DEVELOPMENT THROUGH SKILLS AND ALTRUISM

RACHID LAAJAJ
Universidad de los Andes

Latest version: available on the author's website.

The paper shows theoretically that, in a setting where skills expand what individuals can do and altruism affects the choice of actions, then skills increase and altruism together are sufficient conditions for development, defined as an increase in generated welfare. Greater skills need not increase social welfare, because such expansion combines a positive frontier effect with an ambiguous substitution effect, potentially toward actions that are privately attractive but socially harmful. This ambiguous effect of skills on development disappears when altruism is high enough to ensure that new opportunities created by skills are used only when they raise social welfare. Incorporating technological change and institutions shows that, when endogenous to humans' decisions, they become levers that can crystallize the altruism of some into widespread and durable gains. Finally, allowing altruism to be group-specific emphasizes the need for universalism to ensure that skills contribute to development, in particular when technologies allow for externalities that reach beyond one's inner group. The framework speaks to a broad set of historical and contemporary development successes and failures, leading to the conclusion that ensuring long-term development requires selecting altruistic leaders and learning how to shift the distribution and scope of altruism at scale.

KEYWORDS: Economic development, altruism, other-regarding preferences, technological change, institutions, universalism, social welfare.

JEL CLASSIFICATION. D64, O10, O33.

Rachid Laajaj: r.laajaj@uniandes.edu.co

I am grateful to David Bardey, Juan Camilo Cárdenas, Carolina Castro, Joshua Dean, Garance Genicot, Nathan Nunn, Debraj Ray, James Robinson, Nicolas de Roux, Álvaro Sandroni, María Alejandra Vélez, Chris Udry and Rakesh Vohra for valuable comments and suggestions. I also thank seminar participants at Chicago, Javeriana, Madison, Michigan, Northwestern, U de los Andes and VDEV/CEPR/BREAD seminar for helpful feedback. All errors are my own.

INTRODUCTION

Many fields largely attribute the success of the human species to humans' unique ability to cooperate at scale, itself shaped by cultural values and other-regarding preferences (Hume, Durkheim, HCHLH+, Batson). Economists have developed evidence of rational altruistic preferences (Andreoni and Miller) and explain how cooperation depends on "social emotions such as shame and guilt, and our capacity to internalize social norms so that acting ethically became a personal goal rather than simply a prudent way to avoid punishment" (Bowles and Gintis). Yet economic research rarely presents altruism and pro-social motivations as central drivers of development.

This paper fills this gap by providing a theoretical framework that puts skills and altruism at the heart of development, defined as an increase in generated social welfare. The core argument is straightforward: if skills increase the range of feasible actions and altruism ensures that agents choose actions beneficial to social welfare, then skills and altruism together are sufficient conditions for development. Altruism directs the use of technological progress: it determines whether new capabilities contribute to the common good or secure private gains at the expense of social welfare. The paper thereby offers a common framework for interpreting salient development challenges, including climate change, violent conflict, corruption, lobbying, persistent inequality, and poverty amid abundance. These challenges share a common structure: skills and technology are used in ways that benefit some individuals while imposing large costs on others. As reflected in the model, such outcomes persist only when some individuals' welfare receives greater weight than others'.

In the baseline model, at each period, a single decision maker selects her action from a feasible set of actions that generates payoffs for all individuals. We focus on two outcomes: the payoff that directly affects the decision maker and the effect on social welfare, estimated as the sum of all payoffs. Skills expand the set of actions the individual can implement. Altruism is captured by a parameter between zero and one that measures the relative weight placed on others' payoffs compared to one's own payoff, and therefore determines how the decision maker trades off private benefit against social welfare. The model delivers four main implications. First, greater altruism pushes choices toward higher social welfare. Second, the effect of higher skills on social welfare is ambiguous.We refer to this potentially ambiguous effect of skills as the *lottery of technology*: as skills rise and expand the set of

feasible actions, the welfare consequences depend on whether progress primarily facilitates actions that benefit society or creates new opportunities for private gain at others' expense. Third, for similar reasons, whether skills and altruism are complements depends on the direction of technological change. Fourth, the "guardrail theorem" states that when altruism is sufficiently strong, additional skills translate into higher social welfare. The model captures the fact that technological progress expands both win-win opportunities and the capacity to gain by imposing harm, and that altruism guarantees that actions beneficial for all will be prioritized.

The remainder of the paper addresses the most common limitations to the conclusion that skills and altruism are central to development. First, in dynamic settings, technological change is typically endogenous to choices and economic activity (Acemoglu, Aghion, Bursztyn, and Hemous). We find that once the direction of technological progress responds to behavior, then skills and altruism become complements: altruism not only affects contemporaneous choices, its distribution shapes the trajectory of technological development.

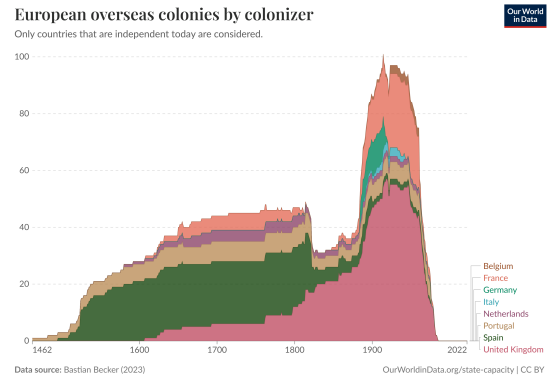
Second, as stated by Connolly, Loewenstein, and Chater, major policy issues require systemic solutions. People do not typically pay taxes out of pure altruism, they pay because institutions require it. Similarly, widespread environmental progress often requires regulation and coordinated action more than private initiatives that might come at a cost to competitiveness. One might therefore infer that development requires institutions that set the right incentives, rather than pure altruism. Yet the baseline model can be interpreted as including regulatory decisions among the actions individuals can take, highlighting that systemic solutions also result from individual choices and preferences. To make this interaction explicit, Section 3 shows that, if institutional decisions are themselves subject to human choices and interests, then, in the absence of altruism among policymakers, the effect of greater institutional power is subject to a *lottery of alignment of interests*: policymakers implement actions that benefit social workers only when doing so aligns with their own interests. By contrast, when policymakers' altruism is sufficiently high, more institutional power is necessarily beneficial. If the same model is used to consider the endogenous choice among types of institutions, then some *lottery of power struggle* might determine the choice of institutions and who is in power, unless sufficient altruism ensures the implementation of institutions, or the choice of leaders, that are most likely to benefit society as a whole. Even the creation and expansion of markets, often viewed as an alternative to

altruism, were championed by actors who explicitly argued for reconciling private and public gains for the benefit of society (Groenewegen, Smith), at a time when trade restrictions and publicly assigned monopolies were highly profitable for policymakers and their direct beneficiaries. In summary, rather than a substitute for altruism, institutions act as levers that crystallize altruism into durable conditions that support stable, large-scale cooperation.

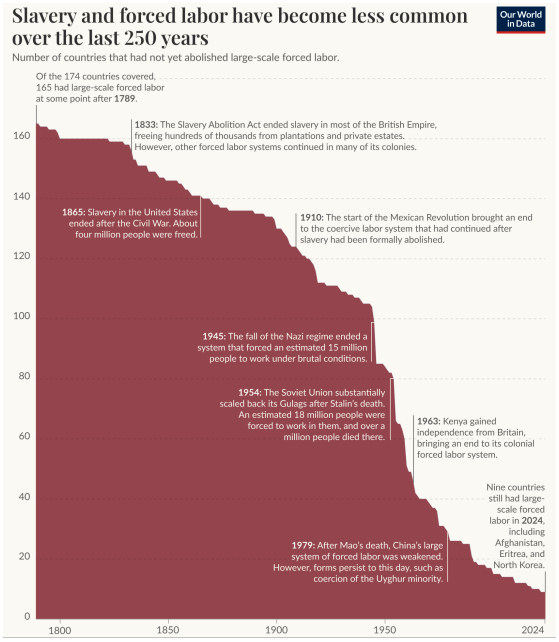
Third, when we distinguish altruism toward the inner circle from altruism toward the outer circle, we show that greater altruism toward one's inner circle can reduce social welfare. A lack of altruism creates a bias toward privately beneficial decisions, yet any non-uniform moral weights can also generate socially harmful choices, favoring some individuals at the expense of others. This highlights that universalism is the form of altruism that matters most for development, connecting to recent work emphasizing its importance for political and policy choices Enke, Rodríguez-Padilla, and Zimmermann, Cappelen, Enke, and Tungodden. In an extension of the model, the reach of externalities expands with technological progress, while altruism toward newly affected individuals takes time to catch up. This creates a window of vulnerability in which society has the capacity to harm some populations, but not yet the breadth of concern required to refrain from actions that are particularly damaging to the common good. This sub-model provides a useful lens for thinking about the rise and fall of dramatic historical episodes, including slavery and colonization.

We also provide a non-formal discussion of additional concerns. First, in some situations, cooperative outcomes can be sustained even among fully self-interested players. Second, we ask whether the model is consistent with evidence that appears to show negative consequences of solidarity, such as the "kin tax" (Jakiela and Ozier), or with what Acemoglu and Robinson describe as a "cage of norms". Third, we acknowledge that, in practice, altruism can take various forms, from unconditional cooperation to warm glow (Andreoni), conditional cooperators and altruistic punishers. These alternative forms are more common than pure altruism and reflect the joint determination of preferences and norms (**?**Ligon and Schechter). Finally, we discuss whether altruism is sufficiently diverse and malleable to be a meaningful policy margin.

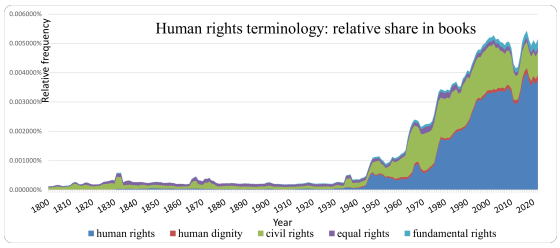
Our broad definition of development as the generation of social welfare, regardless of who is affected, and including economic and non-economic drivers of welfare, is key both to the coherence of the theoretical model and to allow the theory to apply to a broader set of phenomena. To take a specific example, Figure 1 documents four related patterns. Panel A



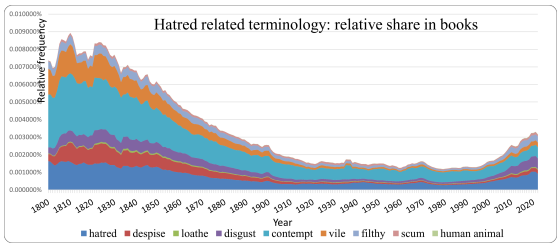
Panel A. European overseas colonies by colonizer



Panel B. Slavery and forced labor



Panel C. Relative share of human-rights terminology in books



Panel D. Relative share of hatred terminology in books

FIGURE 1.—Historical patterns in colonial expansion, forced labor, and the language of rights and hatred.

shows the expansion and contraction of European overseas colonies through the nineteenth century, an expansion made feasible by technological progress (**?**). Panel B shows the progressive retreat of slavery and forced labor. Panels C and D show corresponding changes in discourse: the relative frequency of human-rights terminology rises sharply from the nineteenth century, while hatred terminology declines more gradually. These patterns are consistent with the view that development depends on the interplay between technological progress and the values and motives that direct its use.

Of course, these historical changes reflected many forces beyond altruism, including private interests, strategic conflict, and shifts in political power. Yet the rhetoric and observable sacrifices of the leaders and followers of the movements that produced these changes often

reveal motivations that went beyond self-interest.[1] Beyond these leaders, large-scale social change required mass mobilization. Participants had to overcome the standard free-rider problem in collective action and often faced substantial risks to generate a public good for oppressed groups. During the French Revolution, the Jacobin Club's motto, "Live free or die," illustrates the willingness to incur such costs, and violently repressed protests remain part of contemporary political life.

One possible explanation for worldwide improvements in social welfare is that values changed and practices of the past became morally and politically unacceptable. The mid-nineteenth century marked an important turning point in the international recognition of human rights and associated values as shown in Panel C and Henkin. Such large-scale changes in discourse, and perhaps in values, can embolden movements that contest injustice and, increase the moral and social cost of maintaining oppression. At the same time, the evidence also suggests a recent reversal. Hatred terminology (Figure 1, Panel D) and invasion-or-rejection terminology (Appendix Figure 7, Panel A) have increased since the beginning of the twenty-first century, with the latter reaching levels not observed since before World War II. These patterns coincide with a recent increase in deaths from armed conflict (Appendix Figure 7, Panel B).

This work builds on research in economics that studies the drivers of cooperation (**?**HBBC+), documents the existence and importance of altruism (Andreoni and Miller, Bowles and Gintis, Ashraf and Bandiera), and highlights the role of prosocial motivations in public service, dedication at work, and interpersonal transfers (Bénabou and Tirole, Ashraf, Bandiera, and Jack, Bourlès, Bramoullé, and Perez-Richet, Besley and Ghatak, Ashraf, Bandiera, Davenport, and Lee). Particularly relevant to this paper is the recent evidence that universalism is a common factor behind multiple choices that affect development, ranging from inequality and migration to the environment (Enke, Enke, Rodríguez-Padilla, and Zimmermann, Cappelen, Enke, and Tungodden). These advances provide the

[1] For example, Gandhi claimed that "the good of the individual is contained in the good of all"; Martin Luther King Jr. argued that one had not truly lived until rising above "individualistic concerns" to the "broader concerns of all humanity"; Kate Sheppard, who led the first successful national movement for women's voting rights, maintained that "all that separates, whether of race, class, creed, or sex, is inhuman, and must be overcome"; and Nelson Mandela, who spent 27 years in prison, later suggested that the significance of life lies in "what difference we have made to the lives of others."

basis for considering the broader role of altruism in development. This paper extends this research by providing a micro-founded theoretical environment that where altruism affects development by shaping the direction in which skills, technology, and power are used. By doing so, it connects to research aiming to understanding the sources of development, including Guiso, Sapienza, and Zingales, who emphasize civic capital as a key determinant of development and macro-historical approaches that seek common patterns and explanations for large-scale development phenomena (**?**Acemoglu and Robinson).

The paper has two main policy implications. First, ensuring long-term development requires shifting the distribution of altruism. Most societies have invested heavily in skills and technology, while investments in strengthening altruism, and in learning systematically how to do so, merit more attention. This connects to a growing body of research on endogenous preferences and values and on the co-evolution of preferences, norms, and institutions (Bisin and Verdier, Besley). Second, the results highlight the amplified role of a small number of individuals with high political power or influence on technological progress, connecting to recent research on the importance of selecting prosocial agents (Ashraf, Bandiera, Davenport, and Lee, Besley and Ghatak) or directly affecting the preferences of leaders (Mehmood, Naseer, and Chen).

1. BASELINE MODEL

1.1. *Non-technical summary and applications*

We study a setting with N identical agents in which, at the beginning of each one of the K periods, a single decision maker is randomly chosen to play. This agent selects an action that yields to potential contributions to the payoffs of each individual: $(u_1, u_2, ..., u_N)$. We focus on the private contribution to one's own payoff u_i and an aggregate contribution to the payoffs $U = \sum_{j=1}^{N} u_j$. Feasible pairs (u_i, U) form a strictly convex set $R(s)$ that expands in all directions with skills s (Figure 2, Panel A). Given agent i's altruism $a \in [0, 1]$, she selects her preferred action within the feasible set, which is the one that maximizes her valuation function:

$$v_i \;=\; (1-a)\,u_i + a\,U.$$

Thus skills determine what is achievable and altruism sets how the trade-off between private and aggregate payoffs is evaluated. Since rounds are independent, solving the K-

round problem amounts to solving the single-agent problem for U^* and setting aggregate social welfare to $W = K \times U^*$. In this baseline model, the agent's decision does not affect other agents' preferences or payoff structure, eliminating strategic interaction. Sections 2 and 3 introduce specific forms of strategic interactions.

As shown in Figure 2, Panel A, the unique solution must be located in the upper-right border of the set $R(s)$ and at the unique point that satisfies the tangency that equalizes the marginal rate of technical substitution between u_i and U, given by the slope at the frontier of $R(s)$ and the ratio of their relative valuations from the perspective of player i: $\frac{1-a}{a}$. If for example $a_i = 0.2$, then agent i is willing to trade one unit of u_i for four units of U, hence the solution requires being at the frontier of the possibility set that satisfies this tradeoff.

The first theorem states that more altruism unambiguously raises social welfare, but the effect of skills and the complementarity between skills and altruism cannot be signed. Graphically, increasing a flattens the indifference slope, rotating the choice counterclockwise along the frontier toward outcomes with higher U. The ambiguous effect of skills on U occurs because an increase in s induces (i) a positive *expansion effect*, the frontier moves outward, allowing a simultaneous increase of u_i and U, maintaining the angle fixed, and (ii) a *substitution effect*: the optimal point shifts along the new frontier in a direction that depends on whether U or u_i became relatively easier to reach. Depending on the direction in which the frontier expands faster, substitution can reinforce, reduce or offset the expansion effect, so dU/ds needs not be positive. Panels B and C of Figure 2 illustrate how the effect on welfare of an increase in skills can go in both directions. For similar reasons, the effect of an increase in skills on the complementarity between skills and altruism cannot be signed without adding restrictions to how $R(s)$ expands with s.

This model refers to technological progress as the *crystallization* over time of progress manifested in the durable expansion of $R(s)$ allowed by skills. We refer to this ambiguous effect of skills on development as the *lottery of the technology*: when skills rise, does the expansion of $R(s)$ make it relatively easier to raise u_i in a way that contributes to social welfare as in Figure 2, Panel B, or at the expense of social welfare, as in Figure 2, Panel C?

