial intelligence (AI), as well as the negative social implications that can arise from its development. At the same time, a growing literature has illuminated the risk that developing AI has of leading to global catastrophe[4] — and further pointed out the effect that racing dynamics has on exacerbating this risk.[5] As a result, it is becoming especially vital to understand and develop strategies to manage the human process of developing AI.

[1] Song, K. 2017 June 21. "Jack Ma: Artificial intelligence could set off WWIII, but 'humans will win'". *CNBC*.
[2] Simonite, T. 2017 September 08. "Artificial Intelligence Fuels New Global Arms Race." *Wired*.
[3] Musk, E. 2017 September 04. ""China, Russia, soon all countries w strong computer science. Competition for AI superiority at national level most likely cause of WW3 imo." [Twitter Post].
[4] See, for example, Bostrom (2014).
[5] Armstrong et al. (2016).

The current landscape suggests that AI development is being led by two main international actors: China and the United States. Moreover, speculative accounts of "competition" and "arms races" have begun to increase in prominence[6], while state actors have begun to take steps that seem to support this assessment.[7] If truly present, a racing dynamic[8] between these two actors is a cause for alarm and should inspire strategies to develop an AI Coordination Regime between these two actors.

In order to assess the likelihood of such a Coordination Regime's success, one would have to take into account the two actors' expected payoffs from cooperating or defecting from the regime. In this paper, I develop a simple theory to explain whether two international actors are likely to cooperate or compete in developing AI and analyze what variables factor into this assessment. The paper proceeds as follows:

First, I survey the relevant background of AI development and coordination by summarizing the literature on the expected benefits and harms from developing AI and what actors are relevant in an international safety context. Here, I also examine the main agenda of this paper: to better understand and begin outlining strategies to maximize coordination in AI development, despite relevant actors' varying and uncertain preferences for coordination. I refer to this as the AI Coordination Problem.

[6] For example, New York Times/Reuters. 2017, November 28. "China Racing for AI Military Edge Over U.S.: Report."
[7] Kania, Elsa. 2017 June 28. "Beyond CFIUS: The Strategic Challenge of China's Rise in Artificial Intelligence."
[8] That is, the extent to which competitors prioritize speed of development over safety (Bostrom 2014: 767)

Next, I outline my theory to better understand the dynamics of the AI Coordination Problem between two opposing international actors. In short, the theory suggests that the variables that affect the payoff structure of cooperating or defecting from an AI Coordination Regime determine which model of coordination we see arise between the two actors (modeled after normal-form game setups). Depending on which model is present, we can get a better sense of the likelihood of cooperation or defection, which can in turn inform research and policy agendas to address this. This section defines suggested payoffs variables that impact the theory and simulate the theory for each representative model based on a series of hypothetical scenarios.

Finally, I discuss the relevant policy and strategic implications this theory has on achieving international AI coordination, and assess the strengths and limitations of the theory in practice.

2. AI Development and the Coordination Problem

In this section, I survey the relevant background of AI development and coordination by summarizing the literature on the expected benefits and harms from developing AI and what actors are relevant in an international safety context. I also examine the main agenda of this paper: to better understand and begin outlining strategies to maximize coordination in AI development, despite relevant actors' varying

and uncertain preferences for coordination. I refer to this as the AI Coordination Problem.

2.1 AI Development

In recent years, artificial intelligence has grown notably in its technical capacity and in its prominence in our society. We see this in the media as prominent news sources with greater frequency highlight new developments and social impacts of AI. In the business realm, investments in AI companies are soaring.[9] In our everyday lives, we store AI technology as voice assistants in our pockets and as vehicle controllers in our garages. And impressive victories over humans in chess by AI programs are being dwarfed by AI's ability to compete with and beat humans at exponentially more difficult strategic endeavors like the games of Go and StarCraft.

On one hand, these developments outline a bright future. Advanced AI technologies have the potential to provide transformative social and economic benefits like preventing deaths in auto collisions, drastically improving healthcare, reducing poverty through economic bounty, and potentially even finding solutions to some of our most menacing problems like climate change.

At the same time, there are great harms and challenges that arise from AI's rapid development. Uneven distribution of AI's benefits could exacerbate inequality, resulting

[9] McKinsey Global Institute (2017 [2]: 5).

in higher concentrations of wealth within and among nations. Moreover, racist algorithms and lethal autonomous weapons systems force us to grapple with difficult ethical questions as we apply AI to more society realms.

Perhaps most alarming, however, is the global catastrophic risk that the unchecked development of AI presents. Most prominently addressed in Nick Bostrom's *Superintelligence* (2014), the creation of an artificial superintelligence (ASI)[10] requires exceptional care and safety measures to avoid developing an ASI whose misaligned values and capacity can result in existential risks for mankind.[11] In a particularly telling quote, Stephen Hawking, Stuart Russell, Max Tegmark, and Frank Wilczek (The Independent 2014) foreshadow this stark risk:

> *"One can imagine such technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. Whereas the short-term impact of AI depends on who controls it, the long-term impact depends now whether it can be controlled at all."*

[10] Defined by Bostrom as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills." (Bostrom 2006)
[11] For more on the existential risks of Superintelligence, see Bostrom (2014) at Chapters 6 and 8.

As new technological developments bring us closer and closer to ASI[12] and the beneficial returns to AI become more tangible and lucrative, a race-like competition between key players to develop advanced AI will become acute with potentially severe consequences regarding safety.

First-move advantage will be decisive in determining the winner of the race due to the expected exponential growth in capabilities of an AI system and resulting difficulty of other parties to catch up. As a result, concerns have been raised that such a race could create incentives to skimp on safety (Armstrong et al. 2016). Once this Pandora's Box is opened, it will be difficult to close. But who can we expect to open the Box?

