Artificial Intelligence, Comparative Advantage and International Cooperation

Itai Agur (International Monetary Fund) [1]

December 2018

Abstract

Expert surveys indicate that human-level AI within the coming decades is not unlikely, a development with far-reaching implications for global production and income patterns. Extending a simple Ricardian trade model, this paper shows that technological divergence can rearrange trade away from comparative advantage towards competitive advantage. Next, the paper turns to the question of how to limit a global disparity, formalizing a game on international cooperation to diffuse technology. Multipolarity of AI leadership is found to make cooperation more attainable. Moreover, a high expected AI takeoff speed facilitates sharing. Finally, the game has implications for the structure of AI cooperation negotiations, and cautions against Universal Basic Income schemes.

JEL Classification Numbers: D78, F10, H77, O30.

Keywords: Artificial Intelligence, Global inequality, Ricardian trade, Cooperative games.

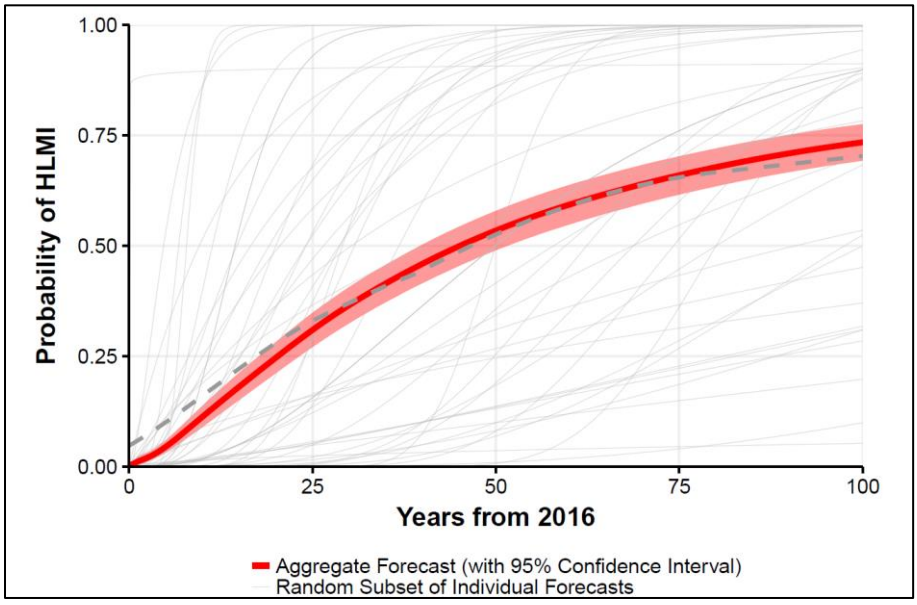
[1] IMF, Research Department, 700 19th Street NW, DC, USA. Email: iagur@imf.org. Tel: +1-202-623-4164. The views expressed in this paper are those of the author and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

I. INTRODUCTION

> *"Agents […] would prefer a sharing agreement that would guarantee them a certain slice of the future to a winner-takes-all struggle […] The presence of big potential gains from collaboration, however, does not imply that collaboration will actually be achieved"* – Nick Bostrom (2014, p. 221)

Rapid developments in Artificial Intelligence (AI) are leading to concerns about the possible emergence of Artificial General Intelligence (AGI). AGI is a form of AI that exceeds human-level abilities in a broad array of tasks, and brings with it the potential for rapid self-improvement (Bostrom, 2014). While AGI may still seem like a remote risk at present, it is arguably not sufficiently remote to ignore (Hawking et al., 2014). Indeed, surveys among AI experts lend credence to the notion of AGI as a non-negligible risk. Figure 1, taken from the survey of AI experts in Grace et al. (2017), places a mean probability of 25% on AGI within the next 20 years, although with extensive heterogeneity in opinions. Baum, Goertzel and Goertzel's (2011) survey of AI experts finds median estimates for a 50% probability of AGI in the 2040s. In Müller and Bostrom (2014) the median respondent similarly foresees a 50% probability of AGI by 2040.

Figure 1: Probability of AGI as forecasted by AI experts



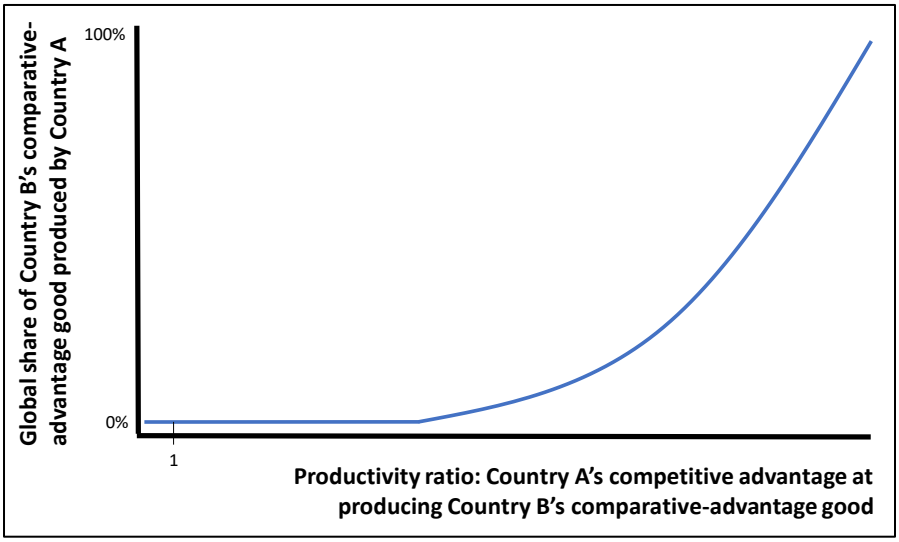
Source: Grace et al. (2017).
Note: Cumulative density functions (cdf) are used represent probabilities (1=100%) of Human Level Machine Intelligence (HLMI), a synonym for AGI. In the survey, HLMI is defined as "unaided machines can perform every task better and more cheaply than human workers". The thick red line is the mean distribution over all individual cdf's (with a subset of individual cdf's shown as the thin grey lines). The dotted grey line is a non-parametric regression over all data points performed by Grace et al. (2017), as an alternative to the mean distribution.

Looking at economic statistics, such as measures of productivity or the pace of increase in capital's share of income, there is little indication that radical change is pending (Nordhaus, 2015). However, AI can grow exponentially for a long time while having limited economic impact, and quite suddenly reach a threshold where the impact is transformational (Brynjolfsson and McAfee, 2014; Brynjolfsson, Rock and Syverson, 2017). Moreover, once

human cognitive levels are breached, the move towards superhuman intelligence may not be a large one (Muehlhauser and Salamon, 2012). The first concerns triggered by such a development are not economic: how to choose AGI's objectives, control it, and ensure an ethical application of its abilities would be the prime challenges (Muehlhauser and Helm, 2012; Bostrom and Yudkowsky, 2014; Tegmark, 2017). But if such challenges could be addressed, and a "well-behaved" AGI can be defined, developed and controlled, economic questions would quickly come to the fore.

The economic literature on AI has largely centered on its implications within countries, including the impact on employment, wages and income inequality.[2] For example, Brynjolfsson and McAfee (2014) list measures to contain the domestic inequality that the development of AI could bring about. However, AI has the potential to significantly affect patterns of comparative advantage and international trade as well (Rodrik 2016a,b, Goldfarb and Trefler 2017). An emergence of AGI could amplify this in extreme and unusual ways.

Figure 2: Competitive differences versus comparative advantage



The paper starts from a simple Ricardian model highlighting how extreme productivity differences can give rise to a dominance of competitive over comparative advantage (Figure 2). The model contains two countries, two goods, perfect competition, a labor-leisure trade-off, and country endowments (and therefore takes Ricardian theory at "face value", without considerations such as supply chains, imperfect competition, or variety preference that naturally lead to intra-industry trade). Production technologies are specified to represent both competitive and comparative advantages between the two countries. For moderate productivity differences, the usual comparative advantage separation results. But as productivity differences diverge further, competitive advantage gains ground. In essence, the usual Ricardian argument

[2] See, for instance, Acemoglu and Restrepo (2017a,b), Aghion, Jones and Jones (2017), Autor (2015a,b), Autor et al. (2017), Berg, Buffie and Zanna (2016, 2018), DeCanio (2016), Ford (2015), Freeman (2015), Frey and Osborne (2017), Hémous and Olsen (2016), Korinek and Stiglitz (2017), Sachs and Kotlikoff (2012), Smith and Andersen (2014), and The White House (2016).

on specialization by comparative advantage, is a limit case for moderate competitive differences.

Next, we introduce technology diffusion. Any technology as radically disruptive as AGI is highly unlikely to remain in private hands for long, and would be taken under a public umbrella quickly (Tegmark, 2014). Abstracting from "unintended" diffusion (i.e., through spying or AGI breaking itself out of imposed bounds) would a government have incentives to allow a supranationalization of the technology?

The paper develops a game theory model that crystallizes the key components of incentive-compatible supranationalization. The model imagines that governments sit around the table today to negotiate on the formation of a supranational holding company, to which the core of AGI technology would be transferred, if and when it emerges. This particular scenario (an AGI holding company) need not be realistic: the main aim is to investigate the incentives relating to international cooperation on AI. The forms of future global cooperation on AI may not be foreseeable at present, but the need for such cooperation is.

Three different negotiation setups are modeled: 1) countries that are ex-ante identical; 2) a single leader in AI development, most likely to host the birth of AGI; 3) a multipolar case. Countries negotiate either on a distribution of shares or on a guaranteed minimum income stream, i.e., Universal Basic Income. Countries' representative agents have concave utility functions, which provides an impetus to share expected future proceeds. The model allows for a degree of altruism – agents care at least somewhat about the world, and not only their own countries – which can soften the time-inconsistency problem.

The analysis arrives at three main results:

1. **Negotiating on Universal Basic Income** is problematic, because its distributions (relative to the global economy) are tilted towards the first years after AGI discovery, thereby amplifying incentive-compatibility problems. Negotiating about shares, conditional on country characteristics (technological leadership), is preferable.
2. **Increased competition for AI leadership** among countries helps make a sharing agreement feasible. The world of AI has arguably been moving from a single leader (US) to a bipolar setting (US and China). Table 1 provides some suggestive evidence for this, based on shifting patterns in the countries of origin for attendees of the world's largest AI conference (Goldfarb and Trefler, 2017). Furthermore, of the seven largest publicly listed corporations in the world (Apple, Amazon, Alphabet, Microsoft, Facebook, Tencent and Alibaba, as of end 2018Q2) are all intense users of AI. Of these, five are based in the US and two in China.[3]
3. **The expected takeoff speed of AGI** matters for cooperation incentives (but the expected *arrival time* of AGI does not). The faster is AGI-driven productivity growth, the sooner declining marginal benefits limit a country's utility gain and the more inclined it is to maintain the sharing arrangement. Foreseeing this, countries have an easier time agreeing when they expect a rapid takeoff.