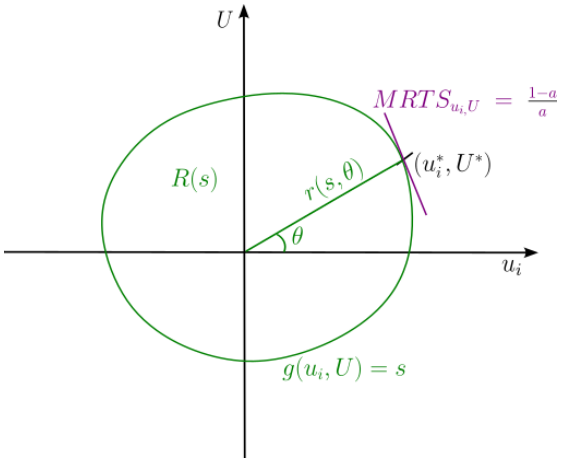
The "guardrail theorem" is the central result behind our statement that skills and altruism are sufficient conditions for development. It shows that, if altruism is high enough, then the effect of an increase in skills on social welfare must be positive. Hence a society with individuals that are sufficiently altruistic can only benefit from the increase in s by benefiting

from the expansion effect allowed by technological progress while refraining from the use of actions brought by technological progress that benefit some at the cost of social welfare. The minimum value of a required does, however, depend on the *lottery of the technology*, meaning the shape and expansion of $R(s)$.

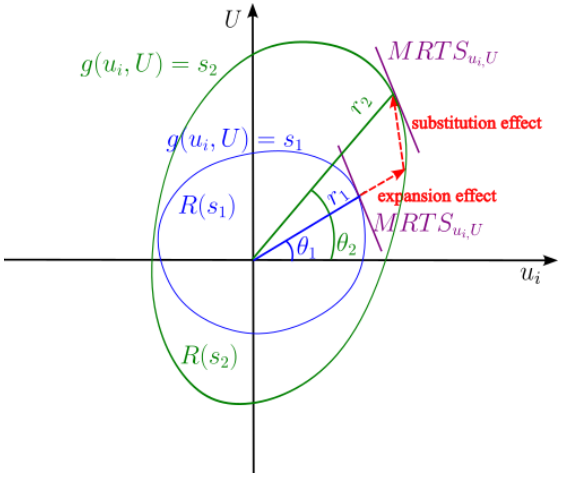
To connect the results to real life interpretations, if technological progress (which, in the model results from skills) drastically reduces the price of a clean energy source, then even an individualistic agent i would move to a cleaner energy and contribute to social welfare, as in Figure 2, Panel B. In this case, skills not only expand the frontier, but also act as a substitute for altruism. By contrast if technological progress lowers the cost of a technology that contaminates more than the initial one, this technological progress allows an increase in private gains at the cost of social welfare and altruism becomes more necessary to renounce such actions with high private payoffs but strong negative externalities. The question of whether the expansion of feasible actions resulting from skills and technological improvement tends to increase the size of the common cake or make it easier to take away from it illustrates a dilemma that is behind many of the most relevant development issues. Will technological progress primarily tend to facilitate more corruption and piracy or strenghten transparency? Will digital tools empower their users or increase opportunities to monetize attention at the cost of their users' wellbeing? Without public supervision, will AI facilitate power concentration, unemployment and existential risk or open a path towards sustainable abundance? The response to these questions drive the "lottery of the technology" and thus the extent to which development will directly result from skills and technological progress or whether altruism is required to avoid falling in the new pitfalls opened by technological progress.

1.2. *The set of achievable actions in Cartesian and Polar coordinates*

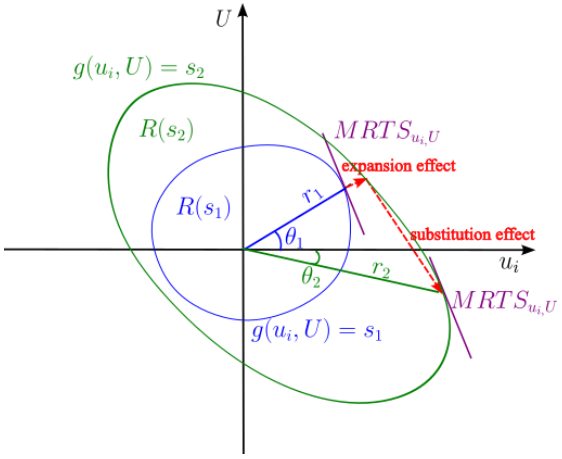
We consider a sequential decision-making environment with N identical individuals. In each of K periods, a randomly selected individual i executes a single action. In this baseline version, there is no strategic interaction since (a) a player's action does not affect other players' payoff structure, (b) payoffs are additive, and (c) agents do not reciprocate, and their preferences or choices are not conditional on others' actions.



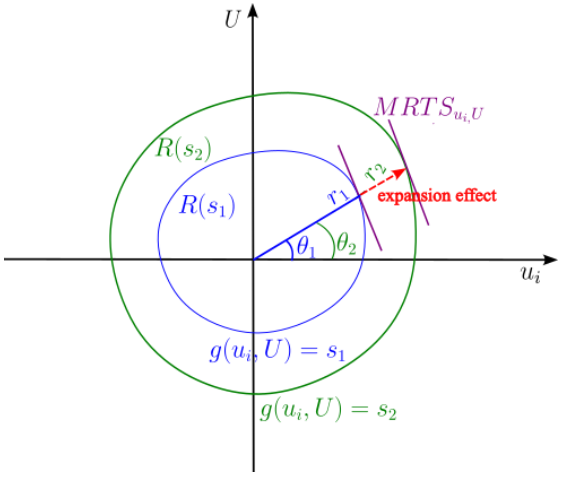
Panel A. The set of achievable actions in Cartesian and polar coordinates



Panel B. Skill increase with a positive substitution effect on U



Panel C. Skill increase with a negative substitution effect on U



Panel D. Skill increase in the absence of technological bias

FIGURE 2.—Representation of the feasible set and the agent's choice in the (u_i, U) space.

To each action, a profile of contributions to payoffs $\left(u_1, \ldots, u_N\right)$ is associated, of which two parameters will drive agent i's decision: her own payoff u_i and total payoff $U = \sum_{j=1}^{N} u_j$. As shown in Figure 2, Panel A, the set $R(s)$ determined by skill level s represents all the pairs $\left(u_i, U\right)$ associated with feasible actions:

$$R(s) = \left\{ (u_i, U) \mid g(u_i, U) \leq s \right\}, \tag{1.1}$$

where $R(s)$ is a strictly convex and compact set, and the skill requirement function $g(u_i, U)$ gives the minimum skill required to reach the set of utilities (u_i, U). The boundary, defined

by $g(u_i, U) = s$, represents the production possibility frontier in terms of direct utilities resulting from feasible actions.

We will also often use polar coordinates, given their convenience and meaningful interpretation in this context. In a two-dimensional space, polar coordinates define the location of any point using its angle (θ), and its distance from the origin (r).

We also define $r(\theta, s)$ as the distance between the origin and the limit of $R(s)$ in the direction θ, thus $r = r(\theta, s)$ if and only if $g\big(u_i(\theta, r), U(\theta, r)\big) = s$. Hence we can also define $R(s)$ in polar coordinates as

$$R(s) \;=\; \big\{\, (\theta, r) \;\mid\; r \;\leq\; r(\theta, s) \big\},$$

and the fact that an increase in s expands $R(s)$ at every point of the contour curve can be expressed as $\frac{dr(\theta,s)}{ds} > 0 \;\forall\theta, \forall s$.

1.3. *Preferences and decisions*

Each player puts a weight of 1 on her own payoff and a on others, where $a \in [0, 1]$ is the altruism parameter. Hence a payoff profile $\{u_1, \ldots, u_N\}$, gives the following value function, maximized by agent i:

$$v_i\big(u_1, \ldots, u_N\big) \;= u_i \;+\; a\, U^{-i} \;= (1 - a)\, u_i \;+\; a\, U = \qquad \text{where } U = \sum_{j=1}^{N} u_j \tag{1.2}$$

In the special case where $a = 1$, the agent cares only about the total payoff U, with no additional relative weighting on her own payoff. In the other extreme $a = 0$, the agent cares only about u_i. Because her payoff depends only on $\{u_i, U\}$, without any reduction in relevant information we collapse $\big(u_1, \ldots, u_N\big)$ into the sufficient statistics $\big(u_i, U\big)$ that characterize any possible action. This model is conceptually a reduced form that directly connects actions to their contribution to final payoffs, including effects through effort, consumption and other channels, except other-regarding preferences captured through a.

1.4. *Description of the solution*

With no intertemporal linkage in this baseline model, the environment is repeated for K independent, identical rounds. Hence if a representative agent chooses payoffs (u_i^*, U^*), social welfare after K rounds is $W = K \times U^*$.

In each round, player i solves:

$$\max_{(u_i,U)\in R(s)} v_i = (1-a)\,u_i \;+\; a\,U \tag{1.3}$$

Because v_i is increasing in u_i and U, the solution must lie in the upper-right portion of the boundary, and since $R(s)$ is strictly convex, the frontier $g\,(u_i, U) = s$ is downward-sloping within this upper-right segment, where we can define the locally decreasing function $U\big(u_i, s\big)$, located on th g function, as the maximum achievable U when player i with skills s reaches private utility u_i.[2]

Using standard microeconomics terminology, we refer to the slope of $U\,(u_i, \overline{s})$ as the marginal rate of technical substitution: [3]

$$MRTS_{u_i,U} \equiv -\frac{dU(u_i, s)}{du_i} \tag{1.4}$$

In this model, the $MRTS_{u_i,U}$ measures how much U must be reduced to gain one more unit of u_i, when fixing s and remaining at the frontier. After replacing $U = U\big(u_i, s\big)$ into the maximization problem (1.3), the first-order condition with respect to u_i is given by:

$$\frac{dv_i}{du_i} = (1-a) \;+\; a\,\frac{dU(u_i, s)}{du_i} \;=\; 0 \quad \Longrightarrow \quad MRTS_{u_i,U} \;=\; \frac{1-a}{a} \tag{1.5}$$

Hence, the unique tangency where the slope of the frontier $U(u_i, s)$ is equal to $\frac{1-a}{a}$ determines the optimal pair $\left(u_i^*, U^*\right)$ that maximizes v_i subject to $(u_i, U) \in R(s)$. The solution exists and is unique because within the range of the upper right contour, the $MRTS_{u_i,U}$ spans $[0, \infty)$ and is strictly increasing in u_i.

We can also connect this definition to the polar coordinates notation by making the Marginal Rate of Substitution a function of θ: $MRTS_{u_i,U}\,(\theta, s)$ is the absolute value of

---

[2] $U = U\big(u_i, s\big)$ must satisfy $g\big(u_i, U(u_i, s)\big) = s$ and is only locally defined. The implicit function theorem ensures the local existence of $U\,(u_i, \overline{s})$ in the upper right set of the curve between its horizontal tangent $\frac{dU(u_i,s)}{du_i} = 0$ and its vertical tangent $\frac{dU(u_i,s)}{du_i} = -\infty$.

[3] Standard microeconomics defines the $MRTS$ as the rate at which a firm can replace one input with another while keeping output constant, in this context, it is the rate at which U can be traded with u_i, keeping skills constant.

the slope of $U\left(u_i, s\right)$ when reaching $g\left(u_i, U\right) = s$ in the direction of θ. Hence the solution can also be written as $MRTS_{u_i,U}\left(\theta, s\right) = \frac{1-a}{a}$, emphasizing that for any skill level s, the solution is located at the θ angle that satisfies $MRTS_{u_i,U} = \frac{1-a}{a}$. Since $\frac{dU(u_i,s)}{du_i} < 0$ within the range $u_i \in \tilde{U}$, an increase in θ while remaining on $g\left(u_i, U\right) = s$ is equivalent to trading more U at the expense of u_i. We use the notations $\frac{dU}{d\theta}\Big|_{g(u_i,U)=s}$ and $\frac{du_i}{d\theta}\Big|_{g(u_i,U)=s}$ to refer to changes in U and u_i under a counterclockwise move along $g\left(u_i, U\right) = s$. This implicitly assumes that $r(\theta, s)$ adjusts to remain on the corresponding g curve. To be more specific: $\frac{dU}{d\theta}\Big|_{g(u_i,U)=s} = \frac{\partial U}{\partial \theta} + \frac{dU}{dr} \cdot \frac{dr(\theta,s)}{d\theta}$.

It may also help the intuition to write $MRTS_{u_i,U}\left(\theta, s\right) = -\frac{\frac{dU}{d\theta}}{\frac{du_i}{d\theta}}\Big|_{g(u_i,U)=s}$: the Marginal Rate of Substitution is the ratio of the number of additional units of U over the number of units of u_i lost when the player increases θ. Hence the solution of equation (1.5) can also be written as

$$MRTS_{u_i,U}\left(\theta, s\right) = -\frac{\dfrac{dU}{d\theta}}{\dfrac{du_i}{d\theta}}\Big|_{g(u_i,U)=s}; = \frac{1-a}{a} = \frac{dv_i/du_i}{dv_i/dU} \tag{1.6}$$

This equation tells us that the ratio of marginal increases of U and u_i when θ increases should be equal to the inverted ratio of their marginal valuations according to the preferences of the player. Equivalently, equation (1.6) can be rewritten as follows:

$$-\frac{dv_i}{dU} \cdot \frac{dU}{d\theta}\Big|_{g(u_i,U)=s} = \frac{dv_i}{du_i} \cdot \frac{du_i}{d\theta}\Big|_{g(u_i,U)=s} \tag{1.7}$$

Equation (1.7) describes the trade-off: the gain in utility obtained from the increase in U resulting from a marginal increase in θ should be equal to the loss in utility caused by the reduction in u_i resulting from the same marginal increase in θ. In other words, at the solution point, player i cannot increase v_i by either increasing or reducing θ.

1.5. *Comparative Statics and propositions*

The first proposition addresses the policy-relevant questions about the effect on U of an increase in a, an increase in s and their cross-derivative. They are then discussed in the following three subsections.

PROPOSITION 1.1: *In the previously described baseline setting,* $\frac{dU}{da} > 0$*, while* $\frac{dU}{ds}$ *and* $\frac{dU^2}{dads}$ *cannot be signed without further assumptions.*

1.5.1. *The effect of an increase in altruism on social welfare:* $\frac{dU}{da}$

Building on the description of the solution in the previous subsection, this effect is quite straightforward. The solution requires the tangency $MRTS_{u_i,U}(\theta, s) = \frac{1-a}{a}$. An increase in a reduces $\frac{1-a}{a}$, meaning the slope becomes flatter. The convexity of the set $R(s)$ ensures that $\frac{dMRTS_{u_i,U}(\theta,s)}{d\theta} < 0$, so when a increases, the slope $MRTS_{u_i,U}$ decreases, requiring θ to increase in order to restore the equality $MRTS_{u_i,U} = \frac{1-a}{a}$ and maintain the tangency. As shown in subsection 1.4 an increase in θ along the g curve results in higher U and lower u_i.

1.5.2. The effect of an increase in skills on social welfare: $\frac{dU}{ds}$

As s increases, the feasible set $R(s)$ expands at all points: $\frac{dr(\theta,s)}{ds} > 0 \forall \theta, s$, which affects U through an expansion effect and a reallocation effect:

$$\frac{dU}{ds} = \underbrace{\frac{dU}{dr} \cdot \frac{\partial r(\theta, s)}{\partial s}}_{Expansion\ effect} + \underbrace{\left.\frac{dU}{d\theta}\right|_{g(u_i,U)=s} \cdot \frac{d\theta}{dMRTS} \cdot \frac{dMRTS(\theta, s)}{ds}}_{Reallocation\ effect} \tag{1.8}$$

To facilitate the intuition, Panels B and C of Figure 2, together with this paragraph, describe a non-marginal change from s_1 to s_2. In the expansion effect, $\frac{\partial r(\theta,s)}{\partial s} > 0$ describes how, holding θ_1 fixed, the length r increases until it reaches $g(u_i, U) = s_2$. As shown in Figure 2, Panel B, the expansion effect can be represented by a move from (θ_1, r_1) to the e point $(\theta_1, r_e))$, where $r_e = r(\theta_1, s_2)$. To obtain the expansion effect on U, this change in r is multiplied by $\frac{dU}{dr}$ which is positive when $\theta > 0$.[4]

Second, the reallocation effect is graphically represented by the move along the $g\big(u_i, U\big) = s_2$ curve from point e, with polar coordinates $(\theta_1, r(\theta_1, s_2)))$ to the new

[4]By definition of polar coordinates, $U = r \sin\theta$, so $\frac{dU}{dr} = \sin\theta$ which is positive in the upper-right quadrant, i.e. when $0 < \theta < \frac{\pi}{2}$. θ might be outside of this range in some scenarios (e.g. when $U^* < 0$), but this subsection's conclusion that $\frac{dU}{ds}$ cannot be signed does not required assuming $\frac{dU}{dr} > 0$.

equilibrium $(\theta_2, r_2) = (\theta^*(s_2, a), r(\theta_2, s_2))$. $\frac{dU}{d\theta}\Big|_{dg(u_i,U)=0} > 0$ reflects the positive effect on U of an increase in θ with $r(\theta, s_2)$ adjusting to remain on $g(u_i, U) = s_2$ while moving counterclockwise along the g curve. Reallocation occurs when, at point e, the $MRTS_{u_i,U}(\theta_1, s_2) \neq \frac{1-a}{a}$. The direction of the move along g depends on the technological bias: $\frac{dMRTS(\theta,s)}{ds} > 0$ if U becomes easier to achieve relative to u_i as s increases and $\frac{dMRTS(\theta,s)}{ds} < 0$ if u_i becomes relatively easier to achieve as s increases. $\frac{d\theta}{dMRTS}$ reflects the extent to which θ needs to adjust until the tangency $MRTS_{u_i,U}(\theta, s) = \frac{1-a}{a}$ is restored. As shown in subsection 1.5.1: $\frac{d\theta}{dMRTS} < 0$.