2.2 Relevant Actors

In this section, I briefly argue that state governments are likely to eventually control the development of AI (either through direct development or intense monitoring and regulation of state-friendly companies)[13], and that the current landscape suggests two states in particular – China and the United States – are most likely to reach development of an advanced AI system first.

[12] A survey conducted by Grace et al (2017) showed that AI experts and researchers believe there is a 50% chance of AI outperforming humans in all tasks in 45 years.

[13] There is a scenario where a private actor might develop AI in secret from the government, but this is unlikely to be the case as government surveillance capabilities improve. See Shulman (2009).

Because of its capacity to radically affect military and intelligence systems, AI research becomes an important consideration in national security and would unlikely be ignored by political and military leaders. In the US, the military and intelligence communities have a long-standing history of supporting transformative technological advancements such as nuclear weapons, aerospace technology, cyber technology and the Internet, and biotechnology (Allen and Chan 2017: 71-110).

Today, government actors have already expressed great interest in AI as a transformative technology. In 2016, the Obama Administration developed two reports on the future of AI.[14] Meanwhile, U.S. military and intelligence agencies like the NSA and DARPA continue to fund public AI research. Furthermore, in June 2017, China unveiled a policy strategy document unveiling grand ambitions to become the world leader in AI by 2030.[15] Notably, discussions among U.S. policymakers to block Chinese investment in U.S. AI companies also began at this time (Kania 2017).

In addition to boasting the world's largest economies, China and the U.S. also lead the world in A.I. publications[16] and host the world's most prominent tech/AI companies (US: Facebook, Amazon, Google, and Tesla; China: Tencent and Baidu). Combining both countries' economic and technical ecosystem with government pressures

[14] "Preparing for the Future of Artificial Intelligence" in October 2016 and "Artificial Intelligence, Automation, and the Economy" in December 2016.
[15] Mozur, Paul. 2017 July 20. "Beijing Wants A.I. to Be Made in China by 2030." The New York Times. Accessed at: https://www.nytimes.com/2017/07/20/business/china-artificial-intelligence.html
[16] McKinsey Global Institute (2017 [2]: 5)

to develop AI, it is reasonable to conceive of an AI race primarily dominated by these two international actors.

2.3 The AI Coordination Problem

As discussed, there are both great benefits and harms to developing AI, and due to the relevance AI development has to national security, it is likely that governments will take over this development (specifically the US and China). Due to the potential global harms developing AI can cause, it would be reasonable to assume that government actors would try impose safety measures and regulations on actors developing AI, and perhaps even coordinate on an international scale to ensure that all actors developing AI might cooperate under an AI Coordination Regime[17] that sets, monitors, and enforces standards to maximize safety.

Despite this, there still might be cases where the expected benefits of pursuing AI development alone outweigh (in the perception of the actor) the potential harms that might arise. As a result, this tradeoff between costs and benefits has the potential to hinder prospects for cooperation under an AI Coordination Regime. This is what I will refer to as the AI Coordination Problem.

[17] Outlining what this Coordination Regime might look like could be the topic of future research, although potential desiderata could include legitimacy, neutrality, accountability, and technical capacity (Dafoe 2016). Such a Coordination Regime could also exist in either a unilateral scenario – where one 'team' consisting of representatives from multiple states develops AI together – or a multilateral scenario – where multiple teams simultaneously develop AI on their own while agreeing to set standards and regulations (and potentially distributive arrangements) in advance.

Solving this problem requires more understanding of its dynamics and strategic implications before hacking at it with policy solutions. It is the goal this paper to shed some light on these, particularly how the structure of preferences that result from states' understandings of the benefits and harms of AI development lead to varying prospects for coordination.

3. A Theory of International AI Coordination

In this section, I outline my theory to better understand the dynamics of the AI Coordination Problem between two opposing international actors. In short, the theory suggests that the variables that affect the payoff structure of cooperating of defecting from an AI Coordination Regime determine which model of coordination we see arise between the two actors (modeled after normal-form game setups). Depending on which model is present, we can get a better sense of the likelihood of cooperation or defection, which can in turn inform research and policy agendas to address this. This section defines suggested payoffs variables that impact the theory and simulate the theory for each representative model based on a series of hypothetical scenarios. Before getting to the theory, I will briefly examine the literature on military technology/arms racing and cooperation.

3.1 Literature review on international arms races and coordination

Arms Races

Gray (1971: 41) defines an arms race as "two or more parties perceiving themselves to be in an adversary relationship, who are increasing or improving their armaments at a rapid rate and structuring their respective military postures with a general attain to the past, current, and anticipated military and political behaviour of the other parties."

Within the arms race literature, scholars have distinguished between types of arms races depending on the nature of arming. Huntington (1958) makes a distinction between qualitative arms races (where technological developments radically transform the nature of a country's military capabilities) and quantitative arms races (where competition is driven by the sheer size of an actor's arsenal).

Intrilligator and Brito (1976) argue that qualitative/technological races can lead to greater instability than quantitative races. They suggest that new weapons (or systems) that derive from radical technological breakthroughs can render a first strike more attractive, whereas basic arms buildups provide deterrence against a first strike.

In the same vein, Sorenson (1980) argues that unexpected technological breakthroughs in weaponry raise instability in arms races. This "technological shock" factor leads actors to increase weapons research and development and maximize their overall arms capacity to guard against uncertainty.

Finally, Jervis (1978) also highlights the "security dilemma" where increases in an actor's security can inherently lead to the decreased security of a rival state. As a result of this, security-seeking actions such as increasing technical capacity (even if this is not explicitly offensive — this is particularly relevant to wide-encompassing capacity of AI) can be perceived as threatening and met with exacerbated race dynamics.