[3] However, projects focused specifically on the development of AGI are still highly geographically concentrated in the US (Baum, 2017).

Some aspects that aid international cooperation, worsen the problem of controlling AGI. Ensuring that the actions of a technology with high or superhuman intelligence remain aligned with the intentions of its designers, is the central problem of the AI control literature (Omohundro, 2008; Russell, Dewey and Tegmark, 2015). A slower AGI takeoff improves control, as control mechanisms have more opportunity to adapt and evolve (Bostrom, 2014). Moreover, competition among AI-developing countries can worsen the control problem, if competitors end up cutting corners on safety (Armstrong, Bostrom and Shulman, 2013; Bostrom, 2017).

Table 1: Affiliation origin of researchers at global AI conference

Country	2017	2012	Change
US	34%	41%	-6%
China	23%	10%	13%
UK	5%	5%	0%
Singapore	4%	2%	2%
Japan	4%	3%	1%
Australia	3%	6%	-2%
Canada	3%	5%	-3%
India	2%	1%	1%
Hong Kong	2%	3%	-1%
Germany	2%	4%	-1%
France	2%	4%	-2%
Israel	2%	4%	-3%
Italy	2%	2%	-1%
Other	10%	10%	0%

Source: Goldfarb and Trefler (2017)
Note: Participation rates by country in which researchers are based at the Association for the Advancement of Artificial Intelligence Conferences in 2017 (San Francisco) and 2012 (Toronto).

II. COMPARATIVE VERSUS COMPETITIVE ADVANTAGE

In our basic Ricardian model there are two countries, indexed by i=1,2, two tradable goods, indexed by j=1,2, and there is perfect competition on both product and labor markets. The countries have representative agents with the following preferences over consumption and leisure:

$$U(c_{ij}, l_i) = \alpha\beta \log c_{i1} + (1-\alpha)\beta \log c_{i2} + (1-\beta) l_i \quad (1)$$

where c_{ij} is the consumption of Good j in Country i, and l_i is leisure enjoyed by the representative agent of Country i.

Every representative agent maximizes his utility function subject to the budget constraint

(2) $$B_i = w_i(1 - l_i) + E_i$$

where w_i is the wage earned by agent *i*, and we have implicitly set his total labor supply (i.e., time devoted to the production of the two goods and to leisure) equal to 1, while E_i is the endowment of agent *i*.

Consumption must satisfy

(3) $$\sum c_{ij} p_{ij} \leq B_i$$

so that, overall

(4) $$\sum c_{ij} p_{ij} \leq w_i(1 - l_i) + E_i$$

Standard optimization then gives optimal consumption and leisure for each country as

(5) $$c_{i1} = \frac{\alpha\beta(w_i(1 - l_i) + E_i)}{p_{i1}}$$

(6) $$c_{i2} = \frac{(1 - \alpha)\beta(w_i(1 - l_i) + E_i)}{p_{i2}}$$

(7) $$l_i = \min\left\{\frac{1 - \beta}{w_i}, 1\right\}$$

Next, we specify the production functions as

(8) $$y_{ij} = A_{ij} L_{ij}$$

Where A_{ij} is the productivity with which Good *j* is produced in Country *i*; L_{ij} is the amount of labor devoted to the production of Good *j* in Country *i*. Therefore, output depends only on labor input and technology here (i.e., we do not model capital as an additional input). In particular, we let

(9) $$A_{11} = x\theta_1 A_{21}$$

(10) $$A_{12} = x\theta_2 A_{22}$$

Where $\theta_1 > \theta_2 > 0$: Good 1 is Country 1's comparative advantage good, and Good 2 is Country 2's comparative advantage good. Moreover, $x \geq 1$ is the 'AGI' factor, where an AGI takeoff in Country 1 means that *x* increases (an implicit assumption being that both goods'

production is similarly affected). That is, *x* represents the competitive advantage of Country 1 over Country 2.

Comparative advantage separation (Country 1 only produces Good 1, and Country 2 only produces Good 2) occurs as long as, for Country 1, its marginal benefit of producing Good 1 is greater than its marginal benefit of producing Good 2. That is, the marginal product of labor (MPL) employed in producing Good 1 in Country 1 would need to be larger than the MPL of employing Country 1's labor in the production of Good 2.

(11) $$p_1 x \theta_1 A_{21} > p_2 x \theta_2 A_{22}$$

If this condition holds for equilibrium prices, then Country 1 optimally allocates $L_{11} = (1 - l_1)$ and $L_{12} = 0$. Instead, when allocating all of Country 1's labor to the production of Good 1, $L_{11} = (1 - l_1)$, implies that the price of Good 1 (p_1) is so low that MPL is larger in the production of Good 2, then labor in Country 1 will shift to bring about

(12) $$p_1 x \theta_1 A_{21} = p_2 x \theta_2 A_{22}$$

In this case, Country 1 would be producing part of Country 2's comparative advantage product in equilibrium.

In solving the model, the other key question is whether l_2 hits the corner ($l_2 = 1$) or not ($l_2 < 1$). Our solution technique is to assume a set of conditions hold (such as comparative advantage separation and interior labor supply), and see whether (or under what conditions) the solution indeed supports those assumptions.

As shown below, the outcome of the model is an endogenous separation into three cases:

- **Case 1)** Moderate competitive differences lead to the usual Ricardian outcome with countries specializing according to comparative advantage.
- **Case 2)** Large competitive differences imply that Country 1 starts producing both goods, but Country 2 continues to produce its comparative advantage good.
- **Case 3)** Under extreme competitive differences, Country 1 produces both goods, while Country 2 produces nothing, living purely off its endowment.

The three cases are examined formally below. The key separating conditions we find (namely equations (17), (33), and (48)) are based on the degree of competitive advantage, as represented by the variable *x*. These separating conditions are summarized in the expression below

$$x \begin{cases} < \dfrac{(1-\beta)^2 + \alpha\beta(E_1 + E_2)}{(1-\beta)^2 + (1-\alpha)\beta(E_1 + E_2)} \dfrac{1}{\theta_2} \Rightarrow Case\ 1 \\ \geq \uparrow \ and \ < \downarrow \Rightarrow Case\ 2 \\ \geq \dfrac{(1-\beta)^2 + \beta(E_1 + E_2)}{(1-\beta)^2} \dfrac{1}{\theta_2} \Rightarrow Case\ 3 \end{cases}$$

Note that in this expression, the denominator for the first (Case 1) condition is larger than for last condition, while the numerator is smaller. Hence, there always exists "space" for Case 2.

As a numerical example, take $\alpha = \frac{1}{2}, \beta = \frac{3}{4}, E_1 = E_2 = 1, A_{21} = A_{22} = 1, \theta_1 = 1, \theta_2 = \frac{1}{4}$. Then:

$$x \begin{cases} < 4 \Rightarrow Case\ 1 \\ \geq 4\ and < 100 \Rightarrow Case\ 2 \\ \geq 100 \Rightarrow Case\ 3 \end{cases}$$

Therefore, in this example, large competitive differences ($x \geq 4$) bring an end to pure comparative advantage specialization. But only extreme ($x \geq 100$) competitive differences make Country 2 stop producing altogether.

Case 1: comparative advantage separation and interior labor supply

Here, we first assume that $p_1 x \theta_1 A_{21} > p_2 x \theta_2 A_{22}$ and $l_2 < 1$. In that case, $L_{11} = (1 - l_1)$, $L_{12} = 0$, $L_{21} = 0$, and $L_{22} = (1 - l_2) > 0$. Moreover, as wages equal the MPL (given the assumption of perfect competition):

(13) $$w_1 = p_1 x \theta_1 A_{21}$$

(14) $$w_2 = p_2 A_{22}$$

Furthermore, from product market clearing

(15) $$c_{11} + c_{21} = y_{11} + y_{21}$$

(16) $$c_{12} + c_{22} = y_{12} + y_{22}$$

This leaves us with a soluble system, represented in (18)-(30) below. Using that solution, we can write the condition on comparative advantage separation (11) to

(17) $$x < \frac{(1-\beta)^2 + \alpha\beta(E_1 + E_2)}{(1-\beta)^2 + (1-\alpha)\beta(E_1 + E_2)} \frac{1}{\theta_2}$$

Therefore, when *x* is large enough, perfect separation by comparative advantage cannot be sustained as an equilibrium. But, as long as *x* is low enough, we certainly get comparative advantage separation, as well as interior labor supply: $l_2 < 1$ can be solved to $(1 - \alpha)(1 - \beta)\beta(E_1 + E_2) > 0$.

Case 1 equilibrium outcomes

(18) $$p_1 = \frac{(1-\beta)^2 + \alpha\beta(E_1 + E_2)}{(1-\beta)(1+x\theta_2)\theta_1 A_{21}}$$

(19) $$p_2 = \frac{2(1-\beta)^2 + \beta(1-\alpha)(E_1 + E_2)}{(1-\beta)A_{22}}$$

(20) $$w_1 = 1 - \beta + \frac{\alpha\beta(E_1 + E_2)}{1-\beta}$$

(21) $$w_2 = 1 - \beta + \frac{(1-\alpha)\beta(E_1 + E_2)}{1-\beta}$$

(22) $$c_{11} = \frac{\alpha\beta((1-(1-\alpha)\beta)E_1 + \alpha\beta E_2)}{(1-\beta)^2 + \alpha\beta(E_1 + E_2)} x\theta_1 A_{21}$$

(23) $$c_{21} = \frac{\alpha\beta((1-\alpha\beta)E_2 + (1-\alpha)\beta E_1)}{(1-\beta)^2 + \alpha\beta(E_1 + E_2)} x\theta_1 A_{21}$$

(24) $$c_{12} = \frac{(1-\alpha)\beta((1-(1-\alpha)\beta)E_1 + \alpha\beta E_2)}{(1-\beta)^2 + (1-\alpha)\beta(E_1 + E_2)} A_{22}$$

(25) $$c_{22} = \frac{(1-\alpha)\beta((1-\alpha\beta)E_2 + (1-\alpha)\beta E_1)}{(1-\beta)^2 + (1-\alpha)\beta(E_1 + E_2)} A_{22}$$

(26) $$l_1 = \frac{(1-\beta)^2}{(1-\beta)^2 + \alpha\beta(E_1 + E_2)}$$

(27) $$l_2 = \frac{(1-\beta)^2}{(1-\beta)^2 + (1-\alpha)\beta(E_1 + E_2)}$$

(28) $$y_{11} = \frac{\alpha\beta(E_1 + E_2)}{(1-\beta)^2 + \alpha\beta(E_1 + E_2)} x\theta_1 A_{21}$$

(29) $$y_{12} = y_{21} = 0$$

(30) $$y_{22} = \frac{(1-\alpha)\beta(E_1 + E_2)}{(1-\beta)^2 + (1-\alpha)\beta(E_1 + E_2)} A_{22}$$

Case 2: Competitive advantage separation and interior labor supply

When comparative advantage separation fails, prices must satisfy (12), because only then will the MPL (and therefore the wage) of a Country 1 worker employed in producing Good 1 equal the MPL of a Country 1 worker employed in producing Good 2. This additional condition (i.e., equation (12)) will allow us to solve for

(31) $$L_{11} + L_{12} = (1 - l_1)$$

That is, (12) and (31) replace $L_{11} = (1 - l_1)$ and $L_{12} = 0$. As before, however, $L_{22} = (1 - l_2)$ and $L_{21} = 0$.