To sign the total effect on U, the expansion effect is positive, since both of its terms are positive, and the sign of the reallocation effect $\frac{dMRTS(\theta,s)}{ds}$ is given by the sign of $\frac{dMRTS(\theta,s)}{ds}$. The absence of restriction on how $R(s)$ expands with s allows the reallocation effect to go in any direction with the possibility that it offsets the expansion effect, in which case the total effect on U of an increase in s becomes negative. Panels B and C of Figure 2 display a case where $\frac{dU}{ds} > 0$ and a case where $\frac{dU}{ds} < 0$, respectively, thus showing graphically that, without more assumptions, $\frac{dU}{ds}$ cannot be signed. The direction of $\frac{dU}{ds}$ depends on the shape of the irregular cone defined by $R(s)$.[5] This is what we refer to as the "lottery of the technology" mentioned in subsection 1.1.

1.5.3. The direction of the cross-derivative: $\frac{dU^2}{dads}$

We know from subsection 1.5.1 that a negatively affects the slope $\frac{1-a}{a}$, which in turn pushes a counterclockwise reallocation along the g curve, toward more U and less u_i. The sign of the cross-derivative $\frac{dU^2}{dads}$ would tell whether this positive effect tends to increase or decrease as s increases. We show in Appendix A.1 that:

$$\frac{dU}{da} = \frac{dU}{dMRTS_{u_i,U}} \frac{dMRTS_{u_i,U}}{da} = \frac{U^{u_i}(u_i, s)}{U^{u_i u_i}(u_i, s)} \frac{1}{a^2} > 0 \tag{1.9}$$

Hence the extent to which a given increase in a translates into an increase in U is increasing in the slope U^{u_i} and decreasing in the local curvature of the frontier $U^{u_i u_i}$. Intuitively,

[5] The "irregular cone, defined over (u_i, U, s) refers to the cone that can be slided horizontally to obtain $R(s)$, the set of feasible payoffs for a given level of skills and the agent moves up the cone as s increases

the less concave the g curve is at that point, the more movement along the g curve is needed to restore the tangency $MRTS_{u_i,U} = \frac{1-a}{a}$. Hence the effect of an increase in s on $\frac{dU}{da}$ depends on how it affects U^{u_i} and $U^{u_i u_i}$. Given that we allow for a flexible irregular cone, these changes could both go in any direction and thus the sign of the cross-derivative $\frac{dU^2}{dads}$ cannot be determined. A more formal proof is provided in Appendix A.1.

1.5.4. *The guardrail theorem: $\frac{dU}{ds}$ is positive for a high enough*

The previous subsections show that, without further assumptions on the structure of the flexible irregular cone that represents $R(s)$ $\forall s$ we cannot say much about the effect of an increase in s on welfare, nor about the complementarity between skills and altruism. This raises the question about what we can tell regarding how these two variables relate and jointly contribute to development, which is answered by Proposition 1.2, the core result that justifies the title of this paper.

PROPOSITION 1.2: *In the baseline setting described, if a is high enough, then $\frac{dU}{ds} > 0$.*

The proof of Proposition 1.2 is quite straightforward: if $a = 1$, then $\frac{dU}{ds} > 0$. Indeed, when $a = 1$, the player maximizes U, so a strict expansion of $R(s)$ can only benefit U.

Additionally one can see from Equation (1.9) that $\frac{dU}{ds}$ is continuous and differentiable in a (since $U(u_i, s)$ is infinitely differentiable). Hence, for any irregular cone $R(s)$, there exists ε small enough that for any $a \in [1-\varepsilon, 1]$, $\frac{dU}{ds} > 0$.

1.6. *The Proportional Cone Model*

To help the reader develop intuition for the model, we present the a variant of the baseline model adding one restriction: $r(\theta, s) = s\ r(\theta, 1)$. As a consequence, the shape of $R(s)$ remains the same but is simply scaled up as skills increase, making the cone more regular, in the sense that it still allows for any concave initial shape $R(1)$, but then requires this shape to be maintained. In this special case of the previous general model, the demonstration that $\frac{dU}{da} > 0$ remains valid. The structure imposed on $R(s)$ greatly facilitates the ability to sign $\frac{dU}{ds}$ and $\frac{dU^2}{dsda}$.

As a consequence of the fixed shape of $R(s)$, the slope of the tangency at angle θ is now independent of s: $\frac{dMRTS(\theta,s)}{ds} = 0$. This implies that the angle θ that satisfies the equilibrium condition $MRTS_{u_i,U}(\theta, s) = \frac{1-a}{a}$ is independent of s. Thus the reallocation effect of

equation (1.8) disappears and $U(a,s) = s\,U(a,1)$. $\frac{dU}{ds} > 0$ if and only if $U(a,1) > 0$ hence it only requires that the initial level of U is positive so that a technologically unbiased scaling up with s implies a proportional increase in U.

Regarding the complementarity between a and s, since we know that $\frac{dU}{ds} = U(a,1)$ and $\frac{dU(a,1)}{da} > 0$, then $\frac{dU^2}{dsda} > 0\ \forall R(1)$. Intuitively, in the absence of the "lottery of the technology", s and a are indeed complementary since a shifts the choice of θ toward higher values of θ and the increase in skills extends $r(\theta, s)$, the distance from the origin.

2. ENDOGENOUS TECHNOLOGICAL CHANGE AND DYNAMIC MODEL

2.1. *Non-technical summary of lessons, with endogenous technological change*

In the baseline model, the "lottery of technology" means that some technologies may be easier to develop. In practice, however, the direction of technological progress is also shaped by agents' choices: Acemoglu, Aghion, Bursztyn, and Hemous show that policies can steer innovation toward cleaner technologies. This extension lets technical progress arise from the choices of agents who share a common technology but differ in altruism. The variable s captures both skills and technological progress, evolving linearly over time. At any given point in time, a continuum of agents select actions. Through learning by doing, the frontier of feasible outcomes extends in the directions chosen.[6] More precisely, the expansion of $R(s)$ at any point of its frontier is proportional to the density of agents who selected that action. Figure 3 displays how the distribution of directions θ chosen by agents translates into expansion of the feasible set.

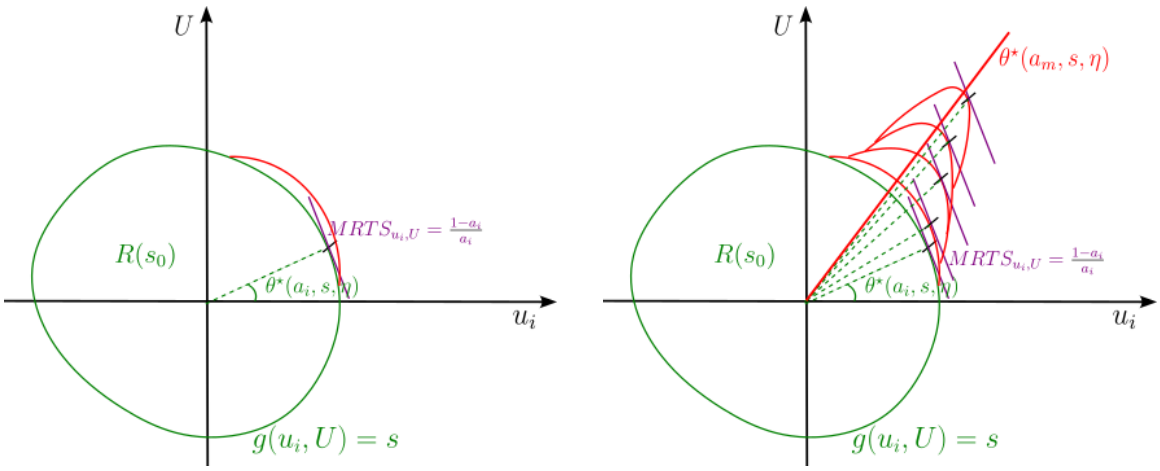
Because altruism is single-peaked, the direction initially chosen by the modal agent a_m, $\theta^*(a_m)$, is pulled ahead fastest and becomes increasingly attractive. This self-reinforcing dynamic makes all agents converge toward that direction. In the short-term, altruism combines an immediate welfare effect with a durable effect on the direction of technological progress. As a result, skills and altruism become complementary, and higher skills benefit society only when a_m is high enough.

Reality probably lies between the "lottery of technology" and this fully endogenous version, which generates extreme concentration in the chosen direction. This extension illus-

[6]The setup is consistent with learning by doing, and with intentional R&D investments if each agent orients her R&D in the same direction as her action.

trates that endogenizing technological progress reinforces the role of altruism: technological progress becomes a lever through which altruism affects societal outcomes. Because of path dependence, early moves can have outsized and durable effects, making later corrections more costly.

2.2. *Model setup with endogenous technological change*



Panel A. Learning by doing and expansion of $R(s)$ proportional to agents' choices

Panel B. Long-term dynamics under endogenous technological change

FIGURE 3.—Model with endogenous technological change.

A unit mass of agents indexed by $i \in [0, 1]$, each with altruism parameter $a_i \in [0, 1]$, is drawn from a fixed distribution with probability density function $\eta(a)$, concave over its domain, with a single peak reached at its mode a_m.[7] The initial achievable set $R(s_0)$, value function and the MRTS-based characterization of the static optimum remain identical to those in section 1. However, in this version learning by doing drives the expansion of the feasibility set: in any direction θ, for any level of skills s, the increase in the distance from the origin $r(\theta, s, \eta)$ is proportional to $\rho(\theta, s, \eta)$, the density of agents who choose this $\theta^\star(a_i, s, \eta)$ as

$$\frac{dr(\theta, s, \eta)}{ds} = \rho(\theta, s, \eta) \tag{2.1}$$

[7] Concavity is not strictly necessary but is a sufficient condition to ensure the local concavity of the function $U(u_i, s, \eta)$ as it expands and thus to ensure the uniqueness of the solution.

The shape of $R(s,\eta)$ is thus a function of the distribution η since it is endogenously determined by the choices $\theta^\star(a_i,s,\eta)$, which themselves depend on each agent's level of altruism:

$$R(s,\eta) \;=\; \big\{\,(\theta,r)\;\mid\; r \;\leq\; r(\theta,s,\eta)\big\},$$

where $r(\theta,s=0,\eta)$ is fixed by $R(s_0)$ and at any other time, the distance to the origin of the feasible set $r(\theta,s=0,\eta)$ results from its endogenous expansion over time.

In a continuously iterative process, over time, s grows linearly, $R(s,\eta)$ expands as described above, hence each agent i continuously re-adjusts her choice $\theta^\star(a_i,s,\eta)$, and these choices continuously shape the expansion of $R(s,\eta)$. In this context, we are interested in how skills and altruism affect the evolution of $W(s,\eta)\equiv\int_{i=0}^{i=1}U(a_i,s,\eta)\;di$, the social welfare at skills s resulting from all agent's decision at that skill level.

2.3. *Characterization of the solution with endogenous technological change*

Building on subsection 1.4, let the angle $\theta^\star(a_i,s,\eta)$ denote the unique solution chosen by an agent with altruism a_i, characterized by the tangency condition

$$\text{MRTS}_{u_i,U}\big(\theta^\star(a_i,s,\eta),s,\eta\big)=\frac{1-a_i}{a_i}. \tag{2.2}$$

As shown in subsection 1.5.1, within the upper-right section of the g curve, where all solutions are located, $\frac{\partial\text{MRTS}}{\partial\theta}<0$. Hence higher altruism maps into a higher θ^*:

$$\frac{d\theta^\star}{da_i}=\frac{1/a_i^2}{-\partial\text{MRTS}/\partial\theta}>0. \tag{2.3}$$

This monotonicity allows us to define the inverse function $a(\theta,s,\eta)$, the level of altruism that leads an agent to choose direction θ for given (s,η). Since $\frac{\partial a(\theta,s,\eta)}{\partial\theta}>0$, the density of agents choosing each direction is the density associated to only level of altruism that leads to this choice of θ^*:

$$\rho(\theta,s,\eta)=\eta\big(a(\theta,s,\eta)\big). \tag{2.4}$$

Combining this with the endogenous expansion rule in equation (2.1) gives

$$r_s(\theta,s,\eta)=\eta\big(a(\theta,s,\eta)\big). \tag{2.5}$$

Thus the distribution of altruism determines where the feasible frontier expands most. Because η is single-peaked at a_m and $a_\theta > 0$, $r_s(\theta, s, \eta)$ is single-peaked at the modal direction $\theta_m \equiv \theta^\star(a_m, s_0, \eta)$. This is the direction chosen by agents with the modal level of altruism a_m, and therefore the direction with the fastest learning-by-doing expansion.

LEMMA 2.1: *Fixing $R(s,\eta)$, any first-order stochastic dominance (FOSD) shift of η toward higher altruism strictly increases immediate social welfare $W(s,\eta)$.*

Proof. Fix the feasible set $R(s,\eta)$. By subsection 1.5.1, the chosen level of social welfare is increasing in altruism: $\frac{dU}{da} > 0$. A FOSD shift of η toward higher altruism therefore raises the average value of the increasing function $U(a,s,\eta)$. Hence $W(s,\eta) = \int_{i=0}^{i=1} U(a_i,s,\eta)\, di$ increases.

LEMMA 2.2: *As s increases, every interior solution $\theta^\star(a_i,s,\eta)$ converges toward the modal direction $\theta_m \equiv \theta^\star(a_m,s_0,\eta)$.*

Proof. We adapt equation (1.8). to the new annotations for a given agent i:

$$\frac{dU(a_i,s,\eta)}{ds} = \underbrace{\frac{dU}{dr}\cdot\frac{\partial r(\theta,s,\eta)}{\partial s}}_{Expansion\ effect} + \underbrace{\left.\frac{dU}{d\theta}\right|_{g(u_i,U)=s}\cdot\frac{d\theta}{dMRTS}\cdot\frac{dMRTS(\theta,s,\eta)}{ds}}_{Reallocation\ effect} \quad \forall a_i \tag{2.6}$$

Because $\frac{d\theta^\star(a_i,s,\eta)}{ds} = \frac{d\theta}{dMRTS}\cdot\frac{dMRTS(\theta,s,\eta)}{ds}$ and, as shown in subsection 1.5.1, $\frac{d\theta}{dMRTS} < 0$, then the reallocation effect is positive if and only if $\frac{d\text{MRTS}_{u_i,U}}{ds} > 0$

In the baseline model, $\frac{\partial MRTS(\theta,s,\eta)}{\partial s}$ could not be signed, but in this endogenous technology version we can use the fact that:

$$\operatorname{sgn}\left(\frac{\partial \text{MRTS}_{u_i,U}}{\partial s}\right) = \operatorname{sgn}\left(\frac{\partial r_s(\theta,s,\eta)}{\partial \theta}\right), \tag{2.7}$$

The formal proof of equation (2.7) appears in Appendix B.1. Intuitively, both elements express a change in the angle of the tangent to the g frontier at the θ angle and both are positive if and only if the slope of the tangent at $(\theta, r(\theta,s,\eta))$ becomes steeper when skills increase. However, the $MRTS$ is defined in Cartesian coordinates, while the right hand side is defined in polar coordinates. This equation allows us to know the sign of the reallocation

effect because the expansion profile of $r(\theta, s, \eta)$ is generated by the distribution of choices:

$$\frac{dr_s(\theta, s, \eta)}{d\theta} = \frac{d\eta(a(\theta, s, \eta))}{d\theta} = \underbrace{\frac{d\eta(a(\theta, s, \eta))}{da}}_{>0 \; iff \; a_i<a_m} \underbrace{\frac{da(\theta, s, \eta)}{d\theta}}_{>0}$$

The sign of $\frac{dMRTS(\theta,s,\eta)}{ds}$ is thus the same as the sign of $\frac{d\eta(a(\theta,s,\eta))}{da}$, meaning that it depends on whether the distribution η is increasing or decreasing at a given a_i. Remembering that η is single-peaked at a_m and $\theta^\star(a_m, s, \eta)$ is the angle with the maximum expansion because it is chosen by most agent, then if $a_i < a_m$, then $\theta^\star(a_i, s, \eta) < \theta^\star(a_m, s, \eta)$ and the reallocation effect drives $\theta^\star(a_i, s, \eta)$ upward as s increases. The reverse is true when $a_i > a_m$, and the reallocation effect is null when $a_i = a_m$.

Hence, as skills increase, a self-reinforcing dynamic concentrates the technological comparative advantage around the unique stationary direction $\theta^\star(a_m, s, \eta) = \theta_m$. This attracts more agents toward that choice, further strengthening the comparative advantage at θ_m. Thus all choices $\theta^\star(a_i, s, \eta)$ converge toward θ_m.