So far, the readings discussed have commented on the unique qualities of technological or qualitative arms races. Additional readings provide insight on arms characteristics that impact race dynamics. Jervis (1978) highlights the distinguishability of offensive-defensive postures as a factor in stability. If security increases can be distinguished as purely defensive, this decreases instability. Additionally, Koubi (1999) develops a model of military technological races that suggests the level of spending on research and development varies with changes in an actor's relative position in a race.

Although the development of AI at present has not yet led to a clear and convincing military arms race (although this has been suggested to be the case [Geist 2016]), the elements of the arms race literature described above suggest that AI's broad and wide-encompassing capacity can lead actors to see AI development as a threatening "technological shock" worth responding to with reinforcements or augmentations in one's own security – perhaps through bolstering one's own AI development program. As described in the previous section, this "arms" race dynamic is particularly worrisome due to the existential risks that arise from AI's development and call for appropriate

measures to mitigate it. To begin exploring this, I now look to the literature on arms control and coordination.

Coordination

In order to mitigate or prevent the deleterious effects of arms races, international relations scholars have also studied the dynamics that surround arms control agreements and under what conditions actors might coordinate with one another. Schelling and Halperin (1985: 3) offer a broad definition of arms control by defining it as "all forms of military cooperation between potential enemies in the interest of reducing the likelihood of war, its scope and violence if it occurs, and the political and economic costs of being prepared for it."

As an advocate of structural realism, Gray (1992) questions the role of arms control, as he views the balance of power as a self-sufficient and self-perpetuating system of international security that is more preferable. On the other hand, Glaser (1995) argues that rational actors under certain conditions might opt for cooperative policies. Under the assumption that actors have a combination of both competing and common interests, those actors may cooperate when those common interests compel such action. As a result, it is conceivable that international actors might agree to certain limitations or cooperative regimes to reduce insecurity and stabilize the balance of power.

Using game theoretical representations of state preferences, Downs et al. (1985) look at different policy responses to arms race de-escalation and find that the model or

'game' that underlies an arms race can affect the success of policies or strategies to mitigate or end the race.

Finally, in a historical survey of international negotiations, Garcia and Herz (2016) propose that international actors might take preventative, multilateral action in scenarios under the "commonly perceived global dimension of future potential harm" (for example the ban on laser weapons or the dedication of Antarctica and outer space solely for peaceful purposes).

Together, these elements in the arms control literature suggest that there may be potential for states as untrusting, rational actors existing in a state of international anarchy to coordinate on AI development in order to reduce future potential global harms. Specifically, it is especially important to understand where preferences of vital actors overlap and how game theory considerations might affect these preferences. The theory outlined in this paper looks at just this and will be expanded upon in the following subsection.

3.2 Four models of coordination

This subsection looks at the four predominant models that describe the situation two international actors might find themselves in when considering cooperation in developing AI, where research and development is costly and its outcome is uncertain. Each model is differentiated primarily by the payoffs to cooperating or defecting for each international actor. As such, it will be useful to consider each model using a

traditional normal-form game setup as seen in Table 1. In these abstractions, we assume two utility-maximizing actors with perfect information about each other's preferences and behaviors.

Table 1. *This table contains a representation of a payoff matrix. As is customary in game theory, the first number in each cell represents how desirable the outcome is for Row (in this case, Actor A), and the second number represents how desirable the same outcome is for Column (Actor B).*

		Actor B	
		Cooperate	Defect
Actor A	Cooperate	C, C	C, D
	Defect	D, C	D, D

Depending on the payoff structures, we can anticipate different likelihoods of and preferences for cooperation or defection on the part of the actors. These differences create four distinct models of scenarios we can expect to occur: Prisoner's Dilemma, Deadlock, Chicken, and Stag Hunt. The remainder of this subsection briefly examines each of these models and its relationship with the AI Coordination Problem.

Prisoner's Dilemma

The familiar Prisoner's Dilemma is a model that involves two actors who must decide whether to cooperate in an agreement or not. While each actor's greatest preference is to defect while their opponent cooperates, the prospect of both actors

defecting is less desirable then both actors cooperating. This is visually represented in Table 2 with each actor's preference order explicitly outlined.

Table 2*. This table contains an ordinal representation of a payoff matrix for a Prisoner's Dilemma game. As will hold for the following tables, the most preferred outcome is indicated with a '4,' and the least preferred outcome is indicated with a '1.'*

		Actor B: Cooperate	Actor B: Defect
Actor A	Cooperate	3, 3	1,4
	Defect	4, 1	2,2

Actor A's preference order: DC > CC > DD > CD
Actor B's preference order: CD > CC > DD > DC

In the context of international relations, this model has been used to describe preferences of actors when deciding to enter an arms treaty or not.[18] For example, by defecting from an arms-reduction treaty to develop more weapons, an actor can gain the upper hand on an opponent who decides to uphold the treaty by covertly continuing or increasing arms production. Meanwhile, the escalation of an arms race where neither side halts or slows progress is less desirable to each actor's safety than both fully entering the agreement.

[18] For example, see Snyder (1971) and Downs et al. (1985).

This same dynamic could hold true in the development of an AI Coordination Regime, where actors can decide whether to abide by the Coordination Regime or find a way to cheat. In this model, each actor's incentives are not fully aligned to support mutual cooperation and thus should present worry for individuals hoping to reduce the possibility of developing a harmful AI.

Chicken

Similar to the Prisoner's Dilemma, Chicken occurs when each actor's greatest preference would be to defect while their opponent cooperates. Additionally, both actors can expect a greater return if they both cooperate rather than both defect. The primary difference between the Prisoner's Dilemma and Chicken, however, is that both actors failing to cooperate is the least desired outcome of the game. As a result, a rational actor should expect to cooperate.[19] This is visually represented in Table 3 with each actor's preference order explicitly outlined.

[19] Snyder (1971).