Solving this system of equations gives as outcome (34) – (47). We can use (45) and (47) to obtain an expression for the production of Good 2 in Country 1 versus Country 2:

(32) $$\frac{y_{12}}{y_{22}} = \frac{(1-\beta)^2(x\theta_2 - 1) + \beta(E_1 + E_2)\big((1-\alpha)x\theta_2 - \alpha\big)}{\beta(E_1 + E_2) - (1-\beta)^2(x\theta_2 - 1)}$$

In this expression the denominator is positive, since we are looking at a case where $l_2 < 1$ (condition below), which means a nonzero output of y_{22}. Then, it follows that the above expression is increasing in *x*: the larger is the AGI factor, the more the share of global production of Country 2's comparative advantage good shifts to Country 1. This is depicted in Figure 2 in the introduction.

From equation (7) we have that $l_2 < 1$ as long as $w_2 > (1 - \beta)$. Replacing for w_2 from (37) below, this condition becomes

(33) $$x < \frac{(1-\beta)^2 + \beta(E_1 + E_2)}{(1-\beta)^2}\frac{1}{\theta_2}$$

Case 2 equilibrium outcomes

(34) $$p_1 = \frac{2\theta_2(1-\beta)^2 + \beta\theta_2(E_1 + E_2)}{(1-\beta)(1 + x\theta_2)\theta_1 A_{21}}$$

(35) $$p_2 = \frac{2(1-\beta)^2 + \beta(E_1 + E_2)}{(1-\beta)(1 + x\theta_2)A_{22}}$$

(36) $$w_1 = 2\frac{1-\beta}{1 + x\theta_2} + \frac{\beta x\theta_2}{1-\beta}\frac{E_1 + E_2}{1 + x\theta_2}$$

(37) $$w_2 = 2\frac{1-\beta}{1+x\theta_1} + \frac{\beta}{1-\beta}\frac{E_1+E_2}{1+x\theta_1}$$

(38) $$c_{11} = \frac{(1-\beta)\big(E_1-(1-\beta)\big) + x\theta_2\big(1+E_1+\beta(E_2+\beta-2)\big)}{2(1-\beta)^2+\beta(E_1+E_2)}\frac{\alpha\beta\theta_1 A_{21}}{\theta_2}$$

(39) $$c_{21} = \frac{(1-\beta)^2+\beta E_1+E_2+x\theta_2(1-\beta)\big(E_2-(1-\beta)\big)}{2(1-\beta)^2+\beta(E_1+E_2)}\frac{\alpha\beta\theta_1 A_{21}}{\theta_2}$$

(40) $$c_{12} = \frac{(1-\beta)\big(E_1-(1-\beta)\big) + x\theta_2\big(1+E_1+\beta(E_2+\beta-2)\big)}{2(1-\beta)^2+\beta(E_1+E_2)}(1-\alpha)\beta A_{22}$$

(41) $$c_{22} = \frac{(1-\beta)^2+\beta E_1+E_2+x\theta_2(1-\beta)\big(E_2-(1-\beta)\big)}{2(1-\beta)^2+\beta(E_1+E_2)}(1-\alpha)A_{22}$$

(42) $$l_1 = \frac{(1-\beta)^2(1+x\theta_2)}{2(1-\beta)^2+\beta(E_1+E_2)}\frac{1}{x\theta_2}$$

(43) $$l_2 = \frac{(1-\beta)^2(1+x\theta_2)}{2(1-\beta)^2+\beta(E_1+E_2)}$$

(44) $$y_{11} = \frac{\alpha\beta(E_1+E_2)}{2(1-\beta)^2+\beta(E_1+E_2)}\frac{\theta_1(1+x\theta_2)A_{21}}{\theta_2}$$

(45) $$y_{12} = \frac{(1-\beta)^2(x\theta_2-1)+\beta(E_1+E_2)\big((1-\alpha)x\theta_2-\alpha\big)}{2(1-\beta)^2+\beta(E_1+E_2)}A_{22}$$

(46) $$y_{21} = 0$$

(47) $$y_{22} = \frac{(1-\beta)^2(1-x\theta_2)+\beta(E_1+E_2)}{2(1-\beta)^2+\beta(E_1+E_2)}A_{22}$$

Case 3: Country 2 ceases production

From the discussion above, we know that when

(48) $$x \geq \frac{(1-\beta)^2 + \beta(E_1 + E_2)}{(1-\beta)^2} \frac{1}{\theta_2}$$

we have that $l_2 = 1$. In this case, Country 2 lives off its endowment income only, and the solution to the system, expressed in (49)-(60), becomes relatively simple.

(49) $$p_1 = \frac{(1-\beta)^2 + \beta(E_1 + E_2)}{1-\beta} \frac{1}{x\theta_1 A_{21}}$$

(50) $$p_2 = \frac{(1-\beta)^2 + \beta(E_1 + E_2)}{1-\beta} \frac{1}{x\theta_2 A_{22}}$$

(51) $$w_1 = 1 - \beta + \frac{\beta}{1-\beta}(E_1 + E_2)$$

(52) $$c_{11} = \frac{\alpha\beta(E_1 + \beta E_2)}{(1-\beta)^2 + \beta(E_1 + E_2)} x\theta_1 A_{21}$$

(53) $$c_{21} = \frac{\alpha(1-\beta)\beta E_2}{(1-\beta)^2 + \beta(E_1 + E_2)} x\theta_1 A_{21}$$

(54) $$c_{12} = \frac{(1-\alpha)\beta(E_1 + \beta E_2)}{(1-\beta)^2 + \beta(E_1 + E_2)} x\theta_2 A_{22}$$

(55) $$c_{22} = \frac{(1-\alpha)(1-\beta)\beta E_2}{(1-\beta)^2 + \beta(E_1 + E_2)} x\theta_2 A_{22}$$

(56) $$l_1 = \frac{(1-\beta)^2}{(1-\beta)^2 + \beta(E_1 + E_2)}$$

(57) $$l_2 = 1$$

(58) $$y_{11} = \frac{\alpha\beta(E_1 + E_2)}{(1-\beta)^2 + \beta(E_1 + E_2)} x\theta_1 A_{21}$$

(59) $$y_{12} = \frac{(1-\alpha)\beta(E_1+E_2)}{(1-\beta)^2+\beta(E_1+E_2)} x\theta_2 A_{22}$$

(60) $$y_{21} = y_{22} = 0$$

Comparing the extremes of Case 1 and Case 3, consumers in Country 2 gain from increased leisure time and lower goods prices in Case 3, but lose the wage income they had in Case 1. The net effect depends on parameter values, as can be seen from comparing equation (23) with (53), and (25) with (55). As *x* rises ever further, eventually the consumption of Good 2 by Country 2 consumers is always larger in Case 3 than in Case 1 (from (25) versus (55)). But for the consumption of Good 1 by Country 2 consumers, this remains ambiguous even in the limit of $x \to \infty$.

III. INTERNATIONAL COOPERATION GAME: EX-ANTE IDENTICAL COUNTRIES

While the net effects of AGI on Country 2 consumers in the model above are ambiguous, the model highlights the potential for AI to reorganize patterns of production and trade, which merits attention. Socio-economic sentiment (and therefore political impact) does not only depend on consumption and leisure. Inequality, and the shifting of much of production to one or a few countries, would deeply affect the countries that are "left behind".

This section builds a game that represents a global effort to limit such a divergence. The game centers on negotiations towards the establishment of a supranational holding company for future AGI technology. As discussed in more detail in the concluding section, the realism of this setup can be questioned on various counts. But the relevance of the game comes from the aim to crystallize the incentives towards supranational cooperation on AI. We believe that thinking about those incentives is important, and needs to be started early on, when such cooperation still seems a remote possibility in the political realm.

The game is built on the choice between nationalizing and supranationalizing an input-free AGI technology. The economy is simpler than in Section II, as the technology is entirely "disembodied" and, once discovered, creates ever greater economic value without labor input or capital investment. This simplicity of the economic setup is needed to keep the game tractable and flexible to a variety of negotiation modes and country setups.

At time $t = 0$, countries negotiate, knowing that (but not when and where) such AGI technology could emerge in the future. If the negotiation is successful, they create a supranational holding company, to which AGI technology is to be transferred if and when it is discovered. However, the countries internalize that such an agreement needs to be time-consistent. The country where AGI is discovered will need to have sufficient incentives to proceed with supranationalization, rather than shirking from the agreement and nationalizing the technology.

We investigate the cases of: countries that are identical at first; a country that is known to be the single AI leader from the outset; and a multipolar environment, where several countries

compete for AI leadership. Moreover, we consider the difference between negotiating about shares in the holding company and a Universal Basic Income agreement, where a fixed income stream is guaranteed.

A useful benchmark case to consider first, is that of *N* completely identical countries. This serves to highlight the essential features of the negotiation game, before adding in the complexity of different country types. An identical country setup is most conducive to reaching a sharing arrangement, because ex-ante, before the emergence of AGI, it is immediately obvious that all countries benefit from a sharing arrangement. With utility that is subject to declining marginal benefits, identical countries will necessarily prefer the certainty of a 1/*N* slice of the future AGI-pie to a 1/*N* chance of gaining the whole pie. Thus, the aspect on which a negotiation between identical countries will hinge, is the ability to design a time-consistent arrangement, one which continues to be implemented when, at a future date, they are no longer identical. Instead, when country types differ, as in Sections IV and V, then an agreement needs to insure an ex-ante desire to participate for all countries, in addition to the ex-post incentive-compatibility.

A. The setup

In each of the *N* identical countries, income in a given time period, t, is given by y_{it}, where i is the index of the countries. Utility derived from this income is $u(y_{it})$, where $u'(y_{it}) > 0$ and $u''(y_{it}) < 0$. That is, utility derived from income is increasing and concave. Global utility derived from income is:

$$U_t = \sum_i u(y_{it}) \tag{61}$$

At the negotiation described below, each country is represented by a policy maker. This policy maker represents the preferences of the country's (identical) population. These preferences do not only include the utility derived from a country's income, $u(y_{it})$. People are also assumed to intrinsically value global well-being. Simon (1992) has used the compatibility of altruism with evolutionary theory to argue for an inclusion of altruistic preferences in economic utility functions.[4] The need to model some degree of caring about global well-being arises in this analysis, because the standard ways to induce cooperation in a Nash game, such as a Tit-for-Tat retaliation strategy, would fail here.