2.4. *Long-term welfare effects*

PROPOSITION 2.1: *With endogenous technological change, in the long-term, that is, for s sufficiently high, 1) $\frac{dW}{da_m} > 0$; 2) $\frac{d^2W}{da_m ds} > 0$ and 3) $\exists\, \tilde{a}$ such that $\frac{dW}{ds} > 0$ iff $a_m > \tilde{a}$.*

An increase in a_m always increases long-term social welfare and, in this version, (the mode of) altruism and skills become complementary, because a_m sets the direction of long-term progress and s sets the distance to the origin. Finally, in parallel with the guardrail proposition 1.2, there exists a threshold $\tilde{a}$ such that the effect of technological progress (or skill increase) on social welfare is positive if and only if $a_m > \tilde{a}$.

Proof: The results from previous subsection show that, as s increases, $\theta(a_i, s, \eta) \to \theta^\star(a_m, s, \eta)\ \forall a_i$ and thus $\frac{d\theta}{ds} \to 0$, hence in the long-term, the second term disappears. As a consequence, in equation (2.6) that decomposes $\frac{dU(a_i,s,\eta)}{ds}$ into its expansion and reallocation effect converges to zero . Hence as s increases only the expansion effect remains:

$$\frac{dU(a_i, s, \eta)}{ds} \to \underbrace{\frac{dU}{dr} \cdot \frac{\partial r(\theta, s, \eta)}{\partial s}}_{Expansion\ effect} \tag{2.8}$$

As shown in last subsection, the extension of the feasible set at θ_m results from the fraction of the population choosing this action: $\frac{\partial r(\theta,s,\eta)}{\partial s} = \eta(a_m)$, where $\eta(a_m$ is fixed by the initial distribution. To analyze the first term, by definition of polar coordinates: $U = r\sin\theta$, hence $\frac{dU(a_i,s,\eta)}{dr} = \sin\theta^\star(a_m, s, \eta)$.

Thus, for any a_i, in the long run: $\frac{dU(a_i,s,\eta)}{ds} \to \sin\theta^\star(a_m, s, \eta)\ \eta(a_m)$.

We now aggregate across agents to demonstrate the effects on social welfare of Proposition 2.1. The first statement of Proposition 2.1 results from:

$$\frac{dW}{da_m} = \int_{i=0}^{i=1} \frac{dU(a_i, s, \eta)}{da_m}\, di \to = \int_{i=0}^{i=1} \frac{dU(a_m, s, \eta)}{da_m} di = \frac{dU(a_m, s, \eta)}{da_m} > 0 \tag{2.9}$$

To show the second statement, building on prior calculations:

$$\frac{dW}{ds} = \int_{i=0}^{i=1} \frac{dU(a_i, s, \eta)}{ds}\, di = \frac{dU(a_i, s, \eta)}{ds} \to \sin\theta^\star(a_m, s, \eta)\ \eta(a_m) \tag{2.10}$$

Hence[8]

$$\frac{d^2W}{ds\, da_m} \to \eta(a_m)\frac{d\sin\theta^\star(a_m, s, \eta)}{da_m} > 0$$

The third statement of Proposition 2.1 results from Equation 2.10, where $\frac{dW}{ds} > 0$ if and only if $\theta_m > 0$ and θ_m is strictly increasing in a_m. Hence there must exist a $\tilde{a}$ such that $\frac{dW}{ds} > 0$ if and only if $a_m > \tilde{a}$. Graphically, if $U(a_m, s_0, \eta) < 0$ then the angle $\theta_m < 0$ and technological progress, which will concentrate in this direction, will only accentuate this social welfare loss. Notice that it is possible that the upper right section of the frontier entirely lies above the horizontal axis, which is consistent with the proposition for $\tilde{a} < 0$ implying $\frac{dW}{ds} > 0\ \forall a_i \in [0, 1]$.

In summary, the increase in s and the mass at the mode drive the length $r(\theta, s, \eta)$, while a_m determines the direction of technological progress. A higher a_m raises θ_m, increasing the extent to which progress translates into higher U.

---

8 $\frac{d\sin\theta^\star(a_m,s,\eta)}{da_m} = \frac{d\sin\theta_m}{d\theta_m}\frac{d\theta_m}{da_m}$. The second element has been shown to be positive and the fist element is positive because the solution is located on the upper right frontier of the feasible set, ensuring that a higher angle increases the contribution of r to U.

2.5. *Real-life examples and implications*

Our results help explain why technological progress does not always benefit society and connect to a literature showing that the path of innovation is shaped by choices and policies. For example, Acemoglu, Aghion, Bursztyn, and Hemous, ADHM+ show that innovation can be redirected from environmentally dirty toward clean technologies through appropriate public incentives, and BSBLS+ highlight the role of Chinese institutional support in the rapid development of strategic sectors such as solar energy. Departing from the canonical model in which profit maximization is the only objective, the motivations of entrepreneurs and workers may vary with their identities and values (**??**). The range of values and related actions is broad: it includes Muhammad Yunus's creation of the Grameen Bank, the 12.8 million jobs in U.S. nonprofit organizations, corresponding to 9.9% of private-sector employment in 2022,[9] and, at the opposite extreme, major cigarette manufacturers' deliberate deception of the public about the health effects and addictiveness of cigarettes.

Artificial intelligence is a salient example: its impacts may range from massive labor displacement and existential risks to human-complementary uses and "sustainable abundance" (Johnson and Acemoglu, Acemoglu and Restrepo). The lottery of technology of section 1 emphasized some technological deterministic component in simply what will turn out to be easier to develop and profit from. This section highlights that individual decisions where altruism plays a key role also shapes such outcomes. Observable actions of the AI sector already illustrate a wide range of values. Firms and researchers differ in their willingness to accept contracts involving mass domestic surveillance or fully autonomous weapons, to incorporate costly guardrails that reduce misuse,[10] and to lobby against regulation. At the upper end of the distribution, Yoshua Bengio, the most cited researcher in AI, launched LawZero, a nonprofit organization focused on developing safer AI systems.

3. THE ROLE OF INSTITUTIONS

3.1. *Non-technical summary when introducing institutions and incentives*

Economists traditionally rely on institutions and incentives more than individual will to induce desirable actions. To discuss this, we introduce institutions, following North's defi-

[9] https://www.bls.gov/bdm/nonprofits/nonprofits.htm

[10] https://darioamodei.com/essay/the-adolescence-of-technology#humanity-s-test

nition of institutions as "the humanly devised constraints that structure political, economic, and social interactions, including formal and informal rules (social norms) and how they are enforced". This version of the model allows a principal or policymaker p to impose constraints that restrain agent i's achievable set through sanctions that make such actions prohibitively costly. However, the policy maker applies such constraints taking into account her own payoffs associated with the different actions and the aggregate payoff, with a weight determined by her own altruism a_p. We find that, if the policymaker has low or limited altruism, then there is no guarantee that the equilibrium with institutions is better than the one without institutions, nor that an increase in institutional power is beneficial for social welfare. We describe this situation as a *lottery of alignment of interests*, meaning that whether institutions improve welfare or not depends on whether the actions that increase social well-being turn out to also be the ones that benefit the policymaker most (under some constraints of implementability driven by the preference of the population). In parallel with the findings of previous sections, *ceteris paribus*, an increase in the altruism of the policymaker translates into an increase in social welfare. The results highlight that institutions empower a certain group of people to implement some rules and set new incentives, yet, without an explicit intention from these empowered people to increase social welfare, there is no guarantee that such power is put to good use. By contrast, highly altruistic people in power can drastically change social welfare.

While institutions can substitute for altruism, recognizing that the set of institutions and incentives in place is itself endogenous restores a fundamental role to altruism. In this setup, institutions become a lever through which altruism can enhance or crystallize an intention to improve social welfare from those in power. An alternative to altruism is that representativity and accountability are sufficiently well distributed to represent the will of the people. In such scenario, one needs to endogenize the institutional arrangement, which would be subject to a *lottery of power struggle* and high risk of power concentration, unless altruism motivates agents to reach a political system that benefits social welfare.

3.2. *The setup with institutions*

In this game, the principal p, who plays first, may prohibit any set $P \subseteq R(s)$ with area $\mathbb{A}(P) \leq I\mathbb{A}(R(s))$ where $I \in [0,1]$ is the institutional capacity, representing the maximum

share of the area $\mathbb{A}(R(s))$ that the principal can credibly prohibit. The agent i then selects his action as in section 1, but restricted to the area of $R(s)$ not prohibited by the principal.

As in the basic model, each action leads to a payoff profile $(u_1, ..., u_N)$, and $R(s)$ is still represented in two dimensions, as defined in subsection 1.2. Since p is also an agent of the economy, to each pair (u_i, U) is associated a principal's direct payoff $u_p(u_i, U)$, which is continuous, differentiable, and strictly concave over $R(s)$.

The policymaker also has a level of altruism a_p that sets her valuation of social welfare and selects the prohibited area P that solves the problem:

$$\begin{aligned} &\max_{P} v_p = (1 - a_p)\, u_p(u_i, U) + a_p U \\ &subject\ to\ the\ agent's\ decision: (u_i, U) = \underset{(u_i,U)\in R(s)\setminus P}{argmax}\ v_i(u_i, U) \end{aligned} \tag{3.1}$$

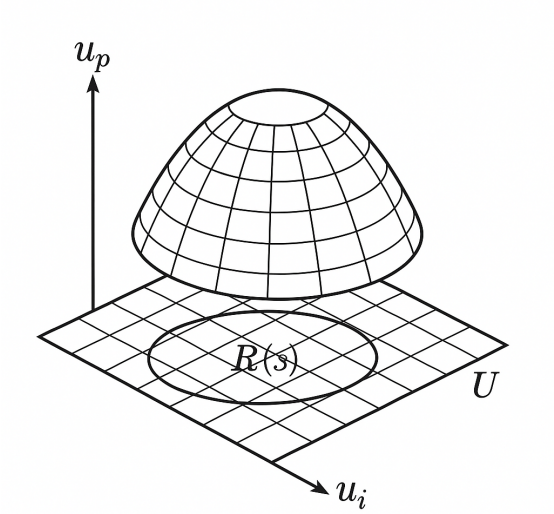
3.3. *Resolution of the problem with institutions*

For conciseness, we provide intuitive explanations of all proofs in the main text with the formal proofs available in Appendix C.

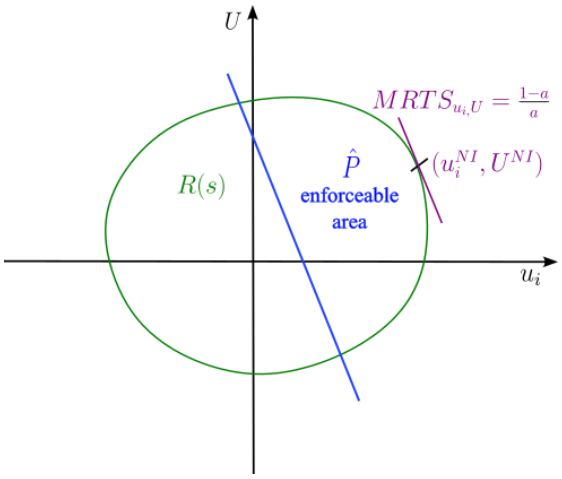
3.3.1. *Description of the solution*

LEMMA 3.1: *$\left(u_i^I, U^I\right)$, the solution to the maximization problem (3.1), is the point within the enforceable area $\widehat{P}$ that maximizes v_p. Where, $\widehat{P}$ is the intersection of $R(s)$ with the open half–plane on the upper right of the slope $-\frac{1-a_i}{a_i}$ such that its area is exactly equal to $I\,\mathbb{A}(R(s))$.*

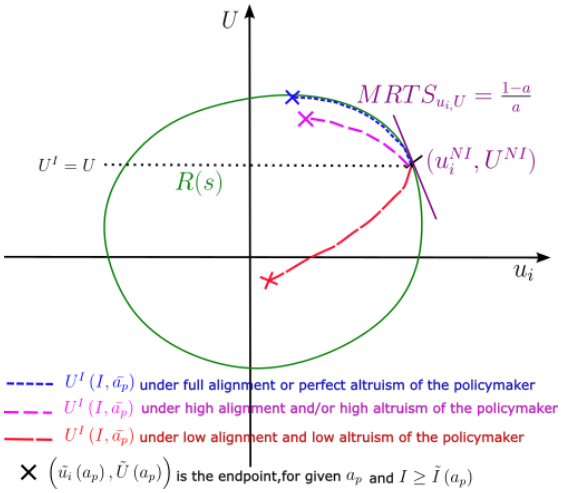
For the principal to have the agent select a given point $(u_i, U) \in R(s)$, she must prohibit all achievable actions that provide an equal or greater payoff $v_i(u_i, U)$ to the agent. As shown in Figure 4, Panel B, these points are located on the upper right of the indifference curve $-\frac{1-a_i}{a_i}$ going through the point that the principal wants to enforce, including the line but not the point (u_i, U). Hence the principal is only able to enforce any point (u_i, U) if the area on the upper right of the agent's indifferent curve with slope $-\frac{1-a_i}{a_i}$ going through (u_i, U) is lower than or equal to $I\,\mathbb{A}(R(s))$. Let $\widehat{P} \in R(s)$ represent the set of enforceable points (u_i, U), represented in Figure 4, Panel B. Then the principal can simply pick the solution to the problem with institutions $\left(u_i^I, U^I\right) \in \hat{P}$ that maximizes v_p and, to enforce it,



Panel A. The dome of u_p associated with each pair (u_i, U)



Panel B. The principal's enforceable area, $\widehat{P}$



Panel C. Different expansion paths as I increases

FIGURE 4.—Model with endogenous institutions.

prohibit the set of other points that generate equal or higher v_i than $v_i\left(u_i^I, U^I\right)$. Notice that, since the principal's indirect utility incorporates U^I with weight a_p, the interior solution is not the highest point in the dome of Figure 4, Panel A but rather the dome's tangency that satisfies $dU = -\frac{(1-a_p)}{a_p} du_p$.

Some basic comparative statics can help understand the solution under different scenarios. Define $\left(\tilde{u_i}(a_p), \tilde{U}(a_p)\right)$ as the points in $R(s)$ that maximize v_p under full institutional

power and $\tilde{I}(a_p)$ as the minimum institutional capacity required to reach $\left(\tilde{u}_i(a_p), \tilde{U}(a_p)\right)$. Also let $\left(u_i^I(I, a_p), U^I(I, a_p)\right)$ represent the expansion path of the solution as I expands. As shown in Figure 4, Panel C it starts at $\left(u_i^I(0, a_p), U^I(0, a_p)\right) = \left(u_i^{NI}, U^{NI}\right)$, where $\left(u_i^{NI}, U^{NI}\right)$ is the solution without institutions, solved in section 1 and ends at $\left(u_i^I(I, a_p), U^I(I, a_p)\right) = \left(\tilde{u}_i(a_p), \tilde{U}(a_p)\right) \ \forall\ I \in \left[\tilde{I}(a_p), 1\right]$. This path is also a function of a_p and its set of solutions is located on the upper frontier of $\hat{P}$ only for $a_p = 1$ or under perfect alignment between the private interest of the principal and social welfare. The strict concavity of $u_p(u_i, U)$ ensures the uniqueness of the solution and that $u_i^I(I, a_p)$ and $U^I(I, a_p)$ are both continuous and differentiable in I and in a_p.

3.3.2. *The ambiguous effect of institutions on social welfare*

PROPOSITION 3.1: *$U^I - U^{NI}$ cannot be signed, nor can $\frac{dU}{dI}$ be signed. Without specifying the altruism of the policymaker, one cannot guarantee that institutions improve social welfare nor that an increase in institutional power is beneficial to social welfare.*

Without institutions, the agent chooses $\left(u_i^{NI}, U^{NI}\right)$, the unique tangency on the upper-right frontier, as derived in section 1 and displayed in Figure 4, Panel B. With institutions of capacity $I \in [0, 1]$, the principal ends up selecting the point within $\hat{P}$ that maximizes $v_p(u_i, U)$. Because the model puts no restriction on $u_p(u_i, U)$ other than its concavity on $R(s)$, the maximizer of $v_p(u_i, U)$ within $\hat{P}$ can be located at any point in $\hat{P}$. Unlike in Figure 4, Panel A, the dome needs not be symmetric and its maximum can be located at any point. The more the peak is located at high values of U, the higher the alignment of interests between the policymaker and social welfare. As shown in Figure 4, Panel C, the expansion path $U^I(I, a_p)$ might or might not lie above the horizontal line $U = U^{NI}$, depending on the policymaker's altruism and alignment of interests with social welfare, since both affect the location of the dome's tangent that satisfies $\frac{dU}{du_p} = -\frac{(1-a_p)}{a_p}$.

The sign of $\frac{dU}{dI}$ reflects the effect of strengthening institutions on social welfare. Increasing I extends the frontier of $\widehat{P}$. If $I \geq \tilde{I}(a_p)$, the interior equilibrium remains unchanged and $\frac{dU}{dI} = 0$. If $I < \tilde{I}$, then $\left(u_i^I(I, a_p), U^I(I, a_p)\right)$ moves with the frontier. Whether the associated U^I rises or falls is again governed by the same alignment consideration: if, in the newly opened band, the principal's private objective u_p tends to pick points with higher

U, then U^I rises; if it tends to pick points with lower U, then U^I falls. In short, without further assumptions, the effect of stronger institutions is also ambiguous.

3.3.3. *The effect of having a more altruistic principal on social welfare :* $\frac{dU^I(I,a_p)}{da_p}$

PROPOSITION 3.2: *For fixed institutional capacity* $I > 0$, $U^I(I, a_p)$ *is weakly increasing in* a_p *and, if the institutional solution is interior, then* $\frac{\partial U^I(I,a_p)}{\partial a_p} > 0$.

