Cooperation as Bulwark: Evolutionary Game Theory and the Internal Institutional Structure of States

L.C.R. Patell

Abstract

This paper challenges the view that advanced AI will inevitably drive states toward low-welfare, high-automation institutional equilibria. After formalizing the basic evolutionary argument, I demonstrate that it omits the possibility that interstate cooperation can sustain a threshold level of welfare. To sustain that possibility, I develop models of three increasingly realistic evolutionary games. After isolating a condition under which cooperation – and hence welfare – is stable in the long run, I canvas several realist objections to the cooperative argument. The paper concludes with several tentative policy recommendations based on the formal models. Far from eroding human welfare, AI can entrench it – provided states embed cooperation at the heart of their interstate strategy.

Introduction

This paper aims to extend and refine the argument in MacInnes et al. (2024). The argument has been used to justify the claim that evolutionary dynamics, combined with the advent of advanced AI, will lead states to drift towards low-welfare-providing institutional arrangements.

Section 1 schematizes that argument and demonstrates that it is invalid without further detail. Section 2 works through three increasingly complex models of interstate competition. Section 3 canvases, and responds to, realist objections to the models. Section 4 crystallizes the argument into policy recommendations and carves out directions for future work.

1 The Basic Evolutionary Argument

The basic evolutionary argument runs as follows.[1]

1. AI enables new institutional arrangements that provide low welfare but are highly competitive.

2. Evolutionary dynamics entail that the population of states will drift towards having competitive traits.

3. Therefore, AI will lead states to drift towards low-welfare-providing institutional arrangements.

This argument is seductive – but invalid. It is not sufficient for these new institutional arrangements to be highly competitive: they must be more competitive than institutional arrangements that provide higher levels of welfare.

The relevant suppressed premiss is:

> 1*. AI enables new institutional arrangements that provide low welfare and are more competitive than states that provide higher welfare.

[1]This argument constitutes a formalized version of the arguments in Assadi (2023) and Drago and Laine (2025). Strictly speaking, MacInnes et al. is entirely consistent with the cooperative models below – but their argument lends itself towards the basic evolutionary argument by omitting cooperative dynamics.

This premiss might seem trivially true. After all, by providing higher welfare, states necessarily take resources away from automated industries. Hence, the reasoning goes, states that do so will be less competitive than states that ruthlessly pursue automation.

Yet the competitiveness of different states depends crucially on the selection mechanism; through a game-theoretic lens, it depends on the structure of the game that is being played. As MacInnes et al. write, selection means that "the units within the population are involved in a competition for existence with one another. The traits of a unit differentially impact its competitive performance; its 'fitness' ".

So what determines a state's fitness on the international stage? MacInnes et al. state that "anarchy imposes competitive pressures on states", but they do not provide a precise anarchic game through which states compete. Instead, they provide two hints concerning the nature of the competitive pressure. First, they invoke Walz's structural realism – including the idea that "war made the state, and the state made war" – and Gourevitch's claim that "war is like the market: it punishes some forms of organization and rewards others".[2] Second, they mention that some technologies provide competitiveness through economic benefits. While they never make the exact structure of the game being played more precise, it is clear that MacInnes et al. are primarily concerned with military and economic competitiveness.

Here lies a lacuna in the basic argument: it omits the possibility that cooperation between states may falsify 1*. To see this, suppose that some states provide at least a welfare level $W*$ and are willing to cooperate on the international stage with all and only those states that provide at least $W*$. Then, depending on the relative payoffs and initial conditions, providing at least $W*$ may be more competitive – higher fitness – than higher-automation, lower-welfare institutional structures even though the latter would dominate in a two-player game. Premiss 1* would turn out false. This claim is consistent with the interstate selection mechanism being based purely on economic and military competition: military alliances, trade agreements, and compute pooling provide ample opportunities for fitness-enhancing cooperation.

[2]Waltz (1979); Gourevitch (1978).

Once we countenance cooperative behavior, we cannot discern its competitiveness without a better model of the game that states are playing. Without more precise modeling, therefore, passing judgment on the claim that cooperation would sustain a minimum welfare threshold is premature.

I turn now to discharging that task.

2 The Anarchic Game

2.1 A Simple Model

Consider the canonical evolutionary game from Maynard Smith and Price (1973). Let there be two types of states that employ the following strategies respectively:

> **High-welfare cooperators (H)**, which meet or exceed the welfare threshold $W*$ and cooperate with other H-types.
>
> **Low-welfare competitors (L)**, which provide less than $W*$ and always defect (compete).