We assume that policy makers place a weight $\gamma > 0$ on altruistic motivations, and that each individual country's policy maker maximizes:

$$\sum_t \delta^t E\left[u(y_{it}) + \gamma\left(U_t - u(y_{it})\right)\right] \tag{62}$$

[4] Altruism and compassion have a solid foundation in evolutionary biology, and can arise as optimal survival strategies in themselves or as necessary by-products of optimal strategies, which can remain in place even if the initial trigger for their evolutionary need has vanished (Dawkins, 2008).

where δ is the time discount rate. Here γ multiplies the term $\left(U_t - u(y_{it})\right)$, as this is global welfare, U_t, excluding Country i's own utility. That is, altruism is the extent to which a country cares about the total welfare of other countries. Let us simplify by rewriting $U_{\dot{\imath}t} = \left(U_t - u(y_{it})\right)$, where $U_{\dot{\imath}t}$ stands for the welfare of all countries excluding Country i:

$$\sum_t \delta^t E[u(y_{it}) + \gamma U_{\dot{\imath}t}] \tag{63}$$

We let y_{it} follow the process:

$$y_{it} = x_{it} y_{it-1} \tag{64}$$

where y_0 is an identical starting income for all countries at $t = 0$. Moreover, x_{it} is the "AGI factor":

$$x_{it} = 1 + \pi_{it}\mathrm{X} \tag{65}$$

where

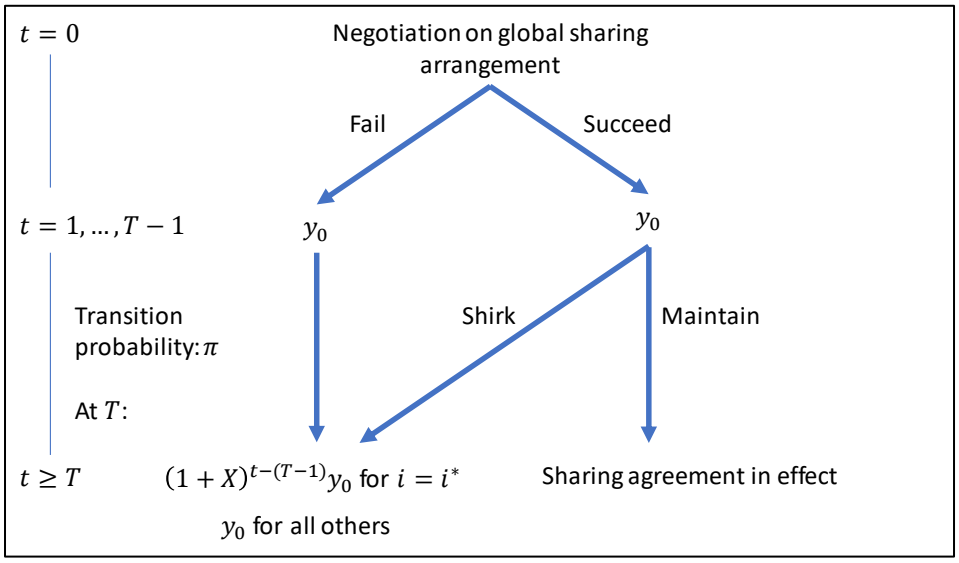
$$\pi_{it} = \begin{cases} \pi \text{ if AGI not realized by } t-1 \\ 0 \text{ if AGI realized by } t-1 \text{ in other country} \\ 1 \text{ if AGI realized by } t-1 \text{ in Country } i \end{cases} \tag{66}$$

That is, there is a given probability, π, of AGI emerging somewhere in the world in a given period. After a country has given birth to AGI, it earns the growth gain *X* for all future periods. For simplicity, we are assuming a constant boost to the growth rate, since this enables comparative statics exercises. We also assume zero real income growth for countries that do not achieve AGI, as this also simplifies the analysis without qualitative loss to the argument.[5] Obviously, in reality we would expect some positive growth externalities across borders. But what this qualitative model tries to capture is the notion of a significant growth *difference* between an AGI-discovering country and others, which is modeled most easily as a positive growth in one country, and zero growth elsewhere.

We define period T as the period in which AGI is realized, and let i^* denote the country in which this occurs. The timing of the negotiation game we will consider is then given by Figure 3.

[5] We could also let *X* be a randomly drawn variable, drawn at time *T*. This would complicate the analysis (adding in expectation terms) without any obvious benefit. As a matter of realism, neither point estimates nor well-defined distributions of *X* are particularly realistic at the current juncture. As such, we prefer the simplicity of a known *X*, which can be varied in comparative statics exercises, capturing the qualitative notion of higher/lower takeoff speed.

Figure 3: Timing of the game



In period $t = 0$ all countries sit around the table and try to negotiate a global sharing agreement, implemented by creating a shell supranational AGI holding company. If they fail, then everything proceeds as per equations (64) and (65). If they agree, then the next crucial point is date T. If the agreement proves incentive compatible, and Country i^* does not shirk, then it enters in effect – the holding company turns from a shell into an active supranational entity, which distributes its income in line with the agreed mechanism. If instead Country i^* does shirk, then we return to the economy described by (64) and (65).

We will consider two types of sharing agreements on which countries can negotiate:

1. **Share Plus**. In this arrangement, each country gets the same fraction of global income, except for Country i^*, which is given a bonus share of the total proceeds to ensure incentive compatibility.
2. **Universal Basic Income (UBI)**. A fixed income stream is defined as the universal minimum for each country to receive.

Thus, Share Plus is a negotiation about a fixed slice of the pie, which grows when the pie grows, whereas UBI is a fixed amount of pie, which becomes a smaller slice as the pie grows.

Another implicit assumption in the setup is that countries negotiate in advance, rather than waiting for period T, and starting negotiations at that time. There are two underlying justifications for this, one practical and one endogenous to the game. Practically, global negotiations take much time and effort and thus need to start well in advance of the event that countries are trying to insure against. Within the game, ex-ante all countries are better off agreeing to negotiate at time $t = 0$, when their types are still identical, rather than waiting for T, when they have diverged and Country i^* can extract a better outcome for itself, as shown for the single leader case in Section IV. That is, at time T, Country i^* is made to face a

predetermined take-it-or-leave-it decision (supranationalize or nationalize), rather than a negotiation about income division under supranationalization. [6]

B. Negotiating on income shares

The Share Plus arrangement is defined as follows:

$$y'_{it} = \begin{cases} y_0 \text{ during } t \in [1, T-1] \text{ and, subsequently, for } t \geq T: \\ y_0 + \dfrac{(1-\varphi)\left((1+X)^{t-(T-1)} - 1\right)y_0}{N} + I\varphi\left((1+X)^{t-(T-1)} - 1\right)y_0 \end{cases} \tag{67}$$

Here I is an indicator variable taking value 1 for Country i^*, and value 0 for all other countries; and $\varphi \in [0,1]$ is the "bonus" parameter: how much extra the AGI-discovering gets. This parameter is the inverse of the "sharing depth" of an agreement, as $\varphi = 1$ implies no sharing at all, and $\varphi = 0$ implies full global income sharing. The prime in y'_{it} is included to differentiate it from the income stream absent an agreement, y_{it}.

Figure 4: Negotiating on income shares

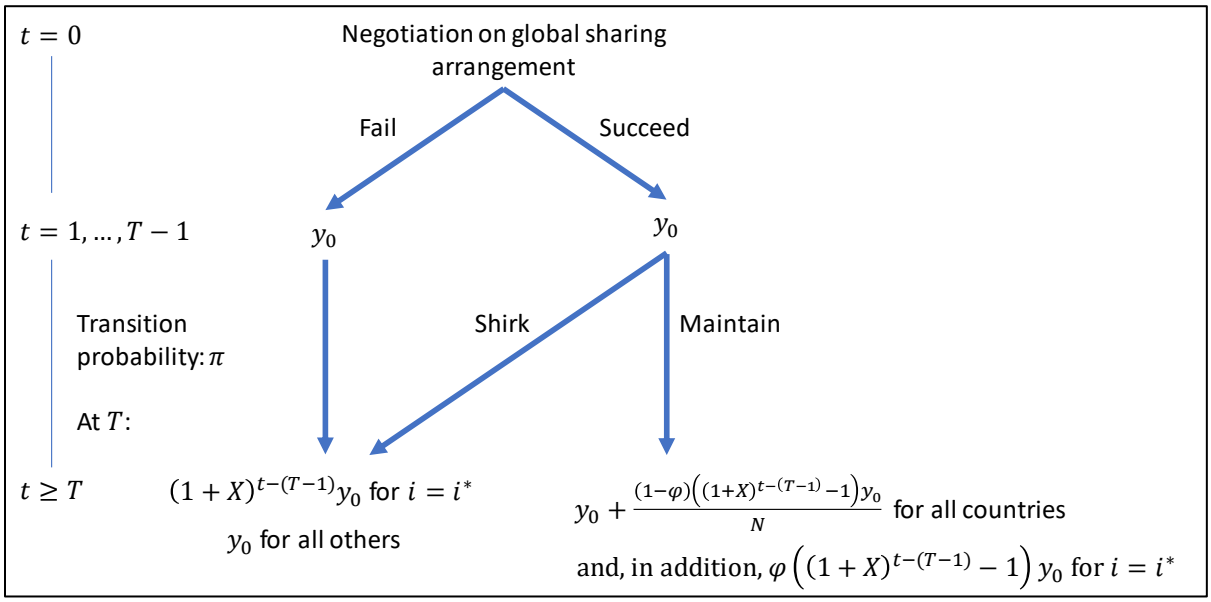


Figure 4 summarizes the Share Plus negotiation game. There are two decision nodes. The first is all countries' decision at time $t = 0$ on whether to join the agreement. The condition that determines if countries join or not is called PC: the Participation Constraint. The second decision is at time T, by Country i^*, on whether to maintain the agreement. The condition on remaining in the agreement is called IC: Incentive Compatibility. Both PC and IC play into the

[6] There are two implicit assumptions here. First, Country i^* cannot use its privileged position to force other countries to renegotiate the previously agreed shares of the holding company. Second, an agreement is the only way to implement global sharing. That is, we abstract from unilateral transfers (i.e., charity) between countries. Arguably, countries would be averse to a bilateral dependency solution.

negotiations at $t = 0$, since countries, if they want to join at all (PC), wish to ensure the future implementation of the agreement (IC).

Participation constraint

With identical countries and concave utility, it is trivial that all countries are at time $t = 0$ better off participating in a sharing agreement. Nonetheless, we write up the formal condition here, as it will be of use further on. For instance, in a multipolar world, participation is no longer trivial (Section V).

From (63), at $t = 0$, the PC of each country is (assuming break-even in favor of the agreement):

$$\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{\dot{\imath}t}] \geq \sum_t \delta^t E[u(y_{it}) + \gamma U_{\dot{\imath}t}] \tag{68}$$

where $U'_{\dot{\imath}t}$ is the total welfare of other countries under the agreement. Here, the concavity of $u(\cdot)$ implies that $\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{\dot{\imath}t}]$ is monotonically decreasing in φ. Therefore, PC is satisfied for any $\varphi \leq 1$ (i.e., always satisfied).