Fixing institutional capacity I pins down the enforceable region $\widehat{P}$ represented in Figure 4. The principal's objective, maximizing v_p, can be pictured as a tilting plane placed over the dome $u_p(u_i, U)$. As a_p increases, the plane tilts toward the U-axis, putting greater weight on social welfare. The proof of Proposition 3.2 shows that this tilt cannot make the implemented point move to a lower value of U. If it did, the earlier point with higher U would be valued relatively more under the larger altruism weight, contradicting the optimality of the later point. Hence $U^I(I, a_p)$ is weakly increasing in a_p. Appendix C.2 formalizes this argument and establishes strictness for interior solutions: $\partial U^I(I, a_p)/\partial a_p > 0$.

3.3.4. *When the policymaker's altruism is high enough, institutions increase social welfare*

PROPOSITION 3.3: *If* $a_i < 1$ *and* a_p *is high enough, then* $U^I > U^{NI}$ *&* $\frac{dU^I(I,a_p)}{dI} \geq 0$.

When the agent's altruism is imperfect, if the policymaker's altruism is high enough, social welfare is greater with institutions than without them and is weakly increasing in institutional capacity. As shown in Figure 4, Panel C, a policymaker with $a_p = 1$, maximizing U^I subject to $(u_i, U) \in \hat{P}$ would move her optimal solution counterclockwise, along the upper contour of $R(s)$. Hence when $a_p = 1$ and $a_i < 1$, then $U^I(I, a_p) > U^{NI}\ \forall I > 0$. The tangency of the frontier of $R(s)$ with slope $-\frac{1-a_i}{a_i}$ at the solution $\left(u_i^{NI}, U^{NI}\right)$ ensures the existence of points with higher U are in $\widehat{P}$ (Figure 4, Panel B), and the convexity of $R(s)$ ensures $\frac{dU^I(I,1)}{dI} > 0$ for $U^{NI} < U^I < U^{max}$. Once $U_I = U^{max}$, $\frac{dU^I(I,1)}{dI} = 0$. Because $U^I(I, a_p)$ is continuous and differentiable in both parameters, if $\frac{dU^I(I,1)}{dI} > 0$ then, for a_p close enough to 1, $\frac{dU^I(I,a_p)}{dI} > 0$. In Proposition 3.3 U^I is only weakly increasing in institutional capacity since the case where $I > \tilde{I}(1)$ and $U_I = U^{max}$, then $\frac{dU^I(I,1)}{dI} = 0$, then for a_p close enough to one, $I > \tilde{I}(a_p)$ and thus $\frac{dU^I(I,a_p)}{dI} = 0$.

Proposition 3.2 shows that $\frac{dU^I(I,a_p)}{da_p} \geq 0 \; \forall I, a_p$, and the previous paragraph shows that $U^I(I,1) > U^{NI}$. It is thus clear that $\forall I, \; \exists \tilde{a_p}(I) \in [0,1)$ such that $U^I(I,a_p) \geq U^{NI}$ if and only if $a_p \geq \tilde{a_p}$. This minimum $\tilde{a_p}$ needed to ensure positive effects of institutions is a function of I and of the alignment of interests: the more the interests of the population and the policymaker are aligned, the less altruism is needed and $\frac{dU^I(I,a_p)}{da_p} = 0$ only when $U^I(I,a_p)$ is already maximizes U within the implementable set $\widehat{P}$.

3.4. *Interpretation and connection to historical examples*

Proposition 3.1 can be interpreted as the *lottery of alignment of interests*: when a_p is null or low, institutional power steers behavior toward what best serves the principal's private objective u_p. If policymaking is controlled by group of individuals with preferences weighted by their political power, social welfare rises only when that group's private incentives align with it. A leader with private stakes in clean energy may raise welfare, while an oil-company lobby is more likely to lower it. A perfectly functioning democracy could substitute for altruism if policymakers represented all individuals equally, but this shifts the question to why such institutions arise and persist. In the absence of altruism, political power is itself subject to a *"lottery of power struggle"*. Modeling that process is beyond this article and is a useful extension.

The lottery of alignment of interests can be used to describe the colonial episodes documented by Acemoglu, Johnson, and Robinson. European powers implemented extractive institutions where feasible and more inclusive institutions only when purely extractive institutions were not feasible (e.g. due to high colon mortality or organized resistance). In this context where colonizers were largely indifferent toward local populations, long-term welfare in the colony rose only when what served the colonizers happened, by circumstance, to align with the policies that contributed to broad-based prosperity.

4. FROM ALTRUISM TO UNIVERSALISM

Up to this section, homogeneous altruism toward all other individuals was always good. Yet in practice, altruism may be stronger toward some individuals and weaker toward outgroups. This section clarifies whether more altruism remains beneficial when directed un-

evenly, then studies a sub-model in which technologies expand the ability to reach people beyond one's parochial circles faster than altruism adapts.

4.1. *Parochial altruism versus universalism*

4.1.1. *Basic setting allowing for parochial altruism*

Assume there are G groups. Fix an individual i belonging to group $G(i)$. Let u_i denote i's utility, u_G the aggregate utility of i's group (including i), and U total utility in society. Altruism is allowed to differ between in-group members and outsiders. Let a^g denote the valuation of in-group welfare and a^o the valuation of outsiders' welfare, with $0 \leq a^o \leq a^g \leq 1$.

The new valuation is given by $v_i\left(U_i, U_{G-i}, U_{-G}\right) = u_i + a^g u_{G-i} + a^o U_{-G}$, which, using $U_G = U_i + U_{G-i}$ and $U = U_G + U_{-G}$ can be redefined over U_i, U_G and U to define agent i's constrained maximization:

$$\begin{aligned} &\max\ v_i\left(U_i, U_{G-i}, U_{-G}\right) = \left(1 - a^g\right) u_i + \left(a^g - a^o\right) u_G + a^o U \\ &\text{s.t.} \quad \left(U_i, U_G, U\right) \in R(s) = \left\{ \left(u_i, U_G, U\right) \in \mathbb{R}^3 \mid g(u_i, U_G, U) \leq s \right\}. \end{aligned} \tag{4.1}$$

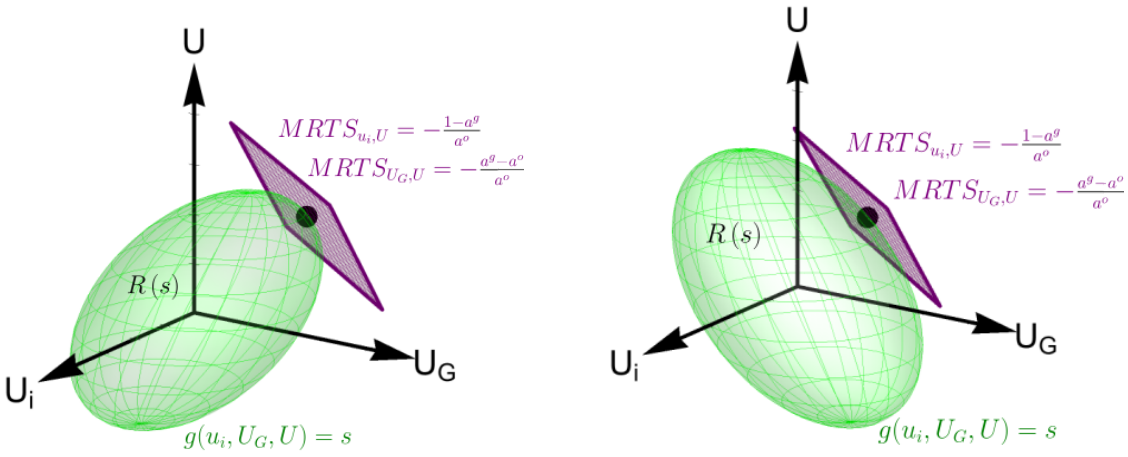
where an increase in skills s expands $R(s)$ in all directions. The problem is thus very similar to the basic problem of Section 1, but in three dimensions, and it will be subject to the same mechanisms. The main text below provides the intuition, and the detailed proofs are in Appendix D.

4.1.2. *Characterization of the solution and its geometry*

Under the same regularity conditions as in the baseline model (compactness, strict convexity of $R(s)$, and a smooth frontier), the maximizer (u_i^*, U_G^*, U^*) exists, is unique, and lies on the boundary of $R(s)$. As in the basic model, on the relevant portion of g, we can represent the frontier g as $U = U(U_i, U_G, s)$, satisfying $g\big(u_i, U_G, U(U_i, U_G, s)\big) = s$, and the two tangency conditions of the optimal solution can be written as

$$MRTS_{u_i,U} \equiv \frac{dU(u_i^*, U_G^*, s)}{du_i} = -\frac{1 - a^g}{a^o} \ ; \ MRTS_{U_G,U} \equiv \frac{dU(u_i^*, U_G^*, s)}{dU_G} = -\frac{a^g - a^o}{a^o} \tag{4.2}$$

Figure 5 illustrates (4.2). The green ellipsoid represents an achievement set $R(s)$, the purple plane is an indifference plane of v_i that satisfies the two first-order conditions, and the black dot is the unique tangency point (u_i^*, U_G^*, U^*).



Panel A. U_G and U are positively aligned

Panel B. U_G and U are negatively aligned

FIGURE 5.—Optimal choice under altruism differentiated toward one's group and others. The green surface is the achievement set $R(s)$, the black dot is the unique solution (u_i^*, U_G^*, U^*), and the purple patch is the local indifference tangent plane. The functional form and parameters are described in Appendix D.

4.1.3. *Comparative statics and the lottery of alignment.*

PROPOSITION 4.1: $\frac{dU^*}{da^o} > 0$ *and the sign of* $\frac{dU}{da^g}$ *is undetermined.*

Proposition 4.1 states that increased altruism always increases social welfare when not directed, but the directed altruism cannot be signed. Graphically, an increase in a^o moves the supporting plane in a direction that places more emphasis on U. By contrast, an increase in a^g pushes the optimum toward points with higher U_G, which might turn out to be positively aligned with U, in which case it would be beneficial for social welfare, as in Figure 5, Panel A. But the effect on social welfare may also go in the opposite direction, as in Figure 5, Panel B. Hence whether this raises or lowers U depends on the technology, distribution of payoffs and "lottery of alignment": an increase in directed altruism only benefits social welfare when the actions that benefit the group turn out to also benefit society. Interpretations and applications are discussed in subsection 4.3.

4.2. *When the reach of the technology expands faster than altruism*

This subsection considers an extension of the universalism model where the reach of externalities increases with skills while the reach of altruism catches up with a lag, creating a catch-up period when a misuse of skills can occur to generate benefits for the agents whose skills improved at the expense of outsiders that can be reached but toward whom altruism has not yet developed.

4.2.1. *The setting*

Individuals are organized in a nested social structure: individuals belong to families, which belong to villages, municipalities, countries, continents, and to the world. We build on the model of the previous subsection, adding the following assumptions. First, assume some exogenous growth in skills over time given by a predetermined function $s(t)$. Second, assume L levels of social structure with L cutoffs $\tilde{s}_l$ strictly increasing in l (including $s_1 = 0$. In this model, the actions of agent i only affect agents that share the same group at level l if $s > \tilde{s}_l$. This is meant to reflect the fact that technological development tends to increase the reach of externalities. Third, we endogenize the level of altruism in the optimistic and extreme way described hereafter.

Agents have no altruism toward outsiders $a^o = 0$, and full altruism toward their inner circle $a^g = 1$. However the inner circle to which a^g applies varies over time: any person i includes in her inner circle all the persons that share level l after the externalities of her actions have been able to reach this level l for at least τ periods, implying a lag between the time when agents dominate a technology that can affect new communities and the time when these agents include these new communities in their inner circle. The rationale behind the assumption is that only when people's actions can potentially affect a certain group does society initiate a public debate and reflection, which takes time before agents start "caring" about these new individuals.

Formally, at the beginning of the game, agent i with skills s_0 applies full altruism to agents that share the same group at level l if $s_0 > \widetilde{s_l}$. After period $s_0 + \tau$, altruism towards another agent switches to $a^g = 1$ only once $s_{t-\tau} > \tilde{s}_l$, meaning after s_t has exceeded $\widetilde{s_l}$ for at least τ periods. To give a concrete example, assume that continent is level 5 and world is level 6, when s_t switches from just below $\widetilde{s_{l=6}}$ to $s_t \geq \widetilde{s_{l=6}}$, then agent i's externalities can

reach people in other continents but during the following τ periods, agent i does not value the welfare of people in other continents.

Based on the tools and results developed previously, the implications are straightforward. When an agent has full altruism toward all agents that she can reach, the setting becomes equivalent to the basic problem of section (1) under full altruism. Hence, we know that the agent will pick the action that maximizes social welfare and skill improvements can only benefit social welfare. However, during the τ periods following $s_t = \widetilde{s}_l$, this is equivalent to the results developed in the last subsection, where the lack of universalism creates a lottery of alignment of interests with technological progress potentially harming the new communities that can be reached and aggregate social welfare.

4.3. *Lessons from the model with differentiated altruism and applications*

The first lesson is that, unlike in Section 1, an increase in altruism need not always raise welfare once it differs across individuals. Bias is introduced whenever agent i assigns different weights to different agents' utilities. Whether she favors herself or group G, uneven weights create a risk of inefficient choices relative to social-welfare maximization. The "lottery of alignment of interests" then asks whether actions that benefit agent i's inner circle also benefit social welfare. A key lesson from Subsection 4.1 is that what guarantees development is not any form of altruism, but universalism: an increase in altruism toward all humans. Subsection 4.2 clarifies how the required scale of altruism depends on technological progress. If each agent's actions can affect only her tribe, tribal altruism is sufficient. Parochial altruism can still support intra-household fairness, information diffusion within networks, local institutions, or informal insurance. The broader lesson is that altruism is required at the system level, where the relevant system is defined by the cluster within which externalities are confined.[11] A system is functional if its decision process incorporates the consequences of its acts within the system.

The conclusions speak to various stylized facts and observations from other disciplines. Bowles and Gintis, Henrich discuss how prosociality and ability to cooperate can determine which societies thrive, as well as the perils of parochial altruism. Platteau argues that leaders often divert resources because social norms require them to share benefits with kin and

[11] We use Von Bertalanffy's definition of a system as "a set of elements standing in interrelations"

community networks, making strict impartial governance difficult. In more extreme scenarios, **?** conclude that parochial altruism and dehumanization of the enemy strongly predict participation in armed groups and severe violence in Congo. Together with our conclusion, the findings highlight the importance of the recently growing literature around universalism (Enke, Enke, Enke, Rodríguez-Padilla, and Zimmermann).

Examples of applications of the model include recurrent European conflicts from the Middle Ages through the early twentieth century, before the creation of a stronger sense of European unity; African nations' struggles to build national unity; and the need to raise "citizens of the world" when facing growing environmental threats, where moral consciousness appears to lag behind the urgent need to adjust our way of living. But perhaps one of the clearest illustrations arose when technological progress allowed colonizers to reach new territories, leading to mass killing, slavery and exploitation in Africa and the Americas.

Religious and philosophical debates emerged over the rational, spiritual, and legal status of colonized peoples. Influential early arguments drew on ideas of natural slavery, just war, forced conversion, and alleged cultural or religious inferiority to legitimate conquest and inhuman treatments.

These classifications were later contested from within European theological, philosophical, and legal traditions, contributing gradually to a language of natural rights, liberty, sovereignty, and human dignity, progressively eroding the legitimacy of colonization and forced labor (**?**).

Subsection 3.4 discusses how, at a time with extremely low altruism towards indigenous populations, the implementation of extractive versus inclusive institutions was largely driven by a lottery of alignment of interests where long-term development was largely driven by what best served the interests of the colonizers at the time. Generations later, most decolonizations were the result of fierce struggles with specific explanatory factors, yet the major drawback in colonization and forced labor were quite concentrated in time, after World War II (as shown in Figure 1, Panels A and B), which coincides well with the growing recognition of human rights.

To the best of our knowledge no current economic-development model provides an explanation for the rise and fall of certain large-scale evils. These include colonization, slavery, wars with mass death, environmental destruction, apartheid systems, and certain forms of corruption. They share one common pattern: first, the technology made a new form of

harm feasible, often reaching new populations, then the harm occurred at scale, before it became morally unacceptable and lost ground.

5. ADDITIONAL DEBATES

5.1. *Cooperation without Altruism*

Cooperation among non-altruistic individuals is possible. Thus, the paper could have been framed as a theory of development through technological progress and cooperation, being agnostic about the source of cooperation. Since the baseline model deliberately abstracts from strategic interaction, characterizing the full set of cooperative equilibria among self-interested agents would be a valuable extension to this work. Still, the standard theory helps anticipate the limits of such equilibria.

Cooperation between non-altruistic agents requires incentive-compatible enforcement, typically through the threat of retaliation. Thus, cooperation can be sustained without altruism when agents can retaliate, or when information about actions reaches agents who have the capacity to retaliate (**?**). These conditions become harder to satisfy as the number of players increases, and also in asymmetric relationships. In larger groups, actions are harder to monitor and individual incentives to free ride are harder to discipline (**?**). If a powerful country can seize resources from a weaker country, a self-enforcing agreement can prevent appropriation only to the extent that the agreement is also preferable for the powerful country. The same limitation applies to economic inequalities, restricting the scope to correct injustice or for welfare-improving redistribution when not incentive compatible for powerful actors. Global environmental problems combine all forms of challenges: the number of players is large, consequences are widely dispersed, and observability is costly. These complex cases tend to require institutions designed to monitor behavior, coordinate responses, and impose sanctions (**?**). But, as discussed in Section 3, institutions are themselves endogenous and, in the absence of altruism, may serve the interests of those with the power to shape them.