In each period, two states are randomly matched and receive net per-period payoffs, where:

> R = payoff when two H's meet (cooperative surplus);
>
> T = payoff to L when matched with H (exploiting or victorious payoff);
>
> S = payoff to H when matched with L (exploited or defeated payoff);
>
> P = payoff when two L's meet (pure competition).

If the proportion of H-types in the population is x, then the expected payoffs for each type are:

$$\pi_H = xR + (1-x)S$$
$$\pi_L = xT + (1-x)P$$

The population share of H-types evolves according to the replicator equation:

$$\frac{\partial x}{\partial t} = x(1-x)(\pi_H - \pi_L)$$

If $R > T$ and $S > P$, then the all-H state ($x = 1$) is the unique attractor. More plausibly, if $R > T$ but $P > S$, then there are two basins of attraction: one at $x = 1$ and the other at $x = 0$. In this case, the equilibrium is determined by the initial proportion of cooperators.

In addition to convergence, we are interested in whether or not H is an evolutionarily stable strategy (ESS).

Definition (ESS).[3] A strategy H is evolutionarily stable if for every distinct strategy L, either:

1. $\pi(H, H) > \pi(L, H)$, or
2. $\pi(H, H) = \pi(L, H)$ and $\pi(H, L) > \pi(L, L)$.

An evolutionarily stable strategy is locally stable: a sufficiently large cooperative bloc can hold off defectors. In order to repudiate the conclusion of the basic evolutionary argument, all that is necessary is that the cooperative, threshold-welfare strategy is evolutionarily stable – we need not obtain the global convergence result.

While this is an extremely toy model, it demonstrates two claims:

1. It is *possible* for the payoff structure of a game to imply that cooperation is more competitive in the long run, even when less cooperative states can exploit more cooperative states in dyadic interactions.
2. The *initial conditions* of the game can determine path-dependent equilibria.

While Drago and Laine canvas "coordination" as a possible solution to the problem of long-run drift, they are referring to a state where "aggregative pressures stop driving history"; by contrast, the cooperative mechanisms at play in the foregoing model are endogenous to interstate competition. This argument further demonstrates that the requirements for "strong interstate coordination" may be weaker – and hence more feasibly accomplished – than Assadi (2023) supposes.

[3] Maynard Smith and Price (1973)

2.2 A Structured Model

We have seen that a toy model can sustain cooperation in the long run. That game's basic structure can be transposed into our original setting by adding more detail concerning the internal structure of states and the payoffs they receive from interaction.

In our new model, each state i chooses an institutional profile:

Welfare provision w_i is an element of $[0, 1]$.

There is a trade-off between welfare provision and automation, such that:

Automation $a_i = g(w_i)$.

I leave $g(w_i)$ unspecified, but note that it need not be linear.

As before, states are randomly paired in each time period. Now, however, two states i and j:

1. Cooperate if and only if $w_i >= W*$ and $w_j > W*$, or
2. Compete otherwise.

The payoffs from each match are given by the following equations:

$$\begin{aligned} \pi_i(w_i, w_j) &= \gamma(\alpha g(w_i) + (1-\alpha)g(w_j)) \text{ if Cooperate,} \\ &= g(w_i) - \delta g(w_j) \text{ otherwise.} \end{aligned}$$

When two high-welfare states meet, they cooperate. I model cooperation as the states jointly deploying their automated resources. Rather than splitting the resulting surplus equally, I assume that each state i captures a fraction alpha of its own automation and a fraction $(1 - \alpha)$ of its partners.

The scalar term γ calibrates the extent to which pooling resources magnifies competitiveness. $\gamma = 1$ means that there is no synergy: pooling simply aggregates capacity. As γ increases past 1, there are economies of scale or network effects; at higher levels, we might think of gamma as indicating superlinear gains – if, for example, pooling resources can unlock entirely new AI capabilities.

If either state provides lower than $W*$, they compete. Each side gains a base payoff of its own automation $a_i = g(w_i)$, but then suffers in proportion to its opponent's automation. The parameter δ tunes the intensity of zero-sum competition. If $\delta = 1$, then we have a symmetric zero sum contest. When $\delta < 1$, absolute gains matter more than relative levels of strength. Finally, if $\delta > 1$, then conflict is intense: relative strength is more important than absolute levels of strength.

As before, supplying $W*$ will be an evolutionarily stable strategy if, and only if, the payoff from cooperation between $W*$-providers exceeds the payoff that adversaries gain from competition with $W*$-providers.