Incentive compatibility

Rewriting from (63) and (67), at $t = T$, Country i^* will choose to abide by the agreement if:

$$\begin{aligned} \sum_{t \geq T} \delta^{t-T} \Bigg[& u\left(y_0 \left[1 + \left((1+X)^{t-(T-1)} - 1 \right) \left(\frac{(1-\varphi)}{N} + \varphi \right) \right] \right) \\ & + \gamma(N-1) u\left(y_0 \left[1 + \frac{(1-\varphi)\left((1+X)^{t-(T-1)} - 1 \right)}{N} \right] \right) \Bigg] \\ & \geq \sum_{t \geq T} \delta^{t-T} \left[u\left(y_0 (1+X)^{t-(T-1)} \right) + \gamma(N-1) u(y_0) \right] \end{aligned} \tag{69}$$

For $\varphi = 1$ the condition in (69) holds with equality per definition, as the left-hand side becomes identical to the right-hand side. Thus, the question is whether there exists a range of values in $\varphi \in [0,1)$ such that the condition in (69) holds. We define $\varphi \in [\varphi^{IC}, 1]$ as the range of incentive-compatible sharing arrangements, where φ^{IC} is the lowest incentive-compatible value of φ.

The corner where $\varphi^{IC} = 1$ implies that no sharing agreement is feasible. For example, $\gamma = 0$ (no altruism) trivially implies that $\varphi^{IC} = 1$ in (69) and countries cannot reach any agreement that will prove time consistent. Similarly, for $\gamma \to \infty$, $\varphi^{IC} = 0$ and any agreement is incentive-compatible. We center attention on parameterizations where $\varphi^{IC} \in (0,1)$, and there is scope for, but no certainty of, an incentive-compatible sharing agreement.

Result of the negotiation

For the negotiation at $t = 0$ we assume that countries can always agree on Pareto improvements that make some countries better off (with identical countries: all countries better off), without making anyone worse off. If among the set of feasible agreements, there is only one that is Pareto dominant, then countries will come to this agreement. In the case of Share Plus, the free parameter to negotiate upon is φ. We denote the result of the negotiation by $\hat{\varphi}$.

Under identical countries it follows directly that $\hat{\varphi} = \varphi^{IC}$. As noted under (68), $\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{it}]$ monotonically decreases in φ. Therefore, at time $t = 0$, the Pareto dominant value of φ to agree upon, is the lowest φ that remains incentive compatible at $t = T$, which is φ^{IC}.

The combination of $\hat{\varphi} = \varphi^{IC}$ and the expression in (69), which defines φ^{IC}, allows us to derive the comparative statics of $\hat{\varphi}$ with respect to the underlying variables. We record this in the following result:

> **Result 1.** For parameterizations where $\hat{\varphi} \in (0,1)$: $\frac{\partial \hat{\varphi}}{\partial \delta} < 0$; $\frac{\partial \hat{\varphi}}{\partial X} < 0$; $\frac{\partial \hat{\varphi}}{\partial \gamma} < 0$.
>
> **Proof of Result 1**. In the appendix.

That increased altruism leads to deeper sharing, $\partial\hat{\varphi}/\partial\gamma < 0$, comes as no surprise. But the other two findings, $\partial\hat{\varphi}/\partial\delta < 0$, $\partial\hat{\varphi}/\partial X < 0$, are more intricate. The intuition behind these is as follows. If there is an interior solution, $\hat{\varphi} \in (0,1)$, then this means that Country i^* faces a trade-off over time, in terms of its utility within each time period. If there was no such trade-off, and intra-period utility would be larger under one alternative (shirk/remain) for all *t*, then either $\hat{\varphi} = 0$ or $\hat{\varphi} = 1$. As utility is concave and Country i^*'s income is growing fast (after period T) the relative benefits of sharing versus shirking grow over time. Hence, for any interior solution, $\hat{\varphi} \in (0,1)$, it must be true that there exists a period, call it $t' > T$, after which the intra-period benefit of sharing is greater than shirking; and vice versa for $t \in [T, t')$: in those periods the intra-period utility is larger under shirking. Therefore, a higher δ (an increased weight on future utility) makes a lower $\hat{\varphi}$ (deeper sharing) incentive compatible. Similarly, a faster takeoff speed, X, brings forward the date, t', when intra-period utility favors maintaining the agreement. For any given φ and discount rate, δ, bringing forward t' brings an agreement closer to (or further beyond) the threshold for incentive-compatibility. Therefore, a faster takeoff speed leads to a deeper sharing arrangement.

A numerical example

To visualize the game's outcomes and the comparative statics of Result 1, a numerical example is useful. Letting $u(\cdot) = ln(\cdot)$ and setting parameters $\delta = 0.9$, $X = 0.1$, $\gamma = 0.3$, $N = 10$ and $y_0 = 1$, gives the outcome in Figure 5.

Figure 5 plots φ on the horizontal axis. The vertical axis represents the utility of remaining in the agreement, that is, the left-hand side of (69). However, to ease comparison across different parameterizations, the figure shows "normalized utility", where the peak value is normalized to 1. In this numerical example, and the comparative statics shown in Figures 6-8 below, we are not interested in the exact value of utility obtained, but rather in the shape of the function. It is this shape that determines the result of the negotiation, $\hat{\varphi}$, and the exercise of interest is relating $\hat{\varphi}$ to parameter values.

Figure 5: A numerical example

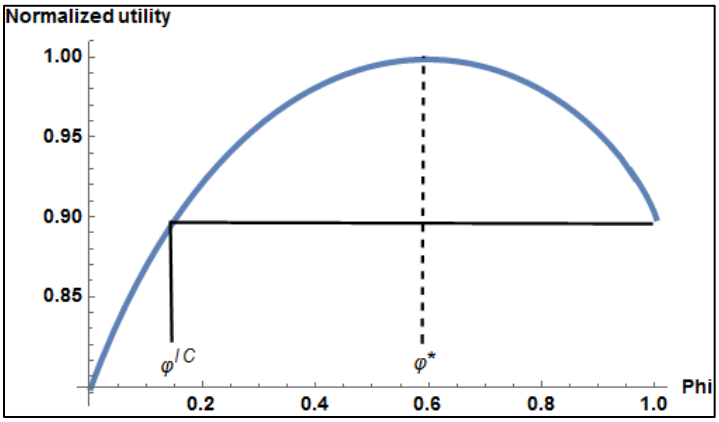


Figure 5 shows how the shape of the function (i.e., the left-hand side of (69)) determines $\hat{\varphi}$. The peak of the function is reached at a value φ^*. This is the φ that Country i^* would prefer at time T. However, if it shirks from the agreement reached at time $t = 0$, it cannot implement φ^*. Rather it implements $\varphi = 1$ with a lower associated utility. At time $t = 0$, all countries are identical, and want to attain the deepest future sharing. Therefore, they set $\hat{\varphi}$ at the lowest value for which the future utility of Country i^* will be identical to what it would get under $\varphi = 1$. This value is φ^{IC} and ensures the incentive-compatibility of the agreement.

Figures 6-8 highlight Result 1. In these figures, the blue dotted line is the parameterization of Figure 5. The red and green lines show different parameterizations with, respectively, less and more sharing as outcomes. In Figure 6, the red line represents setting $\gamma = 0.1$, while keeping all other parameters as in Figure 5. The reduction in altruism leads to a corner solution, whereby $\varphi^{IC} = 1$ and no incentive-compatible sharing can be attained by the negotiating countries at $t = 0$. The green line in Figure 6, raising γ to 0.5, represents the opposite extreme. Utility under $\varphi = 1$ is now so low that any agreement is incentive compatible, and hence countries choose $\hat{\varphi} = \varphi^{IC} = 0$.

Figures 7 and 8 portray similar comparative statics exercises for time discounting and takeoff speed. As proven in Result 1, greater patience among policy makers (higher δ) raises the attainable sharing depth of an agreement, while reduced patience among policy makers makes deep sharing less feasible (Figure 7). Similarly, a higher takeoff speed of AGI (higher X) benefits ex-ante sharing, while slower takeoff hampers such sharing.

Figure 6: Changing altruism

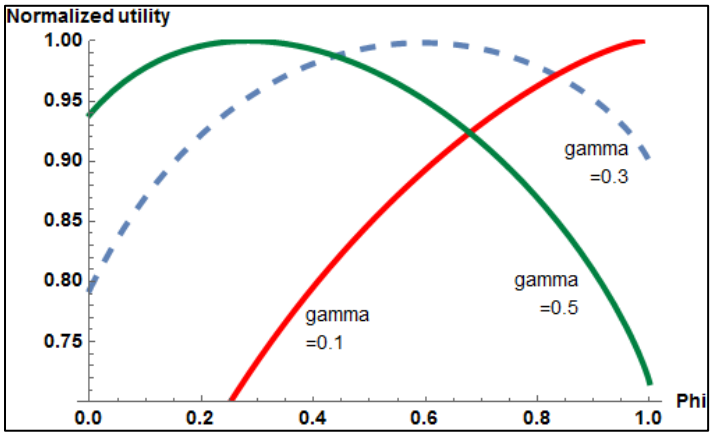


Figure 7: Changing time discounting

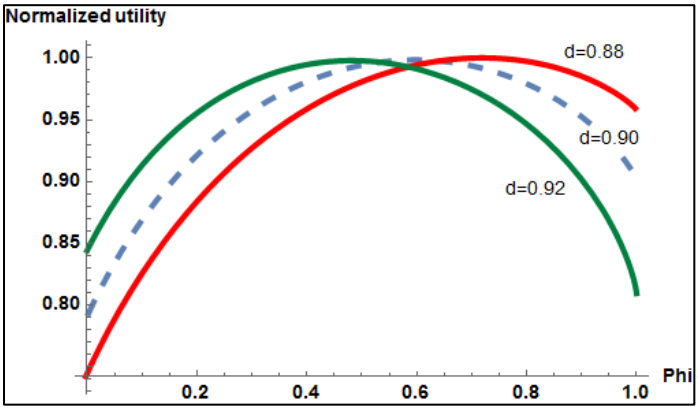
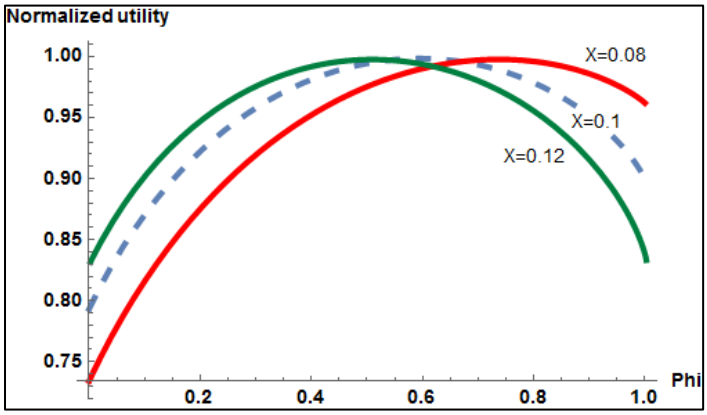