Hence cooperation among non-altruistic agents can solve some social dilemmas, yet the puzzle of human cooperation is precisely that observed cooperation often extends beyond relationships in which direct reciprocity is feasible (Fehr and Fischbacher). A useful interpretation of the baseline model is that the costs and benefits of immediate reactions and sanctions are already incorporated into the payoffs that determine agent i's choices. The

main results should therefore be read as applying to residual dilemmas after self-enforcing cooperation is accounted for. All forms of cooperation matter, but altruism helps guarantee it. The "lottery of technology" and the "guardrail theorem" do not imply that altruism is a necessary condition for development: development may occur without altruism when incentive-compatible cooperation is feasible. Rather, altruism is a sufficient condition because it ensures that higher skills are used well, even when technological bias is unfavorable and enforcement is insufficient.

5.2. *Do Prosocial Norms Sometimes Backfire?*

An economic literature documents settings in which social, sharing, or other prosocial norms appear to generate inefficiency. For example, the "kinship tax" operates as a socially enforced redistributive pressure, creating an implicit tax on observable income that weakens incentives to exert effort or invest in income-generating activities (Jakiela and Ozier). More broadly, Acemoglu and Robinson describe the "cage of norms" as an oppressive set of social rules that restrict individual liberties and economic activity. There are three reasons why such findings are not necessarily in tension with the model.

First, a sharing norm that discourages effort may nevertheless be welfare improving if the benefits from insurance and redistribution exceed the costs. After all, formal redistribution systems suffer from similar limitations but are still regarded as beneficial in many contexts. Second, in practice, the consequences of one's actions may be uncertain. Incorporating random noise into payoff realizations would allow good intentions not always to translate into higher aggregate payoffs. This may occur if a norm adopted to protect vulnerable individuals and increase social welfare actually harms long-run income and social welfare because of unanticipated disincentive effects. Yet, unless all noises are expected to be systematically biased in one direction, in general, the underlying intention still determines expected payoffs. Third, claims of altruistic purpose may mask vested interests. For example, elites may promote sharing obligations or compliance norms to protect their rents and power. False claims of altruism are common and the sincerity of the altruistic motive is central to its expected social benefits.

5.3. *Different Forms of Prosocial Preferences*

This paper models pure altruism: direct concern for the welfare of others or for aggregate social welfare (Andreoni and Miller, Batson), yet it is not the only form of other-regarding preferences. Reciprocity, conditional cooperation, and altruistic punishment are widespread and predict real-life decisions (Ligon and Schechter, **?**). Warm-glow giving reflects private utility from giving itself, focusing on one's own contribution (Andreoni). Also, social norms co-evolve with preferences and can sustain cooperation through internalized obligations, shared expectations, and informal sanctions (Bowles and Gintis). Collective moral values can persist through historical experience and intergenerational transmission (**?**Doepke and Zilibotti).

The model focuses on pure altruism because, by definition, it maps directly into social-welfare maximization. This does not imply that other prosocial motives cannot generate similar qualitative results. In this limited-interaction model, warm glow may be equivalent if each agent focuses on her own contribution to social welfare. Reciprocity-based motives could also support cooperation, but they would introduce equilibrium considerations absent from the baseline model. In particular, they may generate multiple equilibria: cooperation can be sustained when agents expect others to cooperate, while low-trust expectations can lead it to unravel. In some contexts punishment can increase cooperation, including in the presence of some non-altruistic agents. Modeling the consequences of different forms of prosocial preferences are a promising extension of this work.

Different other-regarding preferences matter, but the core mechanism requires sincere intention and care for others. The same observed action can reflect different motives: a donation may reveal genuine concern, social pressure, or a desire to be seen giving (DellaVigna, List, and Malmendier). Punishment can be altruistic, aiming to induce socially desirable behavior, or motivated by vengeance. Different intentions can produce the same actions in some contexts, but without sincere concern one must rely on an alignment of interests that may not always hold. By contrast, genuine and inclusive concern for others is a robust channel through which greater skills become sufficient conditions for broad welfare improvement.

5.4. *Is Altruism Diverse and Malleable?*

This work finds that skills and altruism are sufficient conditions for development, implying that selection into key positions matters and that shifting the distribution of altruism is valuable. Such recommendations make sense only if altruism is diverse and malleable. Multidisciplinary work suggests that some degree of altruism is the norm rather than the exception (HBBC+, Batson, FBDE+).

A growing literature, initially led by psychology suggests that altruism and closely related social preferences are partly endogenous. Empathic concern can elicit altruistic motivation (Batson) and compassion training can increase altruistic redistribution (**?**). Economics literature shows the effects of life events and human interventions on prosociality. Exposure to violence can increase in-group altruism (VNVB+), yet at the cost of universalism (**?**) and historical exposure to the slave trade reducing trust (**?**). Children's prosociality responds to parental mental health (**?**), family environment and randomized mentoring (KDPSH+). Early-childhood education affects later social preferences (Cappelen, List, Samek, and Tungodden), pedagogy that emphasize socio-emotional learning and that promotes cooperation, like Montessori education, promotes altruistic preferences (**?**Lillard and Else-Quest). Alan, Baysan, Gumren, and Kubilay show that perspective-taking curricula can increase trust, reciprocity, cooperation, and altruism toward peers in ethnically mixed schools. Social interaction clearly matters: exposure to poorer classmates increases generosity and egalitarianism (**?**), in particular when intergroup contacts are collaborative (**?**). Targeted interventions among adults are also promising: Mehmood, Naseer, and Chen find that an intervention cultivating prosociality among deputy ministers in Pakistan can increase altruism preferences and decisions. Together, the evidence supports that altruism-related motives and behaviors are quite malleable, but also highlights our limited knowledge of what interventions can durably shape altruism at scale. If altruism is indeed one of the fundamental drivers of development it is certainly worth a lot more investigation to understand its drivers to improve the design of education and large-scale programs toward this purpose.

6. CONCLUSION AND POLICY IMPLICATIONS

This paper shows that skills and altruism together are sufficient conditions for development, defined as an increase in generated social welfare. According to the model, skills

expand the set of feasible payoffs and altruism guides the decision maker's choice. Yet without altruism, the effect of an increase in skills and the resulting technological progress on development is subject to "the lottery of the technology", referring to the uncertainty about whether new technologies that benefit oneself the most will tend to positively affect others or do so at the expense of social welfare. By contrast, an altruism that is sufficiently high to discourage the use of harmful innovations ensures that only the new feasible actions that contribute to social welfare are used. Rather than viewing technological progress, institutions, and incentives as substitutes for altruism, the framework shows that when they are endogenous to human decisions, they become levers through which the good intentions of policymakers and innovators crystallize to enhance their benefits and make them more durable. For example, taxes to fund public goods and redistribution or beneficial market transactions become more likely to occur at scale in the presence of altruistic policymakers. Finally, the research shows that when technology allows the consequences of one's actions to reach others across different groups or countries, then, to ensure development, altruism must be indiscriminate. To support the historical role of universalism we observe that massive reductions in colonization, forced labor, or increases in voting rights empirically coincide with the raise of universalistic values, which accelerated after World War II and deteriorated in the recent years.

The theoretical setup is intentionally broad, with limited assumptions, making the key conclusions relatively insensitive to many modeling choices. Still, the model is deterministic, includes only specific strategic interactions (in sections 2 and 3, restricts heterogeneity, and uses pure altruism as the only other-regarding preference. The framework offers a simple and flexible setting in which to add these complexities to study how capacities, preferences, and choices affect development.

Two policy implications follow. First, societies invest heavily in skills and technological progress, but comparatively little in understanding and fostering altruism and universalism. This is especially important when technology evolves rapidly. More research and policy implementation is needed to design interventions that increase altruism at scale, including early-life, schooling, or parenting interventions, efforts to broaden the moral circle, and interventions that mitigate the preference-damaging effects of violence and trauma. Second, selection into key positions matters. Because pivotal actors can use high capability to shape institutions and technology trajectories, selecting altruistic individuals or promoting altru-

ism among targeted populations should have amplified consequences for development. In democracies, however, leader selection depends on voters' preferences, so selection itself may depend on broader preferences and values.

REFERENCES

Acemoglu, Daron, Philippe Aghion, Leonardo Bursztyn, and David Hemous (????a), "The environment and directed technical change." 102 (1), 131–166, URL https://www.aeaweb.org/articles?id=10.1257/aer.102.1.131. [3, 17, 23]

Acemoglu, Daron, Simon Johnson, and James A Robinson (????b), "The colonial origins of comparative development: An empirical investigation." 91 (5), 1369–1401. [29]

Acemoglu, Daron and Pascual Restrepo (????), "The wrong kind of AI? artificial intelligence and the future of labour demand." 13 (1), 25–35. [23]

Acemoglu, Daron and James A. Robinson (????a), *The Narrow Corridor: States, Societies, and the Fate of Liberty: Winners of the 2024 Nobel Prize in Economics*, Penguin UK. [4, 36]

Acemoglu, Daron and James A Robinson (????b), *Why nations fail: The origins of power, prosperity, and poverty*, Crown Currency. [7]

[ADHM+] Aghion, Philippe, Antoine Dechezleprêtre, David Hemous, Ralf Martin, and John Van Reenen (????), "Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry." 124 (1), 1–51. [23]

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay (????), "Building social cohesion in ethnically mixed schools: An intervention on perspective taking*." 136 (4), 2147–2194, URL https://doi.org/10.1093/qje/qjab009. [38]

Andreoni, James (????), "Impure altruism and donations to public goods: A theory of warm-glow giving." 100 (401), 464–477, URL https://doi.org/10.2307/2234133. [4, 37]

Andreoni, James and John Miller (????), "Giving according to GARP: An experimental test of the consistency of preferences for altruism." 70 (2), 737–753. [2, 6, 37]

Ashraf, Nava and Oriana Bandiera (????), "Altruistic capital." 107 (5), 70–75, URL https://www.aeaweb.org/articles?id=10.1257/aer.p20171097. [6]

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott S. Lee (????a), "Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services." 110 (5), 1355–1394, URL https://www.aeaweb.org/articles?id=10.1257/aer.20180326. [6, 7]

Ashraf, Nava, Oriana Bandiera, and B Kelsey Jack (????b), "No margin, no mission? a field experiment on incentives for public service delivery." 120, 1–17. [6]

[BSBLS+] Banares-Sanchez, Ignacio, Robin Burgess, David Laszlo, Pol Simpson, John Van Reenen, and Yifan Wang (????), "Ray of hope? china and the rise of solar energy." [23]

Batson, Charles Daniel (????), *Altruism in humans*, Oxford University Press. [2, 37, 38]

Besley, Timothy (????), "State capacity, reciprocity, and the social contract." 88 (4), 1307–1335, URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16863. [7]

Besley, Timothy and Maitreesh Ghatak (????), "Prosocial motivation and incentives." 10, 411–438, URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-063016-103739. [6, 7]

Bisin, Alberto and Thierry Verdier (????), "The economics of cultural transmission and the dynamics of preferences." 97 (2), 298–319, URL https://www.sciencedirect.com/science/article/pii/S0022053100926784. [7]

Bourlès, Renaud, Yann Bramoullé, and Eduardo Perez-Richet (????), "Altruism in networks." 85 (2), 675–689, URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA13533. [6]

Bowles, Samuel and Herbert Gintis (????), "A cooperative species: Human reciprocity and its evolution." In *A cooperative species*, Princeton University Press. [2, 6, 33, 37]

Bénabou, Roland and Jean Tirole (????), "Incentives and prosocial behavior." 96 (5), 1652–1678. [6]

Cappelen, Alexander, John List, Anya Samek, and Bertil Tungodden (????a), "The effect of early-childhood education on social preferences." 128 (7), 2739–2758. [38]

Cappelen, Alexander W., Benjamin Enke, and Bertil Tungodden (????b), "Universalism: Global evidence." 115 (1), 43–76, URL https://www.aeaweb.org/articles?id=10.1257/aer.20230038. [4, 6]

Connolly, Daniel J., George Loewenstein, and Nick Chater (????), "An s-frame agenda for behavioral public policy research." 9 (3), 593–613, URL https://www.cambridge.org/core/journals/behavioural-public-policy/article/an-sframe-agenda-for-behavioral-public-policy-research/3817664ED39D8C25FB5B4C04FE13AAC1. [3]

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (????), "Testing for altruism and social pressure in charitable giving *." 127 (1), 1–56, URL https://doi.org/10.1093/qje/qjr050. [37]

Do Carmo, Manfredo P (????), *Differential geometry of curves and surfaces: revised and updated second edition*, Courier Dover Publications. [1]

Doepke, Matthias and Fabrizio Zilibotti (????), "Parenting with style: Altruism and paternalism in intergenerational preference transmission." 85 (5), 1331–1371, URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14634. [37]

Durkheim, Émile (????), *Le suicide: étude de sociologie*, G. Baillière & C. [2]

Enke, Benjamin (????a), "Kinship, cooperation, and the evolution of moral systems*." 134 (2), 953–1019, URL https://doi.org/10.1093/qje/qjz001. [6, 34]

Enke, Benjamin (????b), "Moral values and voting." 128 (10), 3679–3729, URL https://www.journals.uchicago.edu/doi/abs/10.1086/708857. [34]

Enke, Benjamin, Ricardo Rodríguez-Padilla, and Florian Zimmermann (????), "Moral universalism and the structure of ideology." 90 (4), 1934–1962, URL https://doi.org/10.1093/restud/rdac066. [4, 6, 34]

[FBDE+] Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde (????), "Global evidence on economic preferences." 133 (4), 1645–1692. [38]

Fehr, Ernst and Urs Fischbacher (????), "The nature of human altruism." 425 (6960), 785–791, URL https://www.nature.com/articles/nature02043. [35]

Groenewegen, P. D. (????), "In praise of gournay (eloge de gournay) (1759)." In *The Economics of A.R.J. Turgot* (P. D. Groenewegen, ed.), 20–42, Springer Netherlands, URL https://doi.org/10.1007/978-94-010-1073-3_4. [4]

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (????), "Civic capital as the missing link." 1, 417–480. [7]

Henkin, Louis (????), *The age of rights*, Columbia University Press. [6]

Henrich, Joseph (????), "The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter." In *The secret of our success*, princeton University press. [33]

[HBBC+] Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q. Patton, and David Tracer (????), ""economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies." 28 (6), 795–815, URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/economic-man-in-crosscultural-perspective-behavioral-experiments-in-15-smallscale-societies/6EFFD9263D9A5F2FE5DE9DE8FBBA4988. [6, 38]

[HCHLH+] Herrmann, Esther, Josep Call, María Victoria Hernández-Lloreda, Brian Hare, and Michael Tomasello (????), "Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis." 317 (5843), 1360–1366. [2]

Hume, David (????), "A treatise of human nature: Volume 1: Texts." [2]

Jakiela, Pamela and Owen Ozier (????), "Does africa need a rotten kin theorem? experimental evidence from village economies." 83 (1), 231–268, URL https://www.jstor.org/stable/43868462. [4, 36]

Johnson, Simon and Daron Acemoglu (????), *Power and progress: Our thousand-year struggle over technology and prosperity\textbar Winners of the 2024 Nobel Prize for Economics*, Hachette UK. [23]

[KDPSH+] Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk (????), "The formation of prosociality: causal evidence on the role of social environment." 128 (2), 434–467. [38]

Ligon, Ethan and Laura Schechter (????), "Motives for sharing in social networks." 99 (1), 13–26, URL https://www.sciencedirect.com/science/article/pii/S0304387811001179. [4, 37]

Lillard, Angeline and Nicole Else-Quest (????), "Evaluating montessori education." 313 (5795), 1893–1894, URL https://www.science.org/doi/10.1126/science.1132362. [38]

Mehmood, Sultan, Shaheen Naseer, and Daniel L. Chen (????), "Altruism in governance: Insights from randomized training for pakistan's junior ministers." 170, 103317, URL https://www.sciencedirect.com/science/article/pii/S030438782400066X. [7, 38]

North, Douglass C (????), *Institutions, institutional change and economic performance*, Cambridge university press. [23]

Platteau, Jean-Philippe (????), "Monitoring elite capture in community-driven development." 35 (2), 223–246, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7660.2004.00350.x. [33]

Smith, Adam (????), "An inquiry into the nature and causes of the wealth of nations: Volume one." London: printed for W. Strahan; and T. Cadell, 1776. [4]

Von Bertalanffy, Ludwig (????), "General system theory." 41973 (1968), 40. [33]

[VNVB+] Voors, Maarten J, Eleonora E M Nillesen, Philip Verwimp, Erwin H Bulte, Robert Lensink, and Daan P Van Soest (????), "Violent conflict and behavior: a field experiment in burundi." 102 (2), 941–964. [38]

APPENDIX A: APPENDIX OF THE BASELINE MODEL

The core part of the demonstration of the baseline model was kept in the main text.

A.1. *Proof of Proposition 1.1, components 1) $\frac{dU}{da} > 0$ & 3): the cross-derivative $\frac{dU^2}{dads}$ cannot be signed.*

We aim to study the sign of $\frac{dU}{da}$, which after applying the chain rule gives:

$$\frac{dU}{da} = \frac{dU}{dMRTS_{u_i,U}} \frac{dMRTS_{u_i,U}}{da}$$

We solve it making use of the facts that $MRTS_{u_i,U} = -U^{u_i}(u_i, s)$ and $MRTS_{u_i,U} = \frac{1-a}{a}$.