For simplicity, let us suppose that supplying $w_i = 0$ is a feasible institutional profile for a state to implement.[4] Then, $W*$ is an evolutionarily stable institutional profile if, and only if:

$$\pi_i(W*, W*) > \pi_i(0, W*)$$

We know that:

$$\pi_i(W*, W*) = \gamma g(W*)$$

$$\pi_i(0, W*) = g(0) - \delta g(W*)$$

Hence, $W*$ is an evolutionarily stable institutional profile if, and only if:

$$\gamma g(W*) > g(0) - \delta g(W*)$$

$$(\gamma + \delta) g(W*) > g(0)$$

We obtain a continuous condition that is analogous to the simple two-player case. The condition presents twin upshots.

First is that values of these parameters can be empirically investigated: we can investigate the returns to pooling resources, the intensity of competition, and the functional form of the tradeoff between providing welfare and

[4]Below $W*$, decreasing w_i further makes a strategy monotonically more competitive, and so the fittest $w_i < W*$ is $w_i = 0$.

strengthening automation.

Second – and arguably more importantly – these parameters may be under our control. Even if historical or present interstate cooperation does not exhibit significant network effects or increasing returns to scale, we may be able to design cooperative paradigms between states that do have these properties. Moreover, we may be able to shape the competitive pressures between states that do and don't provide $W*$ – modulating the effective value of δ. While the tradeoff between welfare provision and strengthening automation may be determined by technological factors in the limit, the result above raises the tenor of demands to provide benefits more efficiently; it demonstrates that making the provision of welfare more efficient contributes to the long-run preservation of human flourishing.

As I've argued, there are conditions under which $W*$ can be preserved in the long run. To understand whether or not these conditions are likely to hold, however, we need to do more to understand the way in which automation would lead to strength on the international stage.

TODO – add reasonable bounds for gamma and delta
TODO – review continuous games in IR to see if this model is reasonable
TODO – examine how nuclear deterrents function here
TODO – investigate low welfare state collaboration

2.3 A Diachronic Model

Now suppose that each state starts with a stock of automation $a_i(0)$. For simplicity, I begin by assuming that each state starts with the same stock $a_i(0) = 1$. These stocks can grow over time: each state's automation evolves according to the equation:

$$\frac{\partial a_i}{\partial t} = -\rho a_i(t) + \eta g(w_i)\pi_i(w_i, w_j; a_i, a_j),$$

Where:

$\rho > 0$ is the depreciation rate,

$\eta > 0$ is the reinvestment efficiency,

$g(w_i)$ is a decreasing function of a state's welfare provision that determines how efficiently its per-period payoff get turned into new automation, and

π_i is the per-period payoff from interstate interactions:

$$\begin{aligned}\pi_i(t, w_i, w_j) &= \gamma(\alpha a_i(t) + (1-\alpha)a_j(t)) \text{ if Cooperate,} \\ &= a_i(t) - \delta a_j(t) \text{ otherwise.}\end{aligned}$$

Now we can investigate evolutionary stability. We begin with a population of states that play the institutional profile $W*$. To discern stability, we need to consider the effects of the entrance of a mutant state. Let

$x(t)$ = the fraction of states playing the mutant welfare level $W_m < W*$;

$1 - x(t)$ = the fraction of states playing the resident welfare level $W*$.

Each period, I assume that a state's reproductive rate is proportional to its automation growth rate r_i:

$$r_i = \frac{1}{a_i}\frac{\partial a_i}{\partial t} = -\rho + \eta\frac{g(w_i)\pi_i(w_i, w_j; a_i, a_j)}{a_i}$$

So, when a resident meets a resident:

$$\pi_R = \gamma a_R$$

$$r_R = -\rho + \eta g(W*)\gamma$$

And when a mutant meets a resident:

$$\pi_m = a_m - \delta a_R$$

$$= (1-\delta)a_R$$

$$r_m = -\rho + \eta g(W_m)(1-\delta)$$

Thus the standard replicator equation is:

$$\frac{\partial x}{\partial t} = x[r_m - \bar{r}],$$

$$\bar{r} = xr_m + (1 - x)r_R$$

Where:

$$\begin{aligned} r_m = &\text{ mutant's growth rate,} \\ r_R = &\text{ resident's growth rate.} \end{aligned}$$

When x is small, $\bar{r}$ is approximately r_R.

$$\frac{\partial x}{\partial t} = x[r_m - \bar{r}]$$

$$\frac{\partial x}{\partial t} = x[r_m - r_R]$$

$$= x[(-\rho + \eta g(W_m)(1 - \delta)) - (-\rho + \eta g(W*)\gamma)]$$

$$= x\eta[g(W_m)(1 - \delta) - g(W*)\gamma]$$

Since $\eta > 0$, if $[g(W_m)(1 - \delta) - g(W*)\gamma] < 0$, then for sufficiently small $x > 0$, $\frac{\partial x}{\partial t} < 0$. The mutant population shrinks back to 0. Rearranging yields the inequality for the worst case $W_m = 0$:

$$\gamma g(W*) > (1 - \delta)g(0)$$

By definition, an institutional profile that secures $W*$ is an evolutionarily stable strategy under these conditions.