Table 3. *This table contains an ordinal representation of a payoff matrix for a Chicken game.*

		Actor B: Cooperate	Actor B: Defect
Actor A	Cooperate	3,3	2,4
	Defect	4,2	1,1

Actor A's preference order: DC > CC > CD > DD
Actor B's preference order: CD > CC > DC > DD

In international relations, examples of Chicken have included the Cuban Missile Crisis and the concept of Mutually Assured Destruction in nuclear arms development.[20] An analogous scenario in the context of the AI Coordination Problem could be if both international actors have developed, but not yet unleashed an ASI, where knowledge of whether the technology will be beneficial or harmful is still uncertain. Because of the instantaneous nature of this particular game, we can anticipate its occurrence to be rare in the context of technology development, where opportunities to coordinate are continuous. As such, Chicken scenarios are unlikely to greatly affect AI coordination strategies, but are still important to consider as a possibility nonetheless.

[20] Snyder (1971).

Deadlock

Deadlock occurs when each actor's greatest preference would be to defect while their opponent cooperates. However, in Deadlock, the prospect of both actors defecting is more desirable than both actors cooperating. As a result, there is no conflict between self-interest and mutual benefit, and the dominant strategy of both actors would be to defect. This is visually represented in Table 3 with each actor's preference order explicitly outlined.

Table 4. *This table contains an ordinal representation of a payoff matrix for a game in Deadlock.*

		Actor B: Cooperate	Actor B: Defect
Actor A	Cooperate	2,2	1,4
Actor A	Defect	4,1	3,3

Actor A's preference order: DC > DD > CC > CD
Actor B's preference order: CD > DD > CC > DC

Deadlock is a common – if little studied – occurrence in international relations, although knowledge about how deadlocks are solved can be of practical and theoretical importance.[21] In the context of developing an AI Coordination Regime, recognizing that two competing actors are in a state of Deadlock might drive peace-maximizing

[21] Persson (1994)

individuals to pursue de-escalation strategies that differ from other game models. As a result, it is important to consider deadlock as a potential model that might explain the landscape of AI coordination.

Stag Hunt

Finally, a Stag Hunt occurs when the returns for both actors are higher if they cooperate than if either or both defect. As a result, there is no conflict between self-interest and mutual benefit, and the dominant strategy of both actors would be to cooperate. This is visually represented in Table 4 with each actor's preference order explicitly outlined.

Table 5. *This table contains a sample ordinal representation of a payoff matrix for a Stag Hunt game.*

		Actor B	
		Cooperate	Defect
Actor A	Cooperate	4,4	1,3
	Defect	3,1	2,2

Actor A's preference order: CC > DC > DD > CD
Actor B's preference order: CC > CD > DD > DC

An approximation of a Stag Hunt in international relations would be an international treaty such as the Paris Climate Accords, where the protective benefits of environmental

regulation from the harms of climate change (in theory) outweigh the benefits of economic gain from defecting.[22] In the context of the AI Coordination Problem, a Stag Hunt is the most desirable outcome as mutual cooperation results in the lowest risk of racing dynamics and associated risk of developing a harmful AI.

3.3 Determining payoffs in coordination scenarios

As stated, which model (Prisoner's Dilemma, Chicken, Deadlock, or Stag Hunt) you think accurately depicts the AI Coordination Problem (and which resulting policies should be pursued) depends on the structure of payoffs to cooperating or defecting. In each of these models, the payoffs can be most simply described as *the anticipated benefit from developing AI minus the anticipated harm from developing AI.* When looking at these components in detail, however, we see that the anticipated benefits and harms are linked to whether the actors cooperate or defect from an AI Coordination Regime. For example, if the two international actors cooperate with one another, we can expect some reduction in individual payoffs if both sides agree to distribute benefits amongst each other. The remainder of this section looks at these payoffs and the variables that determine them in more detail.[23]

[22] Although there are notable holdouts.
[23] A full list of the variables outlined in this theory can be found in Appendix A.

Both sides cooperate [CC]

If both sides cooperate in an AI Coordination Regime, we can expect their payoffs to be expressed as follows:

Payoffs for A: $P_{b|A}(AB) \cdot b_A \cdot d_A - P_{h|A}(AB) \cdot h_A$

Payoffs for B: $P_{b|B}(AB) \cdot b_B \cdot d_B - P_{h|B}(AB) \cdot h_B$

The benefit that each actor can expect to receive from an AI Coordination Regime consists of the probability that each actor believes such a regime would achieve a beneficial AI – expressed as $P_{b|A}(AB)$ for Actor A's belief and $P_{b|B}(AB)$ for Actor B – times each actor's perceived benefit of AI – expressed as b_A and b_B. Additionally, this model accounts for an AI Coordination Regime that might result in variable distribution of benefits for each actor. If the regime allows for multilateral development, for example, the actors might agree that whoever reaches AI first receives 60% of the benefit, while the other actor receives 40% of the benefit. This *distribution variable* is expressed in the model as d, where differing effects of distribution are expressed for Actors A and B as d_A and d_B respectively.[24]

Meanwhile, the harm that each actor can expect to receive from an AI Coordination Regime consists of the actor's perceived likelihood that such a regime

[24] In a bilateral AI development scenario, the distribution variable can be described as an actor's likelihood of winning * percent of benefits gained by winner (this would be reflected in the terms of the Coordination Regime). Together, the likelihood of winning and the likelihood of lagging = 1.

would create a harmful AI – expressed as $P_{h|A}(AB)$ for Actor A and $P_{h|B}(AB)$ for Actor B – times each actor's perceived harm– expressed as h_A and h_B. Here, we assume that the harm of an AI-related catastrophe would be evenly distributed amongst actors.