Figure 8: Changing the takeoff speed

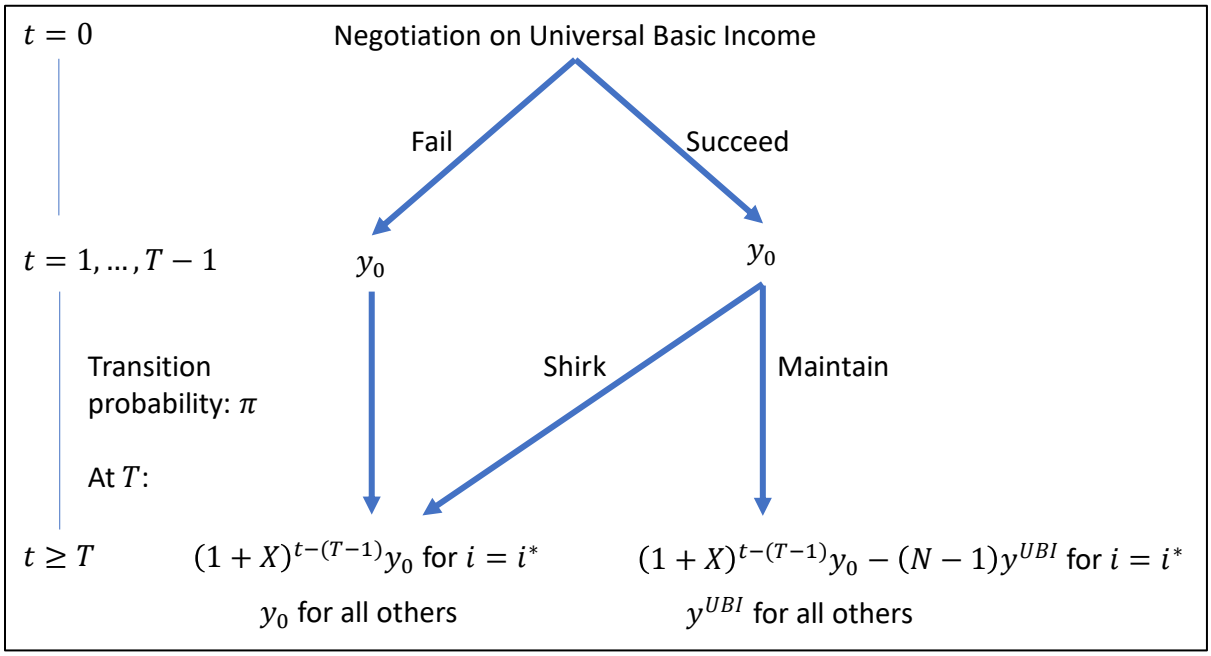


C. Negotiating on Universal Basic Income

An alternative to negotiating on income shares, is negotiating on an income stream. The notion of UBI, a government-guaranteed minimum "basic" income per person, is frequently raised in policy discussions (Brynjolfsson and McAfee, 2014). We define y^{UBI} as basic income, and consider a game where countries negotiate on y^{UBI}. This is depicted in Figure 9.

The assumption here is that y^{UBI} is constant over time, not relative to global income. This is where the difference between UBI and Share Plus (S^+) comes in. UBI offers a constant income stream. This implies a *decreasing* share of the global pie for all countries, except i^*. To "compensate", i.e., attain the same global welfare as under S^+, UBI would have to tax i^* relatively higher in early periods (after T). That is, the share of AGI proceeds being redistributed would need to be relatively high shortly after T, and would gradually decrease afterwards.

Figure 9: Negotiating on Universal Basic Income



The participation constraint at $t = 0$ is trivially satisfied for any $y^{UBI} \geq y_0$. With the same arguments as before, moreover, it directly follows that at $t = 0$ the identical countries will aim maximize y^{UBI}, subject to time consistency. That is, the outcome of the negotiation, $\hat{y}^{UBI}$, is the largest y^{UBI} that is incentive-compatible at T.

Incentive compatibility for Country i^* at T is given by:

$$\sum_{t\geq T}\delta^{t-T}\left[u\left((1+X)^{t-(T-1)}y_0 - (N-1)y^{UBI}\right) + \gamma(N-1)u(y^{UBI})\right] \geq \sum_{t\geq T}\delta^{t-T}\left[u\left(y_0(1+X)^{t-(T-1)}\right) + \gamma(N-1)u(y_0)\right] \tag{70}$$

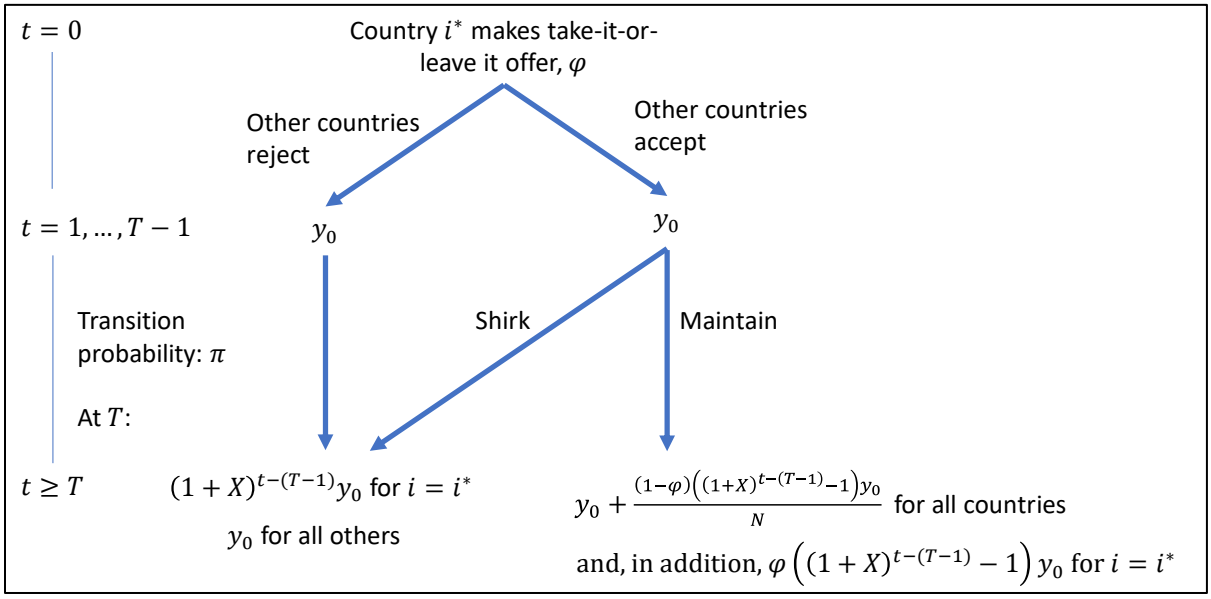
From this we can derive our result on global welfare under UBI versus Share Plus. Based on Result 2, we focus on Share Plus in the remainder of the paper.

> **Result 2**. Negotiating on UBI lowers global welfare, as compared to negotiating under S^+. For any parameterization with $\hat{\varphi} \in (0,1)$ under S^+: $\sum_{t\geq T}\delta^{t-T}U_t'(S^+) > \sum_{t\geq T}\delta^{t-T}U_t'(UBI)$.
>
> **Proof of Result 2**. In the appendix.

IV. A SINGLE LEADER

Compared to the ex-ante identical countries in Section III, we now consider the opposite extreme: a single technology leader. In this case, one country is so far in the lead in terms of AI development that, if AGI ever arises, it is certain to be in this country. In effect, countries are sorted in types (i^* and others) at time $t = 0$, and we will refer to the single leader as Country i^*. Since the single leader has full bargaining power, the negotiation game is now modelled as a take-it-or-leave-it offer by Country i^* to others. This game is depicted in Figure 10.

Figure 10: Negotiation game with a single leader



The outcome of the game depicted in Figure 10 is straightforward. Country i^* makes offer φ^*. That is, Country i^* offers the sharing parameter that maximizes its expected utility (including altruism), as given by the left-hand side of (69). Other countries are always better off accepting some sharing than no sharing. Moreover, this offer is necessarily incentive-compatible at time T. This is proven formally in Result 3. However, the intuition is directly apparent from Figure 5. The concavity of the utility function implies that there can be only one global maximum. As before, we center attention on parameterizations with interior solutions, $\varphi^{IC} \in (0,1)$, for which it follows that $\varphi^* \in (\varphi^{IC}, 1)$.

Result 3. For parameterizations where $\varphi^{IC} \in (0,1)$: $\varphi^* > \varphi^{IC}$.

Proof of Result 3. In the appendix.

The result for the single leader case is straightforward, but provides a useful benchmark when considering the multipolar case. That is, the cases of identical countries and a single leader delineate the extremes of maximally and minimally attainable sharing, respectively. Figure 5 highlights this. Identical countries can "extract" the best incentive-compatible agreement, φ^{IC}, from their own future types. They all agree that deep sharing is optimal, and only struggle against time consistency. Instead, the single leader knows its type from the outset, and extracts the share bonus that serves its preferences best, φ^*.

V. THE MULTIPOLAR CASE

We now assume that there are M countries (out of the total N countries), which are technology leaders, and have some likelihood of achieving AGI. For simplicity, the assumption is that these M countries are identical to each other, and that the other countries are also identical to each other. Thus, there are two groups of (ex-ante) identical countries. We will arrange these conveniently as $i = 1, \dots, m$ and $i = m+1, \dots, n$. That is, the countries numbered 1 through m (the first M countries) are tech leaders, while the remainder are not.

We can still use equations (64) and (65), except that the definition of π_{it} needs to be adjusted to:

$$\text{for } i \in [1, m], \pi_{it} = \begin{cases} \pi \text{ if AGI not realized by } t-1 \\ 0 \text{ if AGI realized by } t-1 \text{ in other country} \\ 1 \text{ if AGI realized by } t-1 \text{ in Country } i \end{cases} \tag{71}$$

and

$$\text{for } i \in [m+1, n], \pi_{it} = 0 \forall t \tag{72}$$

As there are now two types of countries ex-ante (AI leaders and others), we need two separate participation constraints. That is, the negotiation needs to consider countries' different starting positions, not only their different ending positions (i^* and others). We denote by φ_1 and φ_2 the two different sharing parameters that are needed here: φ_1 is the bonus offered to AI leaders (to ensure that they are willing to participate ex-ante), while φ_2 is the usual bonus earned by the AGI discovering country, to maintain incentive compatibility. Thus, parameters φ_1 and φ_2 form the core of the negotiation in the multipolar case.

The multipolar negotiated arrangement is given by (73), and the game is summarized in Figure 11.