A standard result in differential geometry from Do Carmo's *Differential Geometry of Curves and Surfaces.* allows to estimate how an exogenous change in the slope of a function affects the value of this function, which we use to demonstrate that

$$\frac{dU}{d\text{MRTS}_{u_i,U}} = -\frac{U^{u_i}(u_i, s)}{U^{u_i u_i}(u_i, s)}. \tag{A.1}$$

To allow for an exogenous change in the slope, we define

$$\phi \equiv -\text{MRTS}_{u_i,U} = U^{u_i}(u_i, s), \tag{A.2}$$

We want to estimate the effect on U of an exogenous change in the slope ϕ. First, we compute the effect of a change in ϕ on u_i. By total differentiation of $U^{u_i}(u_i, s) = \phi$ with respect to ϕ, we have

$$\frac{dU^{u_i}(u_i, s)}{d\phi} = \frac{d\phi}{d\phi}. \tag{A.3}$$

By the chain rule,

$$U^{u_i u_i}(u_i, s) \frac{du_i}{d\phi} = 1. \tag{A.4}$$

Thus,

$$\frac{du_i}{d\phi} = \frac{1}{U^{u_i u_i}(u_i, s)}. \tag{A.5}$$

Now to compute the effect on U, we apply the chain rule again:

$$\frac{dU}{d\phi} = \frac{d}{d\phi} U(u_i, s) \tag{A.6}$$

$$= U^{u_i}(u_i, s)\frac{du_i}{d\phi} \tag{A.7}$$

$$= \frac{U^{u_i}(u_i, s)}{U^{u_i u_i}(u_i, s)}. \tag{A.8}$$

Since $\mathrm{MRTS}_{u_i,U} = -\phi$, it follows that

$$\frac{dU}{d\mathrm{MRTS}_{u_i,U}} = -\frac{U^{u_i}(u_i, s)}{U^{u_i u_i}(u_i, s)}. \tag{A.9}$$

Back to the initial equation:

$$\frac{dU}{da} = \frac{dU}{dMRTS_{u_i,U}}\frac{dMRTS_{u_i,U}}{da} = \frac{U^{u_i}(u_i, s)}{U^{u_i u_i}(u_i, s)}\frac{1}{a^2} > 0 \tag{A.10}$$

Here we have shown mathematically the first component of Proposition 1.1: that $\frac{dU}{da}$ is always positive: more altruism always leads to more social welfare.

Returning to the sign of $\frac{dU^2}{dads}$, it will thus depend on whether a change in s increases or decreases $\frac{U^{u_i}(u_i,s)}{U^{u_i u_i}(u_i,s)}$. As expected, the magnitude is increasing in the absolute value of $U^{u_i}(u_i, s)$ and decreasing in the absolute value of $U^{u_i u_i}(u_i, s)$ (the lower the curvature, the larger the counterclockwise move over the g frontier ϕ when a increases and the slope ϕ decreases. however, as mentioned in the main text, without further assumptions, the effect of a change in s and the slope and curvature cannot be signed.

To take the result a step further, we estimate the cross-derivative:

$$\frac{d^2U}{ds\,da} = \frac{d}{ds}\left(\frac{dU}{da}\right) = \frac{1}{a^2}\frac{d}{ds}\left[\frac{U^{u_i}(u_i, s)}{U^{u_i u_i}(u_i, s)}\right].$$

Using the quotient rule,

$$\frac{d^2U}{ds\,da} = \frac{U^{u_i s}(u_i, s)U^{u_i u_i}(u_i, s) - U^{u_i}(u_i, s)U^{u_i u_i s}(u_i, s)}{a^2\left[U^{u_i u_i}(u_i, s)\right]^2}$$

Hence, consistent with the preceding analysis, the sign of $\frac{d^2U}{ds\,da}$ depends on the sign and magnitude of $U^{u_i s}$ (how the slope of $U(u_i, \bar{s})$ varies when s increases) and $U^{u_i u_i s}$ (how the curvature of $U(u_i, \bar{s})$ varies when s increases). However, their signs depend on the shape of

the flexible irregular cone $R(s)$ at that point, which the model allows to be flexible, hence without further assumption, $\frac{dU^2}{dads}$ cannot be signed.

A.2. *Ellipsoidal example to confirm Proposition 1.1*

Using the fact that one counter-example is sufficient to show that an effect cannot be signed, we illustrate the sign ambiguity with a two-dimensional analogue of the ellipsoidal achievement set. Let $x = (u_i, U)^\top$ and define

$$\Sigma(\rho) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad |\rho| < 1.$$

For $s \geq 0$, consider

$$R(s) = \left\{ x \in \mathbb{R}^2 : x^\top \Sigma(\rho(s))^{-1} x \leq \kappa(s)^2 \right\}, \qquad \kappa(s) > 0.$$

Each $R(s)$ is a strictly convex ellipsoid. Its upper boundary is

$$U(u_i, s) = \rho(s) u_i + \sqrt{1 - \rho(s)^2} \sqrt{\kappa(s)^2 - u_i^2}.$$

On the relevant branch this frontier is decreasing in u_i, and

$$U^{u_i u_i}(u_i, s) = -\frac{\sqrt{1 - \rho(s)^2}\, \kappa(s)^2}{\left(\kappa(s)^2 - u_i^2\right)^{3/2}} < 0.$$

Thus the frontier is strictly concave, and the feasible set is convex.

The agent maximizes

$$v_i = (1 - a) u_i + a U.$$

Let $w(a) = (1 - a, a)^\top$. The solution is

$$x^*(a, s) = \kappa(s) \frac{\Sigma(\rho(s)) w(a)}{\sqrt{w(a)^\top \Sigma(\rho(s)) w(a)}}.$$

Therefore,

$$U^*(a, s) = \kappa(s) \frac{a + \rho(s)(1 - a)}{\sqrt{(1 - a)^2 + a^2 + 2\rho(s) a (1 - a)}}.$$

Define

$$F(a,\rho) \equiv \frac{a+\rho(1-a)}{\sqrt{(1-a)^2+a^2+2\rho a(1-a)}}.$$

Then

$$U^*(a,s) = \kappa(s)F(a,\rho(s)).$$

The direct effect of altruism is positive:

$$F_a(a,\rho) = \frac{(1-a)(1-\rho^2)}{\left[(1-a)^2+a^2+2\rho a(1-a)\right]^{3/2}} > 0,$$

so

$$\frac{\partial U^*}{\partial a} = \kappa(s)F_a(a,\rho(s)) > 0.$$

Now let

$$\kappa(s) = e^s.$$

The scale of the ellipsoid therefore increases with s. The parameter $\rho(s)$ controls the direction of the expansion: positive ρ aligns private utility u_i and social welfare U, while negative ρ makes them trade off more strongly.

Figure 6 reports two cases. In Case A,

$$a = \frac{1}{2}, \qquad \rho(s) = \tanh(s).$$

Skills expand the feasible set and improve alignment. The selected point moves upward in U, so

$$\frac{\partial U^*}{\partial s} > 0.$$

However, the marginal effect of altruism falls with s, so

$$\frac{\partial^2 U^*}{\partial s\,\partial a} < 0.$$

In Case B,

$$a = \frac{1}{3}, \qquad \rho(s) = -\tanh(s).$$

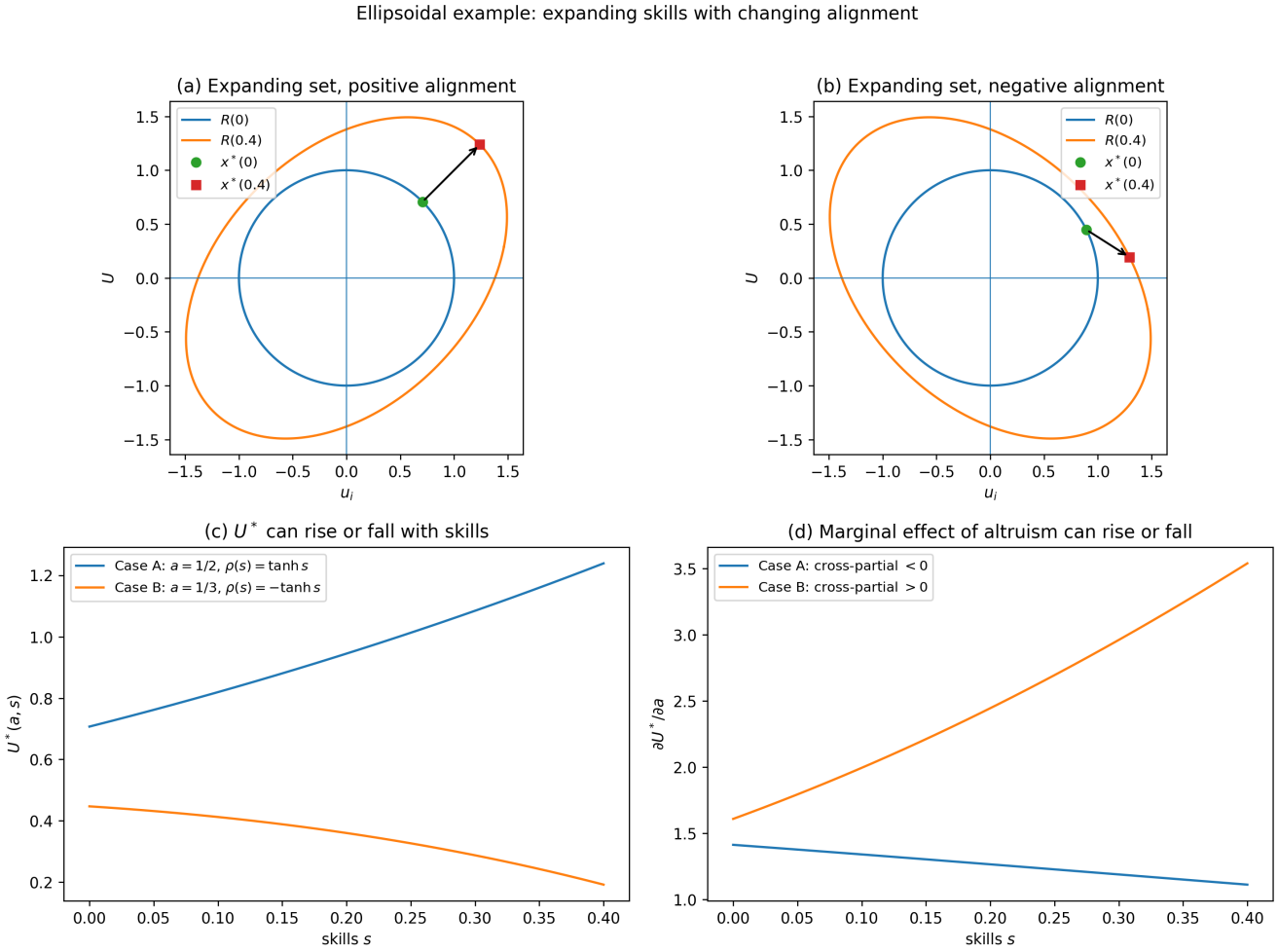


FIGURE 6.—Ellipsoidal example. Panels (a)–(b) show expanding feasible sets. In Case A, skills improve alignment and the optimum moves upward in U. In Case B, skills worsen alignment and the optimum moves toward higher u_i but lower U. Panel (c) shows that $U^*(a, s)$ can rise or fall with skills. Panel (d) shows that the marginal effect of altruism, $\partial U^*/\partial a$, can also rise or fall with skills. Hence both $\partial U^*/\partial s$ and $\partial^2 U^*/\partial s\,\partial a$ are not signed in general.

Skills again expand the feasible set, but now they worsen alignment. The selected point moves toward higher u_i and lower U, so

$$\frac{\partial U^*}{\partial s} < 0.$$

At the same time, the marginal effect of altruism rises with s, implying

$$\frac{\partial^2 U^*}{\partial s\,\partial a} > 0.$$

The example therefore shows that, even when $R(s)$ expands and remains convex, neither

$$\frac{\partial U^*}{\partial s} \qquad \text{nor} \qquad \frac{\partial^2 U^*}{\partial s\,\partial a}$$

is signed under the maintained assumptions.

APPENDIX B: SECTION WITH ENDOGENOUS TECHNICAL PROGRESS

B.1. *Proof of Lemma 2.2:* $\operatorname{sgn}\left(\frac{d\,\mathrm{MRS}_{u_i,U}}{ds}\right) = -\operatorname{sgn}\left(\frac{dr_s}{d\theta}\right)$

By the definition of polar coordinates: $u_i = r(\theta, s, \eta)\cos\theta$ and $U = r(\theta, s, \eta)\sin\theta$, along $ds = 0$, thus allowing us to rewrite the MRS:

$$\mathrm{MRS}_{u_i,U} = \left.\frac{dU}{du_i}\right|_{ds=0} = \frac{\sin\theta\, r_\theta(\theta, s, \eta) + r(\theta, s, \eta)\cos\theta}{\cos\theta\, r_\theta(\theta, s, \eta) - r(\theta, s, \eta)\sin\theta}.$$

Define

$$D(\theta, s, \eta) := \cos\theta\, r_\theta - r\sin\theta.$$

Derivative in s (using $r_{s\theta}$). Assume r is C^2 so mixed partials commute: $r_{\theta s} = r_{s\theta} = \dfrac{dr_s}{d\theta}$, where $r_s := \partial r/\partial s$. Then, holding θ and η fixed,

$$\frac{d\,\mathrm{MRS}_{u_i,U}}{ds} = \frac{r_\theta\, r_s - r\, r_{\theta s}}{D^2} = \frac{r_\theta\, r_s - r\,\dfrac{dr_s}{d\theta}}{D^2}.$$

Fixing r, then

$$\frac{d\,\mathrm{MRS}_{u_i,U}}{ds} = -\frac{r}{D^2}\,\frac{dr_s}{d\theta}.$$

Hence, for $r > 0$ and $D \neq 0$,

$$\operatorname{sgn}\left(\frac{d\,\mathrm{MRS}_{u_i,U}}{ds}\right) = -\operatorname{sgn}\left(\frac{dr_s}{d\theta}\right).$$

APPENDIX C: SECTION WITH INSTITUTIONS

C.1. *Proof of Lemma 3.1*

For any utility level $\tau \in \mathbb{R}$, define its *upper–contour set:*

$$P(\tau) \equiv \left\{ (u_i, U) \in R(s) : v_i (u_i, U) \geq \tau \right\},$$

and its relative area:

$$\Phi(\tau) \equiv \frac{\mathbb{A}(P(\tau))}{\mathbb{A}(R(s))} \in [0, 1]$$

$P(\tau)$ is the intersection of $R(s)$ with the open half–plane $\{v_i \geq \tau\}$, excluding the point (u_i, U), indicating the set that needs to be prohibited to enforce an action that generates utility v_i.

The function Φ gives the share of the area R(s) that would need to be prohibited to enforce a decision (u_i, U) such that $v_i (u_i, U) = \tau$. $\Phi(\tau)$ is continuous and strictly decreasing and varies from 0 to enforce the highest v_i in $R(s)$, because it does not require prohibiting any set since it is already agent i's preferred action, to 1 for the lowest v_i in $R(s)$ because enforcing this action requires prohibiting the entire set. Formally, $\lim_{\tau \uparrow \sup v_i} \Phi(\tau) = 0$ and $\lim_{\tau \downarrow \inf v_i} \Phi(\tau) = 1$. Thus, for each $I \in (0, 1)$ there exists a unique $\tau(I)$ satisfying $\Phi(\tau(I)) = I$. Hence for a given institutional level I, $\tau(I)$ represents the minimum level of v_i that the principal can enforce. We can now define the enforceable area $\hat{P} \equiv P(\tau(I))$ which represents the set of values (u_i, U) that the principal can enforce, located on the upper right of the slope $-\frac{1-a}{a}$ of which each points ensures utility $v_i = \tau(I)$ (as illustrated in Figure 4, Panel B).

$$\widehat{P} \equiv \{(u_i, U) \in R(s) : \ v_i(u_i, U) \geq \tau(I)\},$$

Including the feasible pairs (u_i, U) that generate a valuation $v_i \geq \tau(I)$ (Section 3.3). The principal can enforce any single point in $\widehat{P}$ (by banning all other points weakly preferred by agent i).

Equivalently, with capacity I, a point $(u_i, U) \in R(s)$ is enforceable if and only if $v_i(u_i, U) \geq \tau(I)$. We have shown that $\hat{P}$ represents the enforceable area. Hence the Principal's problem is reduced to selecting the solution $(u_i^I, U^I) \in \hat{P}$ that maximizes $v_p(u_i, U)$:

$$(u_i^I, U^I) = arg \max_{(u_i,U)\in\widehat{P}} v_p(u_p, U, a_p)$$

Once this point is identified, the principal can enforce it by removing all the points in R(s) that agent i would weakly prefer over (u_i^I, U^I):

$$P^* = \{(u_i, U) \in \widehat{P} : v_i(u_i, U) \geq v_i(u_i, U)^*\} \setminus \{(u_i, U)^*\}$$

Then the agent plays and selects $(u_i^I, U^I) = \underset{(u_i,U)\in R(s)\setminus P^*}{argmax} v_i(u_i, U)$, where (u_i^I, U^I) is his only way to reach a utility equal or higher to $v_i\left(u_i^I, U^I\right)$:[12]

The concavity of $v_p(u_i, U)$ and compactness of $R(s)$ ensure that (u_i^I, U^I) exists and is unique.