One side cooperates and one side defects [CD or DC]

If one side cooperates with and one side defects from the AI Coordination Regime, we can expect their payoffs to be expressed as follows (here we assume Actor A defects while Actor B cooperates):

Payoffs for A: $P_{b|A}(AB) \cdot b_A \cdot d_A + P_{b|A}(A) \cdot b_A - P_{h|A}(AB) \cdot h_A - P_{h|A}(A) \cdot h_A$

Payoffs for B: $P_{b|B}(AB) \cdot b_B \cdot d_B - P_{h|B}(AB) \cdot h_B - P_{h|B}(A) \cdot h_B$

For the defector (here, Actor A), the benefit from an AI Coordination Regime consists of the probability that they believe such a regime would achieve a beneficial AI times Actor A's perceived benefit of receiving AI with distributional considerations $[P_{b|A}(AB) \cdot b_A \cdot d_A]$. Additionally, the defector can expect to receive the additional expected benefit of defecting and covertly pursuing AI development outside of the Coordination Regime. This additional benefit is expressed here as $P_{b|A}(A) \cdot b_A$. For the cooperator (here, Actor B), the benefit they can expect to receive from cooperating would be the same as if both actors cooperated $[P_{b|B}(AB) \cdot b_B \cdot d_B]$.

Meanwhile, both actors can still expect to receive the anticipated harm that arises from a Coordination Regime $[P_{h|A\ or\ B}(AB) \cdot h_{A\ or\ B}]$. In this scenario, however, *both* actors can also anticipate to the receive additional anticipated harm from the defector pursuing their own AI development outside of the regime. Here, this is expressed as $P_{h|A\ or\ B}(A) \cdot h_{A\ or\ B}$.

Both sides defect [DD]

Finally, if both sides defect or effectively choose not to enter an AI Coordination Regime, we can expect their payoffs to be expressed as follows:

Payoffs for A: $P_{b|A}(A) \cdot b_A - P_h(A) \cdot h_A - P_{h|A}(B) \cdot h_A$

Payoffs for B: $P_{b|B}(B) \cdot b_B - P_{h|B}(B) \cdot h_B - P_{h|B}(A) \cdot h_B$

The benefit that each actor can expect to receive from this scenario is solely the probability that they achieve a beneficial AI times each actor's perceived benefit of receiving AI (without distributional considerations): $P_{b|A}(A) \cdot b_A$ for Actor A and $P_{b|B}(B) \cdot b_B$ for Actor B.

Meanwhile, the harm that each actor can expect to receive from an AI Coordination Regime consists of both the likelihood that the actor themselves will develop a harmful AI times that harm, as well as the expected harm of their opponent

developing a harmful AI. Together, this is expressed as $P_{h|A\ or\ B}(A) \cdot h_{A\ or\ B} + P_{h|A\ or\ B}(B) \cdot h_{A\ or\ B}$.

Relative risks of developing a harmful AI

One last consideration to take into account is the relationship between the probabilities of developing a harmful AI for each of these scenarios. Namely, the probability of developing a harmful AI is greatest in a scenario where both actors defect, while the probability of developing a harmful AI is lowest in a scenario where both actors cooperate. This is expressed in the following way:

$$P_h(A) \cdot h\ + P_h(B) \cdot h\ [D, D] > P_h(A) \cdot h\ [D, C]^*\ + P_h(AB) \cdot h > P_h(AB) \cdot h\ [C, C]$$

*where A is the defecting actor

The intuition behind this is laid out in Armstrong et al.'s (2016) "Racing to the precipice: a model of artificial intelligence."[25] The authors suggest each actor would be incentivized to skimp on safety precautions in order to attain the transformative and powerful benefits of AI before an opponent. By failing to agree to a Coordination Regime at all [D,D], we can expect the chance of developing a harmful AI to be highest as both actors are sparing in applying safety precautions to development.

[25] See also Bostrom (2014) at Chapter 14.

***

Altogether, the considerations discussed are displayed in Table 6 as a payoff matrix. Based on the values that each actor assigns to their payoff variables, we can expect different coordination models (Prisoner's Dilemma, Chicken, Deadlock, or Stag Hunt) to arise. The following subsection further examines these relationships and simulates scenarios in which each coordination model would be most likely.

Table 6 Payoff Matrix for AI Coordination Scenarios

		Actor B: Cooperate	Actor B: Defect
Actor A	Cooperate	$P_{b\|A}(AB) \cdot b_A \cdot d_A - P_{h\|A}(AB) \cdot h_A$, $P_{b\|B}(AB) \cdot b_B \cdot d_B - P_{h\|B}(AB) \cdot h_B$	$P_{b\|A}(AB) \cdot b_A \cdot d_A - P_{h\|A}(AB) \cdot h_A - P_h(B) \cdot h_A$, $P_{b\|B}(AB) \cdot b_B \cdot d_B + P_{b\|B}(B) \cdot b_B - P_{h\|B}(AB) \cdot h_B - P_{h\|B}(B) \cdot h_B$
	Defect	$P_{b\|A}(AB) \cdot b_A \cdot d_A + P_{b\|A}(A) \cdot b_A - P_{h\|A}(AB) \cdot h_A - P_h(A) \cdot h_A$, $P_{b\|B}(AB) \cdot b_B \cdot d_B - P_{h\|B}(AB) \cdot h_B - P_{h\|B}(A) \cdot h_B$	$P_{b\|A}(A) \cdot b_A - P_{h\|A}(A) \cdot h_A - P_{h\|A}(B) \cdot h_A$, $P_{b\|B}(B) \cdot b_B - P_{h\|B}(B) \cdot h_B - P_{h\|B}(A) \cdot h_B$

Where $P_h(A) \cdot h\,[D, D] > P_h(A) \cdot h\,[D, C] > P_h(AB) \cdot h\,[C, C]$

3.4 Simulating the theory

Using the payoff matrix in Table 6, we can simulate scenarios for AI coordination by assigning numerical values to the payoff variables.[26] The remainder of this subsection looks at numerical simulations that result in each of the four models, and discusses potential real-world hypotheticals these simulations might reflect.