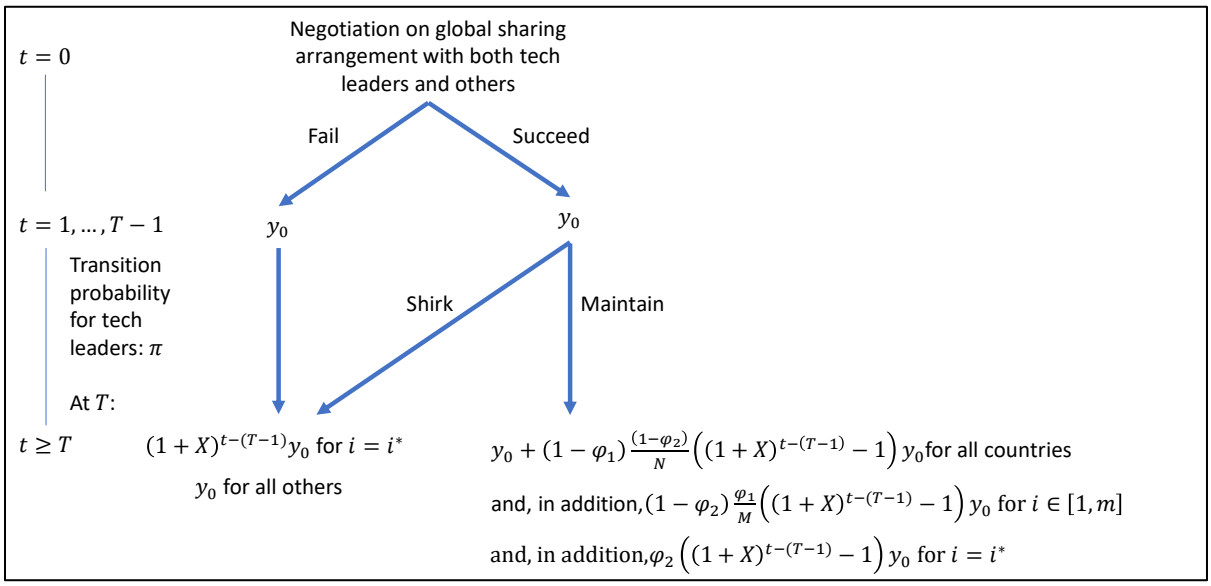
$$(73)\quad y'_{it} = \begin{cases} y_0 \text{ during } t \in [1, T-1] \text{ and, subsequently, for } t \geq T: \\ y_0 + \left[(1-\varphi_1)\dfrac{(1-\varphi_2)}{N} + I_1(1-\varphi_2)\dfrac{\varphi_1}{M} + I_2\varphi_2\right]\left((1+X)^{t-(T-1)} - 1\right)y_0 \end{cases}$$

Where $I_1 = 1$ for $i \in [1, m]$ and $I_1 = 0$ for $i \in [m+1, n]$, while I_2 takes value 1 for Country i^*, and value 0 for all other countries. Notice here that

$$(74)\qquad N(1-\varphi_1)\frac{(1-\varphi_2)}{N} + M(1-\varphi_2)\frac{\varphi_1}{M} + \varphi_2 = 1$$

and we are therefore correctly dividing up the "pie".

Figure 11: The multipolar negotiation game



Participation constraint

PC is given by (68), which is necessarily satisfied for $i \in [m+1, n]$. However, for $i \in [1, m]$, (68) can be written to

$$(75)\quad \begin{aligned} &\sum_{t\geq T}\delta^{t-T}E\left[(1+\gamma(M-1))u\left(y_0 + \left[(1-\varphi_1)\frac{(1-\varphi_2)}{N} + (1-\varphi_2)\frac{\varphi_1}{M} + \frac{1}{M}\varphi_2\right]\left((1+X)^{t-(T-1)} - 1\right)y_0\right)\right. \\ &\qquad \left. + \gamma(N-M)u\left(y_0 + (1-\varphi_1)\frac{(1-\varphi_2)}{N}\left((1+X)^{t-(T-1)} - 1\right)y_0\right)\right] \\ &\geq \sum_{t\geq T}\delta^{t-T}E\left[(1+\gamma(M-1))u\left(\frac{1}{M}y_0(1+X)^{t-(T-1)} + \frac{M-1}{M}y_0\right) + \gamma(N-M)u(y_0)\right] \end{aligned}$$

where the term $\frac{1}{M}$ comes from the ex-ante probability of becoming i^*.

Incentive compatibility
IC for the eventual Country i^* becomes

$$
\begin{aligned}
&\sum_{t\geq T}\delta^{t-T}\Bigg[u\Bigg(y_0+\left[(1-\varphi_1)\frac{(1-\varphi_2)}{N}+(1-\varphi_2)\frac{\varphi_1}{M}+\varphi_2\right]\left((1+X)^{t-(T-1)}-1\right)y_0\Bigg)\\
&\quad+\gamma(M-1)u\Bigg(y_0+\left[(1-\varphi_1)\frac{(1-\varphi_2)}{N}+(1-\varphi_2)\frac{\varphi_1}{M}\right]\left((1+X)^{t-(T-1)}-1\right)y_0\Bigg)\\
&\quad+\gamma(N-M)u\Bigg(y_0+(1-\varphi_1)\frac{(1-\varphi_2)}{N}\left((1+X)^{t-(T-1)}-1\right)y_0\Bigg)\Bigg]\\
&\quad\geq\sum_{t\geq T}\delta^{t-T}\left[u\left(y_0(1+X)^{t-(T-1)}\right)+\gamma(N-1)u(y_0)\right]
\end{aligned}
\tag{76}
$$

In the extreme where $M=N$ (all countries are tech leaders), (76) collapses to (69). Moreover, in that extreme, the left-hand side of (75) is always larger than the right-hand side, and hence PC is always satisfied. That is, for $M=N$ we return to the case of identical countries, and therefore $\hat{\varphi}_1=0$ and $\hat{\varphi}_2=\varphi^{IC}$. Instead, in the extreme where $M=1$, we return to the single leader setting, which implies $\hat{\varphi}_1+\hat{\varphi}_2=\varphi^*$ (the distinction between φ_1 and φ_2 vanishes, and the single leader extracts its optimal payoff). Overall:

> **Result 4**. The multipolar case yields an agreement, $(\hat{\varphi}_1,\hat{\varphi}_2)$, with $(\hat{\varphi}_1+\hat{\varphi}_2)\in[\varphi^{IC},\varphi^*]$.

Intuitively, the multipolar case finds itself between the extremes of identical countries and a single leader. The total "extraction" of bonuses (for the initial tech leaders and final AGI country) is at most φ^* and at least φ^{IC}. The benchmark cases of Sections III and IV define the outer reaches, and the higher is *M*, the deeper is the extent of sharing (i.e., φ moves further away (down) from φ^* and closer towards φ^{IC}).

From the perspective of AI lagging countries, a rise in *M* is beneficial. Greater competition among AI leaders essentially has a positive externality on the rest of the world, as such competition facilitates the formation of a deeper sharing arrangement.

VI. POLICY IMPLICATIONS

The conceivably boundless potential of AI, especially if it becomes self-improving, has long been the staple of a relatively small group of futurists and science fiction writers (Vinge, 1993; Kurzweil, 2005, 2013). However, recently the growth in the capacity of AI has surprised many seasoned observers, and has led to widespread debate about the potential risks of explosive technological development. For example, over eight thousand verified experts have signed the Open Letter on "Research Priorities for Robust and Beneficial Artificial Intelligence".[7] Given the possible ramifications of AGI, one need not believe that it is imminent or particularly likely to justify spending effort in mitigating some of its potentially adverse consequences.

This paper centers on the cross-border economic implications of AGI, and what may be done to mitigate them. First, extending a Ricardian trade model, comparative advantage

[7] https://futureoflife.org/ai-open-letter/.

specialization is shown to emerge only when productivity differences between countries are contained. If AGI makes such differences overwhelming, then comparative advantage theory breaks down, and competitive advantage dictates global production patterns. One could argue that, since the lagging country in the model does not necessarily become worse off, there is no deep problem here. But that would ignore the very real social and political impact of global inequality. This underlies the paper's focus on international cooperation arrangements to deal with the economic impact of AGI.

Taken at face value, negotiating and implementing an agreement on transferring AGI technology to a supranational holding company would be extremely challenging. Challenges include: convincing politicians of the relevance of such a difficult diplomatic undertaking well in advance of witnessing the technology itself; designing an agreement that is not only incentive compatible, but also considered politically palatable; defining in advance what is the appropriate trigger determining if an AGI "event" has occurred; determining (in advance) what would constitute a transferable technological core of AGI; managing the governance of a supranational holding company, and the downstream governance of the income received from that company; and, of course, ensuring that the holding company's use of AGI remains safe and unbiased.

However, even if the modeled negotiation is not necessarily realistic, it captures elements that can help in understanding global cooperation incentives more broadly. Indeed, it may be inherently impossible to think of a realistic form for global AI cooperation at the current juncture. But the potential need for such cooperation is visible on the horizon, and that justifies an incentives analysis of stepping stones and stumbling blocks on the road to cooperation. In turn, if we can identify such stepping stones and stumbling blocks, then it may be feasible to discuss policy options towards facilitating future cooperation, whatever form such cooperation may take.

Perhaps the most immediate of the implications that emerge from this paper's game theory analysis, is that multipolarity in AI leadership is important for future cooperation. As argued in the introduction, the world may already be moving from unipolarity to bipolarity in AI leadership. But the emergence of more (blocs of) countries with a claim to AI leadership, may be important. Given its enormous potential economies of scale and global externalities, the AI industry is one where the usual argument about letting countries specialize does not carry sway. It may be better for most advanced economies to attempt specialization in this field.

As discussed in the introduction, the argument on multipolarity abstracts from the risk of a race to the bottom on AI safety. This trade-off between safety concerns and cooperation incentives is seen in another of the paper's results as well. Bostrom (2014) forcefully lays out the case to (at least attempt to) slow down an AGI takeoff, as time is a crucial variable in safety, especially when learning to deal with intelligence beyond our own. Instead, in terms of global cooperation incentives, a fast takeoff is beneficial.[8]

[8] However, as our model considers long-term growth, while safety concerns center on the initial stages of AGI, the trade-off between safety and cooperation incentives is probably less pertinent for takeoff speed than for multipolarity.

Finally, our analysis cautions against global Universal Basic Income policies. The attraction of Universal Basic Income is its fairness, the fact that everyone is assured of the same minimum. But fairness does not necessarily square well with cooperation incentives. To bring about a time-consistent cooperation scheme between unequal countries, policy makers in the lagging countries may need to tolerate a degree of unfairness.