C.2. *Proof of Proposition 3.2*

We provide a proof by contradiction of Proposition 3.2, according to which, in the presence of institutions, social welfare is increasing in the altruism of the principal. The intuition is straightforwards: when the principal puts more weight on the common good U, she cannot end up implementing a point with a lower U, otherwise, the earlier, higher U point would beat the new one once U is more heavily weighted.

Fix the institutional feasible set $\widehat{P}$. For each $a_p \in [0, 1]$, let

$$v_p(u_i, U; a_p) \equiv (1 - a_p)u_p(u_i, U) + a_p U$$

denote the principal's objective, and let

$$\left(u_i(a_p), U(a_p)\right) \in \arg \max_{(u_i,U)\in\widehat{P}} v_p(u_i, U; a_p)$$

[12]Note that removing tie-breaking with same v_i as $v_i(u_i, U)^*$ is not strictly necesary, but it makes sense to assume that the principal prefers removing it, since it comes at no cost and removes actions with each the agent is indifferent, but that reduce the princpial's valuation compared to $v_p(u_i, U)^*$.

be an optimal institutional choice. We assume that a maximizer exists for each a_p under consideration.

Take $0 \leq a_p' < a_p" \leq 1$. We show that

$$U(a_p") \geq U(a_p').$$

Suppose, toward a contradiction, that

$$U(a_p") < U(a_p').$$

Assuming an increase in a_p define the post-minus-pre differences

$$\Delta U \equiv U(a_p") - U(a_p') < 0,$$

and

$$\Delta u_p \equiv u_p\big(u_i(a_p"), U(a_p")\big) - u_p\big(u_i(a_p'), U(a_p')\big).$$

The sign of Δu_p is unrestricted.

Optimality of $\big(u_i(a_p'), U(a_p')\big)$ at weight a_p' implies

$$\begin{aligned}(1-a_p')u_p\big(u_i(a_p'), U(a_p')\big) + a_p' U(a_p') \\ \geq (1-a_p')u_p\big(u_i(a_p"), U(a_p")\big) + a_p' U(a_p").\end{aligned}$$

Equivalently,

$$(1-a_p')\Delta u_p + a_p' \Delta U \leq 0.$$

Since $a_p' < a_p" \leq 1$, we have $a_p' < 1$. Hence

$$\Delta u_p \leq -\frac{a_p'}{1-a_p'}\Delta U.$$

Now compare the new and old choices under the larger altruism weight $a_p"$. The difference between the principal's objective at the new choice and at the old choice is

$$\begin{aligned}v_p\big(u_i(a_p"), U(a_p"); a_p"\big) - v_p\big(u_i(a_p'), U(a_p'); a_p"\big) \\ = (1-a_p")\Delta u_p + a_p" \Delta U.\end{aligned}$$

Using the bound on Δu_p, we obtain

$$\begin{aligned} & v_p\big(u_i(a_p"), U(a_p"); a_p"\big) - v_p\big(u_i(a_p'), U(a_p'); a_p"\big) \\ & \quad \leq (1 - a_p") \left(-\frac{a_p'}{1 - a_p'} \Delta U \right) + a_p" \Delta U \\ & \quad = \Delta U \left[a_p" - \frac{a_p'(1 - a_p")}{1 - a_p'} \right] \\ & \quad = \Delta U \frac{a_p" - a_p'}{1 - a_p'} < 0. \end{aligned}$$

The last inequality follows because

$$\Delta U < 0, \qquad a_p" - a_p' > 0, \qquad 1 - a_p' > 0.$$

Thus, under weight $a_p"$, the new choice $\big(u_i(a_p"), U(a_p")\big)$ gives the principal a strictly lower objective value than the old choice $\big(u_i(a_p'), U(a_p')\big)$. This contradicts the optimality of $\big(u_i(a_p"), U(a_p")\big)$ at $a_p"$. Therefore,

$$U(a_p") \geq U(a_p').$$

Since $a_p' < a_p"$ were arbitrary, $U(a_p)$ is weakly increasing in a_p. If $U(a_p)$ is differentiable at a given value of a_p, this implies

$$\frac{dU}{da_p} \geq 0.$$

The proof uses only optimality at each parameter value over the fixed feasible set $\widehat{P}$. It does not rely on interiority, differentiability, or curvature of the feasible set, and therefore also covers solutions on the boundary of $\widehat{P}$.

Strict positivity can be shown when the solution is interior.

APPENDIX D: APPENDIX OF THE UNIVERSALISM SECTION

Information to allow the replication of Figure 5: the feasible set is defined as $R(s) = \Big\{(U_G, U_i, U) : \; u_i^2 + U_G^2 + U^2 + 2\rho U_G U \leq s^2 \Big\}$, resulting in an ellipsoidal achievement set, strictly convex for $|\rho| < 1$. The parameter ρ governs the alignment between U_G and U in

the feasible set, which is positive when $\rho < 0$ (the ellipsoid is elongated along the $U_G = U$) and negative when $\rho > 0$. As parameters, both panels use $s = 1$, $a^o = 1/3$ and $a^g = 3/5$. Panel A of Figure 5 has $\rho = -1/2$ (positive alignment between U_G and U), and Panel B has $\rho = 1/2$ (negative alignment between U_G and U).

D.1. *Proof of Proposition 4.1*

D.1.1. *Frontier representation and a curvature restriction*

Let $U = \bar{U}(U_i, U_G; s)$ be the upper frontier induced by $R(s)$:

$$\bar{U}(U_i, U_G; s) \equiv \max\{U : \ (U_i, U_G, U) \in R(s)\}.$$

Strict convexity and smoothness of $R(s)$ imply that $\bar{U}$ is twice continuously differentiable and strictly concave in (U_i, U_G) on the relevant domain. Write $x \equiv (U_i, U_G)$ and denote by $H(x; s)$ the Hessian of $\bar{U}$ with respect to x.

We impose the following mild restriction (standard in comparative statics on concave frontiers).

ASSUMPTION D.1—Weak complementarity on the frontier: On the relevant portion of the frontier,

$$\bar{U}_{U_i U_G}(U_i, U_G; s) \geq 0.$$

D.2. *First-order conditions and implicit-function derivatives*

Consider the reduced problem

$$\max_{x=(U_i,U_G)} (1 - a^g)U_i + (a^g - a^o)U_G + a^o\, \bar{U}(U_i, U_G; s).$$

Define the first-order conditions as

$$F_1(x; a^g, a^o) \equiv (1 - a^g) + a^o\, \bar{U}_{U_i}(x; s) = 0, \tag{D.1}$$

$$F_2(x; a^g, a^o) \equiv (a^g - a^o) + a^o \bar{U}_{U_G}(x; s) = 0. \tag{D.2}$$

At an interior optimum $x^*(a^g, a^o; s)$ we have

$$\bar{U}_{U_i}(x^*; s) = -\frac{1 - a^g}{a^o} < 0, \qquad \bar{U}_{U_G}(x^*; s) = -\frac{a^g - a^o}{a^o} \leq 0, \qquad \bar{U}_{U_G}(x^*; s) - 1 = -\frac{a^g}{a^o} < 0. \tag{D.3}$$

Moreover, the Jacobian of (F_1, F_2) with respect to x is

$$D_x F(x^*; a^g, a^o) = a^o \, H(x^*; s),$$

which is invertible because $H(x^*; s)$ is negative definite (strict concavity). Hence the implicit function theorem applies and x^* is differentiable in (a^g, a^o).

D.3. *Part (ii): $\partial U^*/\partial a^o > 0$ under Assumption D.1*

Differentiate (D.1)–(D.2) with respect to a^o, holding x fixed:

$$\partial_{a^o} F_1(x; a^g, a^o) = \bar{U}_{U_i}(x; s), \qquad \partial_{a^o} F_2(x; a^g, a^o) = \bar{U}_{U_G}(x; s) - 1.$$

The implicit function theorem yields

$$\frac{dx^*}{da^o} = -\big(D_x F(x^*; a^g, a^o)\big)^{-1} \begin{pmatrix} \bar{U}_{U_i}(x^*; s) \\ \bar{U}_{U_G}(x^*; s) - 1 \end{pmatrix} = -\frac{1}{a^o} \, H(x^*; s)^{-1} \begin{pmatrix} \bar{U}_{U_i}(x^*; s) \\ \bar{U}_{U_G}(x^*; s) - 1 \end{pmatrix}.$$

Since $U^* = \bar{U}(x^*; s)$, the chain rule implies

$$\frac{\partial U^*}{\partial a^o} = \nabla_x \bar{U}(x^*; s)^\top \frac{dx^*}{da^o}.$$

Using the explicit inverse of a 2×2 Hessian (and letting $D(x^*) \equiv \bar{U}_{U_i U_i} \bar{U}_{U_G U_G} - \bar{U}^2_{U_i U_G} > 0$), one obtains the closed-form expression

$$\frac{\partial U^*}{\partial a^o} = -\frac{1}{a^o\, D(x^*)}\Big[\bar{U}_{U_G U_G}(x^*;s)\,\bar{U}_{U_i}(x^*;s)^2 - \bar{U}_{U_i U_G}(x^*;s)\,\bar{U}_{U_i}(x^*;s)\big(2\bar{U}_{U_G}(x^*;s)-1\big) + \bar{U}_{U_i U_i}(x^*;s)\,\bar{U}_{U}$$

(D.4)

Strict concavity implies $\bar{U}_{U_iU_i} < 0$ and $\bar{U}_{U_GU_G} < 0$. From (D.3), $\bar{U}^2_{U_i} > 0$ and $\bar{U}_{U_G}(\bar{U}_{U_G} - 1) > 0$, so the first and third terms in brackets are strictly negative. Moreover, (D.3) implies $\bar{U}_{U_i} < 0$ and $2\bar{U}_{U_G} - 1 < 0$, hence $\bar{U}_{U_i}(2\bar{U}_{U_G} - 1) > 0$. Under Assumption D.1, $\bar{U}_{U_iU_G} \geq 0$, so the middle term in (D.4) is weakly negative. Therefore the bracketed expression in (D.4) is strictly negative, implying

$$\frac{\partial U^*}{\partial a^o} > 0.$$

This proves Proposition 4.1 (ii).

D.4. *Part (i): $\partial U^*/\partial a^g$ is not signable (ellipsoid example)*

We establish indeterminacy by constructing a family of strictly convex achievement sets for which the sign of $\partial U^*/\partial a^g$ changes with the *alignment* between U_G and U (as in Figure 5).

Ellipsoidal achievement set with an alignment parameter. Fix $|\rho| < 1$ and consider the ellipsoid

$$R_\rho(s) \equiv \Big\{x \in \mathbb{R}^3 :\; x^\top \Sigma(\rho)^{-1} x \leq s^2\Big\}, \qquad \Sigma(\rho) \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}. \tag{D.5}$$

The off-diagonal term ρ controls whether directions that increase U_G tend to also increase U (positive ρ) or instead trade off against U (negative ρ). This is precisely the "alignment" depicted in Figure 5.

Maximizing the linear objective (4.1) over (D.5) yields the closed-form optimizer

$$x^*(a^g, a^o; s) = s \frac{\Sigma(\rho)\, w(a^g, a^o)}{\sqrt{w(a^g, a^o)^\top \Sigma(\rho)\, w(a^g, a^o)}}, \qquad w(a^g, a^o) = (1 - a^g,\ a^g - a^o,\ a^o). \tag{D.6}$$

Therefore,

$$U^*(a^g, a^o; s) = s \frac{a^o + \rho\,(a^g - a^o)}{\sqrt{(1 - a^g)^2 + (a^g - a^o)^2 + (a^o)^2 + 2\rho\, a^o (a^g - a^o)}}. \tag{D.7}$$

Sign reversal with alignment. Differentiating (D.7) with respect to a^g gives an expression whose sign depends on ρ (and on (a^g, a^o)). A particularly transparent case is "universalism at the margin", $a^g = a^o \equiv a$. Then (D.7) simplifies and the derivative has the closed form

$$\left.\frac{\partial U^*}{\partial a^g}\right|_{a^g = a^o = a} = \frac{s(1-a)\big(a + \rho(1-a)\big)}{\big(a^2 + (1-a)^2\big)^{3/2}}. \tag{D.8}$$

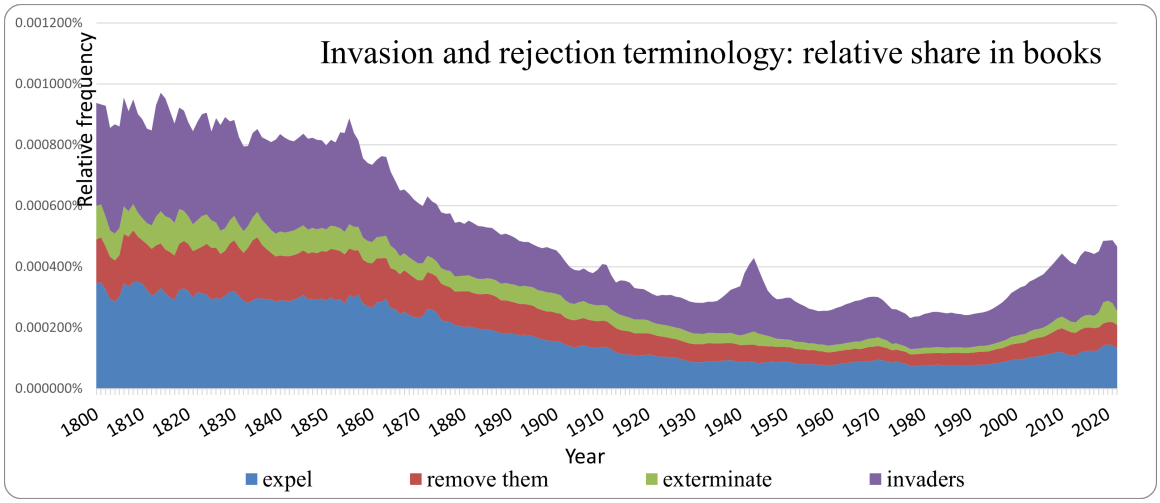
Hence the sign is governed by the alignment term $a + \rho(1 - a)$. In particular, for the value used in the figures ($a^o = 1/3$), (D.8) implies

$$\left.\operatorname{sign}\left(\frac{\partial U^*}{\partial a^g}\right)\right|_{a^g = a^o = 1/3} = \operatorname{sign}\left(\rho + \frac{1}{2}\right).$$

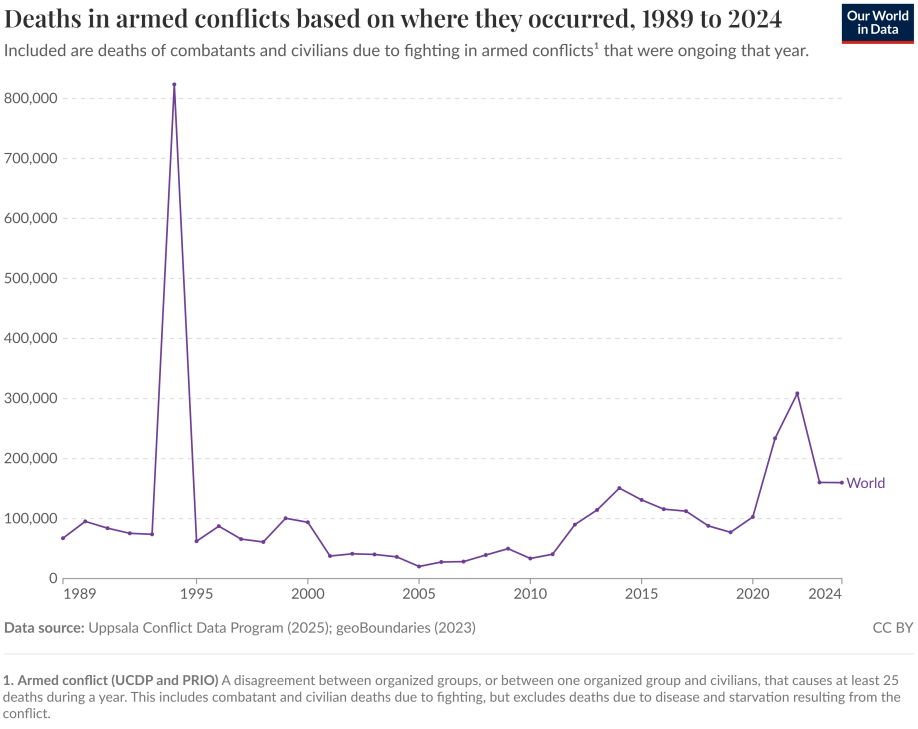
Thus, for sufficiently positive alignment (e.g. $\rho = 1/2$ as in Figure 5, Panel A) the derivative is positive, while for sufficiently negative alignment (e.g. $\rho < -1/2$) the derivative is negative.

Moreover, even at the knife-edge $\rho = -1/2$, the derivative becomes negative for a^g slightly above a^o, consistent with Figure 5, Panel B (which fixes $a^o = 1/3$ and uses $a^g \approx 0.34$): plugging $a^o = 1/3$, $\rho = -1/2$, and $a^g = 0.34$ into (D.7) yields $\partial U^*/\partial a^g < 0$. Therefore, $\partial U^*/\partial a^g$ cannot be signed in general. This proves Proposition 4.1 (i). *Q.E.D.*

APPENDIX E: APPENDIX FIGURES ON INVASION TERMINOLOGY AND ARMED CONFLICT DEATHS



Panel A. Invasion and rejection terminology: relative share in books



Panel B. Deaths in armed conflicts based on where they occurred, 1989 to 2024

FIGURE 7.—Invasion terminology and deaths in armed conflicts.