[26] A link to download the theory simulator (.xlsx file) can be found here: https://goo.gl/nMSife The link is also included in the Supplementary Materials section of this paper with additional information.

Example payoff structure resulting in Prisoner's Dilemma

One example payoff structure that results in a Prisoner's Dilemma is outlined in Table 7. Here, both actors demonstrate varying uncertainty about whether they will develop a beneficial or harmful AI alone, but they both equally perceive the potential benefits of AI to be greater than the potential harms. Moreover, each actor is more confident in their own capability to develop a beneficial AI than their opponent's. The corresponding payoff matrix is displayed as Table 8.

Table 7. Payoff variables for simulated Prisoner's Dilemma

	$P_{b/A}(A)$	$P_{b/B}(A)$	$P_{b/A}(B)$	$P_{b/B}(B)$	$P_{b/A}(AB)$	$P_{b/B}(AB)$	d_A	d_B	b_A	b_B
Perceived Benefits	0.6	0.4	0.5	0.7	0.8	0.8	0.5	0.5	10	10
	$P_{h/A}(A)$	$P_{h/B}(A)$	$P_{h/A}(B)$	$P_{h/B}(B)$	$P_{h/A}(AB)$	$P_{h/B}(AB)$	h_A	h_B		
Perceived Harms	0.4	0.6	0.5	0.3	0.3	0.2	5	7	4	4

Table 8. Payoff matrix for simulated Prisoner's Dilemma. Here, values are measured in utility.

		Actor B	
		Cooperate	Defect
Actor A	Cooperate	3.2 , 3.2	1.2 , 9
	Defect	7.6 , 0.8	2.4, 2.6

Example payoff structure resulting in Deadlock

One example payoff structure that results in a Deadlock is outlined in Table 9. Here, both actors demonstrate a high degree of optimism in both their and their

opponent's ability to develop a beneficial AI, while this likelihood would only be slightly greater under a cooperation regime. Additionally, both actors perceive the potential returns to developing AI to be greater than the potential harms. The corresponding payoff matrix is displayed as Table 10.

Table 9. Payoff variables for simulated Deadlock

	$P_{b/A}(A)$	$P_{b/B}(A)$	$P_{b/A}(B)$	$P_{b/B}(B)$	$P_{b/A}(AB)$	$P_{b/B}(AB)$	d_A	d_B	b_A	b_B
Perceived Benefits	0.7	0.7	0.7	0.7	0.75	0.75	0.5	0.5	13	13
	$P_{h/A}(A)$	$P_{h/B}(A)$	$P_{h/A}(B)$	$P_{h/B}(B)$	$P_{h/A}(AB)$	$P_{h/B}(AB)$	h_A	h_B		
Perceived Harms	0.3	0.3	0.3	0.3	0.25	0.25	7	7		

Table 10. Payoff matrix for simulated Deadlock

		Actor B	
		Cooperate	Defect
Actor A	Cooperate	3.125 , 1.25	1.025 , 10.125
	Defect	10.125 , 1.025	4.9, 4.9

Example payoff structure resulting in Chicken game

One example payoff structure that results in a Chicken game is outlined in Table 11. Here, both actors demonstrate high uncertainty about whether they will develop a beneficial or harmful AI alone (both Actors see the likelihood as a 50/50 split), but they

perceive the potential benefits of AI to be slightly greater than the potential harms. The payoff matrix is displayed as Table 12.

Table 11. Payoff variables for simulated Chicken game

	$P_{b/A}(A)$	$P_{b/B}(A)$	$P_{b/A}(B)$	$P_{b/B}(B)$	$P_{b/A}(AB)$	$P_{b/B}(AB)$	d_A	d_B	b_A	b_B
Perceived Benefits	0.5	0.5	0.5	0.5	0.9	0.9	0.5	0.5	10	10
	$P_{h/A}(A)$	$P_{h/B}(A)$	$P_{h/A}(B)$	$P_{h/B}(B)$	$P_{h/A}(AB)$	$P_{h/B}(AB)$	h_A	h_B		
Perceived Harms	0.5	0.5	0.5	0.5	0.1	0.1	9	9		

Table 4. Payoff matrix for simulated Chicken game. Here, values are measured in utility.

		Actor B	
		Cooperate	Defect
Actor A	Cooperate	3.6 , 3.6	-0.9 , 4.1
	Defect	4.1 , -0.9	-4, -4

Example structure resulting in Stag Hunt

Finally, Table 13 outlines an example payoff structure that results in a Stag Hunt. Both actors are more optimistic in Actor B's chances of developing a beneficial AI, but also agree that entering an AI Coordination Regime would result in the highest chances of a beneficial AI. Moreover, the AI Coordination Regime is arranged such that Actor B is more likely to gain a higher distribution of AI's benefits. Both actors see the potential harms from developing AI to be significant greater than the potential benefits,

but expect that cooperating to develop AI could still result in a positive benefit for both parties. The corresponding payoff matrix is displayed as Table 14.

Table 13. Payoff variables for simulated Stag Hunt

	$P_{b/A}(A)$	$P_{b/B}(A)$	$P_{b/A}(B)$	$P_{b/B}(B)$	$P_{b/A}(AB)$	$P_{b/B}(AB)$	d_A	d_B	b_A	b_B
Perceived Benefits	0.5	0.4	0.7	0.7	0.9	0.9	0.3	0.7	7	7
	$P_{h/A}(A)$	$P_{h/B}(A)$	$P_{h/A}(B)$	$P_{h/B}(B)$	$P_{h/A}(AB)$	$P_{h/B}(AB)$	h_A	h_B		
Perceived Harms	0.5	0.6	0.3	0.3	0.1	0.1	17	17		

Table 14. Payoff matrix for simulated Stag Hunt

		Actor B	
		Cooperate	Defect
Actor A	Cooperate	0.19 , 2.71	-4.91 , 2.51
	Defect	-4.81 , -7.49	-10.1 , -10.4

4. Policy Implications and Discussion

I discuss in this final section the relevant policy and strategic implications this theory has on achieving international AI coordination, and assess the strengths and limitations of the theory outlined above in practice. To reiterate, the primary function of this theory is to lay out a structure for identifying what game models best represent the AI Coordination Problem, and as a result, what strategies should be applied to encourage coordination and stability.