APPENDIX: PROOFS

Proof of Result 1. Consider $t \geq T$. First note that from $\left(\frac{(1-\varphi)}{N} + \varphi\right) < 1$ and $u''(\cdot) < 0$ it follows that $\frac{\partial u(y'_{i^*t})}{\partial t} > \frac{\partial u(y_{i^*t})}{\partial t}$. Moreover, $\frac{\partial U'_{i^*t}}{\partial t} > \frac{\partial U_{i^*t}}{\partial t}$ as $\frac{\partial U'_{i^*t}}{\partial t} > 0$ and $\frac{\partial U_{i^*t}}{\partial t} = 0$. Therefore $u(y'_{i^*t}) + \gamma U'_{i^*t}$ rises faster in t than does $u(y_{i^*t}) + \gamma U_{i^*t}$. A parameterization with $\hat{\varphi} \in (0,1)$ (i.e., and interior solution) must therefore have that $u(y'_{i^*T}) + \gamma U'_{i^*T} < u(y_{i^*T}) + \gamma U_{i^*T}$ since otherwise $\hat{\varphi} \to 0$. That is, at period T the intra-period utility from shirking is higher than from maintaining the agreement, and from some period $t' > T$ onwards the intra-period utility from having maintained the agreement will be larger than from having shirked.

A larger weight, through a higher δ, on periods $t \in [t', \infty)$ relative to $t \in [T, t']$ therefore makes maintaining the agreement more attractive for any given φ. By the monotonicity properties discussed below (69), this means that $\frac{\partial \varphi^{IC}}{\partial \delta} < 0$, which implies $\frac{\partial \hat{\varphi}}{\partial \delta} < 0$.

Furthermore, by $u''(\cdot) < 0$ we have that t' declines in X. By the same argument as for $\frac{\partial \varphi^{IC}}{\partial \delta} < 0$ we then also have $\frac{\partial \varphi^{IC}}{\partial X} < 0$ and $\frac{\partial \hat{\varphi}}{\partial X} < 0$.

Finally, Since $U'_{i^*t} > U_{i^*t}$ for all $t \geq T$, an increase in γ always raises the left-hand side of (69) more than the right-hand side. It follows that $\frac{\partial \varphi^{IC}}{\partial \gamma} < 0$ and hence $\frac{\partial \hat{\varphi}}{\partial \gamma} < 0$.

Proof of Result 2. Convert y^{UBI} into a path of φ_t. That is, for every period $t \geq T$ identify the value of φ that yields y^{UBI}, i.e., the value of φ such that $y_0\left[1 + \frac{(1-\varphi)\left((1+X)^{t-(T-1)}-1\right)}{N}\right] = y^{UBI}$. Note that $\sum_{t \geq T} \delta^{t-T} U'_t(S^+)$ is global utility under agreement $\hat{\varphi}$. Now define $\tilde{y}^{UBI}$ as the value of y^{UBI} which gives the same expected utility as $\sum_{t \geq T} \delta^{t-T} U'_t(S^+)$. Convert $\tilde{y}^{UBI}$ into a path of φ_t. This path has $\varphi_t < \hat{\varphi}$ (deeper sharing) at $t = T$ and for a given number of subsequent periods. Hence, by $u''(\cdot) < 0$ and $\hat{\varphi} = \varphi^{IC}$, it follows that $\tilde{y}^{UBI}$ cannot be incentive compatible. That is, $\hat{y}^{UBI} < \tilde{y}^{UBI}$, which means less global sharing under UBI than under S^+.

Proof of Result 3. From (68) and (69), an interior solution for φ^{IC} is the value of φ such that $\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{it}] = \sum_t \delta^t E[u(y_{it}) + \gamma U_{it}]$. That is, if $\varphi^{IC} \in (0,1)$ then $\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{it}]$ given $\varphi = \varphi^{IC}$ exactly equals $\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{it}]$ given $\varphi = 1$, as the latter is, per definition, identical to $\sum_t \delta^t E[u(y_{it}) + \gamma U_{it}]$.

Since $u''(\cdot) < 0$ for all values of φ, $\sum_t \delta^t E[u(y'_{it}) + \gamma U'_{it}]$ can have only one global maximum. If that maximum is at $\varphi = 1$, then $\varphi^{IC} = 1$; if that maximum is at $\varphi = 0$, then $\varphi^{IC} = 0$; hence $\varphi^{IC} \in (0,1)$ implies a parameterization for which the maximum is interior: $\varphi^{IC} \in (0,1) \Rightarrow \varphi^* \in (0,1)$. Together with $u'(\cdot) > 0$, $u''(\cdot) < 0$ this directly implies $\varphi^* \in (\varphi^{IC}, 1)$, i.e., $\varphi^* > \varphi^{IC}$.

REFERENCES

Acemoglu, Daron, and Pascual Restrepo (2017a) "Low-Skill and High-Skill Automation" NBER Working Paper 24119.

Acemoglu, Daron, and Pascual Restrepo (2017b) "Robots and Jobs: Evidence from US Labor Markets" NBER Working Paper 23285.

Aghion, Philippe, Benjamin F. Jones, Charles I. Jones (2017) "Artificial Intelligence and Economic Growth" NBER Working Paper No. 23982.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman (2013) "Racing to the Precipice: a Model of Artificial Intelligence Development" Future of Humanity Institute Technical Report #2013-1.

Autor, David H., (2015a) "Why Are There Still So Many Jobs? The History and Future of Workplace Automation" *The Journal of Economic Perspectives* 29(3), 3-30.

Autor, David H., (2015b) "The Paradox of Abundance" in Subramanian Rangan (ed.) *Performance and Progress: Essays on Capitalism, Business, and Society*, (Oxford: Oxford University Press).

Autor, David H., David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen (2017) "Concentrating on the Fall of the Labor Share" *American Economic Review: Papers & Proceedings* 107(5), 180–185.

Baum, Seth D., (2017) "A Survey of Artificial General Intelligence Projects for Ethics, Risk and Policy," Global Catastrophic Risk Institute WP 17-1.

Baum, Seth D., Ben Goertzel, and Ted G. Goertzel (2011) "How Long Until Human-Level AI? Results from an Expert Assessment," *Technological Forecasting & Social Change*, 78(1), 185-195.

Berg, Andrew, Ed Buffie, and Felipe Zanna (2016) "Robots, Growth and Inequality" *Finance & Development* Sept. 2016, 10-13.

Berg, Andrew, Ed Buffie, and Felipe Zanna (2018) "Should We Fear the Robot Revolution? (The Correct Answer Is Yes)" IMF Working Paper 18/116.

Bostrom, Nick (2014) *Superintelligence: Paths, Dangers and Strategies* (Oxford: Oxford University Press).

Bostrom, Nick, and Eliezer Yudkowsky (2014) "The Ethics of Artificial Intelligence" in *The Cambridge Handbook of Artificial Intelligence* (Cambridge, UK: Cambridge University Press).

Bostrom, Nick (2017) "Strategic Implications of Openness in AI Development" *Global Policy* 8(2), 135-148.

Brynjolfsson, Erik, and Andrew McAfee (2014) *The Second Machine Age* (New York: W.W. Norton).

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson (2017) "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics" NBER Working Paper No. 24001.

Dawkins, Richard (2008) *The God Delusion* (New York: Mariner Books).

DeCanio, Stephen J., (2016) "Robots and Humans - Complements or Substitutes?" *Journal of Macroeconomics* 49, 280-291.

Ford, Martin (2015) *Rise of the Robots: Technology and the Threat of a Jobless Future*, (New York: Basic Books).

Freeman, Richard B., (2015) "Who Owns the Robots Rules the World" IZA World of Labor 2015:5.

Frey, Carl B., and Michael A. Osborne (2017), "The Future of Employment: How Susceptible Are Jobs To Computerisation?" *Technological Forecasting and Social Change* 114, 254-280.

Goldfarb, Avi, and Daniel Trefler (2017) "AI and International Trade", Forthcoming in Ajay Agrawal, Josh Gans, and Avi Goldfarb (eds.) *The Economics of AI*, NBER and University of Chicago Press.

Good, I.J., (1965) "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers*, June 1965, 6.

Gordon, Robert J., (2012) "Is U.S. Economic Growth Over? Faltering Innovation Confronts the Six Headwinds," NBER Working Paper 18315.

Grace, Katja, John Salvatier, Allan Dafoe, Babao Zhang, and Owain Evans (2017) "When Will AI Exceed Human Performance? Evidence from AI Experts" https://arxiv.org/pdf/1705.08807.pdf

Hawking, Stephen, Stuart Russell, Max Tegmark, and Frank Wilczek (2014) "Transcendence Looks at the Implications of Artificial Intelligence - But Are We Taking AI Seriously Enough?" The Independent, 01.05.2014.

Hémous, David, and Morten Olsen (2016) "The Rise of the Machines: Automation, Horizontal Innovation and Income Inequality" mimeo.

Korinek, Anton, and Joseph E. Stiglitz (2017) "Artificial Intelligence and Its Implications for Income Distribution and Unemployment" forthcoming in Ajay Agrawal, Josh Gans, and Avi Goldfarb (eds.) *The Economics of AI*, NBER and University of Chicago Press. http://www.nber.org/chapters/c14018

Kurzweil, Ray (2005) *The Singularity is Near: When Humans Transcend Biology* (New York: Viking).

Kurzweil, Ray (2013) *How to Create a Mind: The Secret of Human Thought Revealed* (New York: Penguin Books).

Muehlhauser, Luke, and Anna Salamon (2012) "Intelligence Explosion: Evidence and Import" in: Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart (eds.) *Singularity Hypotheses: A Scientific and Philosophical Assessment* (Berlin: Springer).

Muehlhauser, Luke, and Louie Helm (2012) "Intelligence Explosion and Machine Ethics" in: Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart (eds.) *Singularity Hypotheses: A Scientific and Philosophical Assessment* (Berlin: Springer).

Müller, Vincent C., and Nick Bostrom (2014) "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Berlin: Springer).

Nordhaus, William D., (2015) "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth" NBER Working Paper 21547.

Omohundro, Stephen M., (2008) "The Basic AI Drives" in Pei Wang, Ben Goertzel, and Stan Franklin (eds.) *Artificial General Intelligence: Proceedings of the First AGI Conference (Frontiers in Artificial Intelligence and Applications)* (Amsterdam: IOS Press).

Rodrik, Dani (2016a) "Innovation is Not Enough" Project Syndicate, June 9, 2016.

Rodrik, Dani (2016b) "Premature Deindustrialization" *Journal of Economic Growth* 12, 1-33.

Russell, Stuart, Daniel Dewey, and Max Tegmark (2015) "Research Priorities for Robust and Beneficial Artificial Intelligence" *AI Magazine* 36(4). https://arxiv.org/abs/1602.03506v1

Sachs, Jeffrey D., and Laurence J. Kotlikoff (2012) "Smart Machines and Long-Term Misery" NBER Working Paper 18629.

Simon, Herbert (1992) "Altruism and Economics" *Eastern Economic Journal* 18(1), 73-83.

Smith, Aaron, and Janna Andersen (2014) "AI, Robotics and the Future of Growth" Pew Research Center, http://www.pewinternet.org/2014/08/06/future-of-jobs/

Tegmark, Max (2014) *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality* (New York: Vintage).

Tegmark, Max (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf).

The White House (2016) “Artificial Intelligence, Automation, and the Economy” Technical Report.

Vinge, Vernor (1993) "The Coming Technological Singularity: How to Survive in the Post-Human Era", in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, G. A. Landis, ed., NASA Publication CP-10129.