Downs et al. (1985) look at three different types of strategies governments can take to reduce the level of arms competition with a rival: (1) a *unilateral* strategy where an actor's individual actions impact race dynamics (for example, by focusing on shifting to defensive weapons[27]), (2) a *tacit bargaining* strategy that ties defensive expenditures to those of a rival, and (3) a *negotiation* strategy composed of formal arms talks. In their paper, the authors suggest "Both the 'game' that underlies an arms race and the conditions under which it is conducted can dramatically affect the success of any strategy designed to end it" (143-144). Moreover, they also argue that pursuing all strategies at once would also be suboptimal (or even impossible due to mutual exclusivity), making it even more important to know what sort of game you're playing before pursuing a strategy (145-146). Using their intuition, the remainder of this paper looks at strategy and policy considerations relevant to some game models in the context of the AI Coordination Problem. These strategies are not meant to be exhaustive by any means, but hopefully show how the outlined theory might provide practical use and motivate further research and analysis. Finally, the paper will consider some of the practical limitations of the theory.

[27] This is additionally explored in Jervis (1978).

Strategic considerations in Prisoner's Dilemma

Continuous coordination through negotiation in a Prisoner's Dilemma is somewhat promising, although a cooperating actor runs the risk of a rival defecting if there is not an effective way to ensure and enforce cooperation in an AI Cooperation Regime. Therefore, if it is likely that both actors perceive to be in a state of Prisoner's Dilemma when deciding whether to agree on AI, strategic resources should be especially allocated to addressing this vulnerability. Furthermore, a *unilateral* strategy could be employed under a Prisoner's Dilemma in order to effect cooperation. This could be achieved through signaling lack of effort to increase an actor's military capacity (perhaps by domestic bans on AI weapon development, for example). As a result, this could reduce a rival actor's perceived relative benefits gained from developing AI.

Strategic considerations in Stag Hunt

As stated before, achieving a scenario where both actors perceive to be in a Stag Hunt is the most desirable situation from the perspective of maximizing safety from an AI catastrophe, since both actors are primed to cooperate and will maximize their benefits from doing so. In the event that both actors are in a Stag Hunt, all efforts should be made to pursue negotiations and persuade rivals of peaceful intent before the window of opportunity closes. This can be facilitated, for example, through a state leader publicly and dramatically expressing understanding of danger and willingness to negotiate with other states to achieve this.

Strategies to shift game scenarios

One final strategy that a safety-maximizing actor can employ in order to maximize chances for cooperation is to change the type of game that exists by using strategies or policies to affect the payoff variables in play. For example, Stag Hunts are likely to occur when the perceived harm of developing a harmful AI is significantly greater than the perceived benefit that comes from a beneficial AI. A relevant strategy to this insight would be to focus strategic resources on shifting public or elite opinion to recognize the catastrophic risks of AI.

Similar strategic analyses can be done on variables and variable relationships outlined in this model. For example, can the structure of distribution impact an actor's perception of the game as cooperation or defection dominated (if so, should we focus strategic resources on developing accountability strategies that can effectively enforce distribution)? Does a more optimistic/pessimistic perception of an actor's own or opponent's capabilities affect which game model they adopt? Especially as prospects of coordinating are continuous, this can be a promising strategy to pursue with the support of further landscape research to more accurately assess payoff variables and what might cause them to change.

Limitations

One significant limitation of this theory is that it assumes that the AI Coordination Problem will involve two key actors. Although Section 2 describes to some

capacity that this might be a likely event with the U.S. and China, it is still conceivable that an additional international actor can move into the fray and complicate coordination efforts.

Moreover, the usefulness of this model requires accurately gauging or forecasting variables that are hard to work with. For example, it is unlikely that even the actor themselves will be able to effectively quantify their perception of capacity, riskiness, magnitude of risk, or magnitude of benefits. Still, predicting these values and forecasting probabilities based on information we do have is valuable and should not be ignored solely because it is not perfect information.

Finally, there are a plethora of other assuredly relevant factors that this theory does not account for or fully consider such as multiple iterations of game playing, degrees of perfect information, or how other diplomacy-affecting spheres (economic policy, ideology, political institutional setup, etc.) might complicate coordination efforts. Despite the large number of variables addressed in this paper, this is at its core a simple theory with the aims of motivating additional analysis and research to branch off. While there is certainly theoretical value in creating a single model that can account for all factors and answer all questions inherent to the AI Coordination Problem, this is likely not tractable or useful to attempt (at least with human hands and minds alone).

Appendix A: Theory Variables

Independent Variables

$P_{b/A}(A)$	Probability Actor A believes it will develop a beneficial AI
$P_{b/B}(A)$	Probability Actor B believes Actor A will develop a beneficial AI
$P_{b/A}(B)$	Probability Actor A believes Actor B will develop a beneficial AI
$Pb/B(B)$	Probability Actor B believes it will develop a beneficial AI
$P_{b/A}(AB)$	Probability Actor A believes AI Coordination Regime will develop a beneficial AI
$P_{b/B}(AB)$	Probability Actor B believes AI Coordination Regime will develop a beneficial AI
d_A	Percent of benefits Actor A can expect to receive from an AI Coordination Regime
d_B	Percent of benefits Actor B can expect to receive from an AI Coordination Regime
b_A	Actor A's perceived utility from developing beneficial AI
b_B	Actor B's perceived utility from developing beneficial AI
$P_{h/A}(A)$	Probability Actor A believes it will develop a harmful AI
$P_{h/B}(A)$	Probability Actor B believes Actor A will develop a harmful AI
$P_{h/A}(B)$	Probability Actor A believes Actor B will develop a harmful AI
$P_{h/B}(B)$	Probability Actor B believes it will develop a harmful AI
$P_{h/A}(AB)$	Probability Actor A believes AI Coordination Regime will develop a harmful AI
$P_{h/B}(AB)$	Probability Actor B believes AI Coordination Regime will develop a harmful AI
h_A	Actor A's perceived harm from developing a harmful AI
h_B	Actor B's perceived harm from developing a harmful AI

Dependent Variable

Type of game model and prospect of coordination

Appendix B: Theory Simulator

A theory simulator can be found and downloaded at: https://goo.gl/nMSife

The simulator is an .xlsx file that allows the user to input values for payoff variables and see resulting payoff matrix and game types.

Bibliography

Allen, G. and Chan, T. 2017. "Artificial Intelligence and National Security." Report for Harvard Kennedy School: Belfer Center for Science and International Affairs.

Armstrong, S., Bostrom, N., & Shulman, C. 2016. "Racing to the precipice: a model of artificial intelligence development." *AI and Society*, *31*(2), 201–206.

Bostrom, N. 2006. "How long before suerintelligence?" *Linguistic and Philosophical Investigations,* 2006, Vol. 5, No. 1, pp. 11-30.

Bostrom, N. 2014. *Superintelligence: paths, dangers, strategies.* Oxford, United Kingdom: Oxford University Press.

Dafoe, A. 2016. "Cooperation, Legitimacy, and Governance in AI Development." Working Paper.

Downs, G. W., Rocke, D. M., & Siverson, R. M. 1985. "Arms Races and Cooperation." *World Politics*, *38*(1), 118–146.

Executive Office of the President. 2016. "Artificial Intelligence, Automation, and the Economy." Accessed at: https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

Executive Office of the President and National Science and Technology Council Committee on Technology. 2016. "Preparing for the Future of Artificial Intelligence." Accessed at: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Garcia, D. and Herz, M. 2016. "Preventive Action in World Politics." *Global Policy*, *7*(3), 370–379.

Geist, E. M. 2016. "It's already too late to stop the AI arms race - We must manage it instead." *Bulletin of the Atomic Scientists*, *72*(5), 318–321.

Glaser, C. 1994. "Realists as Optimists: Cooperation as Self-Help." *International Security* 19, no. 3, 50-90.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When Will AI Exceed Human Performance? Evidence from AI Experts, 1–21. Retrieved from http://arxiv.org/abs/1705.08807

Gray, C. S. 1971. "The Arms Race Phenomenon." *World Politics, 24*(1), 39-79.

Gray, C. S. 1992. *House of Cards: Why Arms Control Must Fail.* Ithaca u.a.: Cornell Univ. Press.

Huntington, S. P. 1958. "Arms Races: Prerequisites and Results." *Public Policy* 8, 41–86.

The Independent. 2014 May 01. "Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'". Accessed at http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

Intriligator, M. D., & Brito, D. L. 1976. "Formal Models of Arms Races." *Journal of Peace Science*, *2*(1), 77–88.

Jervis, R. 1978. "Cooperation Under the Security Dilemma." *World Politics, 30*(2), 167-214.

Kania, E. 2017 June 28. "Beyond CFIUS: The Strategic Challenge of China's Rise in Artificial Intelligence." Retrieved December 08, 2017, from https://www.lawfareblog.com/beyond-cfius-strategic-challenge-chinas-rise-artificial-intelligence

Koubi, V. 1999. Military Technology Races. *International Organization*, *53*(3), 537–565.

McKinsey Global Institute. 2017 [1]. "Artificial Intelligence: Implications for China." Discussion Paper.

McKinsey Global Institute. 2017 [2]. "Artificial Intelligence: The Next Digital Frontier?" Discussion Paper.

Mozur, Paul. 2017 July 20. "Beijing Wants A.I. to Be Made in China by 2030." The New York Times. Accessed at: https://www.nytimes.com/2017/07/20/business/china-artificial-intelligence.html

Musk, E. 2017 September 04. ""China, Russia, soon all countries w strong computer science. Competition for AI superiority at national level most likely cause of WW3 imo." [Twitter Post]. Retrieved from https://twitter.com/elonmusk/status/904638455761612800

New York Times/Reuters. 2017, November 28. "China Racing for AI Military Edge Over U.S.: Report.". Retrieved December 08, 2017, from https://www.nytimes.com/reuters/2017/11/28/technology/28reuters-usa-china-ai.html

Persson, S. 1994. Deadlocks in International Negotiations. *Cooperation and Conflict*, *29*(3), 211–244.

Schelling, T. C., & Halperin, M. H. 1985. *Strategy and Arms Control*. Washington: Pergamon Press.

Shulman, C. 2009. "Arms Control and Intelligence Explosions." *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.*, 6.

Snyder, Glenn H. 1971. "'Prisoner's Dilemma' and 'Chicken' Models in International Politics." *International Studies Quarterly* 15(1):66–103.

Song, K. 2017 June 21. "Jack Ma: Artificial intelligence could set off WWIII, but 'humans will win'". *CNBC*. Retrieved December 08, 2017, from

https://www.cnbc.com/2017/06/21/jack-ma-artificial-intelligence-could-set-off-a-third-world-war-but-humans-will-win.html

Sorenson, D. S. 1980. "Modeling the Nuclear Arms Race: A Search for Stability." *Journal of Peace Science* 4:169–85.

Simonite, T. 2017 September 08. "Artificial Intelligence Fuels New Global Arms Race." *Wired*. Retrieved December 08, 2017, from https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/