TODO – explore under conditions when the sub-W* states can cooperate; explore under conditions when the starting proportion and welfare-provision of states varies – include one instance where we approximate existing conditions and maybe perform a computational parameter sweep]
TODO – note that $g(0) = 1$; $g(w*) > 0$; for plausible lower bounds of lambda

and delta (1 and 0.5), the inequality reduces to $g(W*) > 0.5$
TODO – note that the idea that providing welfare is a proportional cost of automation gains is an EXTREMELY strong condition – more plausibly, there is some fixed per-capita cost of providing W*, in which case the inequality will scale with -w*/a, which decreases as a increases. This result would be extremely interesting: it suggests that early growth dynamics are more critical

3 A Realist Rejoinder

There are several objections to this cooperative argument that a staunch realist might pursue. Realists do not deny that alliances are possible on the international stage. Instead, they claim that these alliances have a particular character – one that threatens to scupper the possibility of long-run cooperation. I elucidate these objections and canvas responses in turn.

3.1 The Fragility of Cooperation

Objection. Under anarchy there is no ultimate arbiter and enforcer of agreements.[5] Any cooperative scheme that diverts resources toward welfare can, therefore, be abandoned the moment a state perceives a shift in the balance of power. In our model, increases in an adversary's automation stock will tempt defectors to abandon the welfare coalition in favor of increasing its unilateral strength.

Response. *Per contra*, cooperation is self-reinforcing: cooperation is tied to continued access to pooled automation surplus. As long as coalition members switch to competing against prospective defectors as soon as their welfare provision drops, then there is an endogenous sanction that no external arbiter is needed to enforce.

3.2 Relative Gains Concerns

Objection. Absolute payoffs matter more than relative ones.[6] If two high-welfare states pool automation and share surplus, then, if one state's gains

[5] Waltz (1979).
[6] Mearsheimer (2001).

outpace the other's, the weaker party will fear that future concessions will imperil its security. States cannot commit not to exploit future power shifts.

Response. Relative gains anxieties weaken under more realistic asymmetric stock assumptions due to the surplus-split parameter α. Even if members can accrue strictly more than their peers due to greater initial stocks of automation or higher individual growth rates, then, combined with the response to objection 3.1, weaker states will lose out from defection in the long run. Technological developments may also allow for credible commitments concerning future uses of automated resources or even link them to providing certain levels of welfare.[7]

3.3 Balance of Power Logic

Objection. Alliances form only to counter a credible threat and dissolve once that threat abates.[8]

Response. Under certain starting conditions, it is plausible that there would be a persistent credible threat: namely highly automated states that underprovide welfare.

3.4 Informational Uncertainty

Objection. The model assumes that states are able to know the institutional profile of states they interact with.[9] If states cannot verify that rivals actually meet $W*$, then rational states may defect preemptively. The strategies that sustain cooperation in repeated games can break down under uncertainty.

Response. Verification need not be perfect to sustain cooperation. Even so, the development of better metrics and technical transparency measures may enable new methods of low-cost, high-trust verification.[10] As I discuss in the next section, however, this objection correctly identifies that a more precise model would involve a signaling game; without further modeling, the

[7] Assadi (2023), pp. 16-17.
[8] Waltz (1979); Glaser (1994).
[9] Fearon (1994); Fearon (1997).
[10] Assadi (2023), pp. 15-16.

realist cannot rule out that the signaling game has a cooperative evolutionarily stable strategy.

3.5 Statistical Accuracy

Objection. The replicator dynamics assume large, well-mixed populations. In reality, the number of major powers is small ($N < 10$), and the strategic interactions are highly networked.[11] Stochastic shocks can also tip the balance of power abruptly, overriding smooth evolutionary attractors. A coalition that appears stable in expectation in the statistical limit may not manifest in practice.

Response. It is correct that stochasticity and small-N dynamics mean that an ESS may not be achieved in practice. Yet this response cuts equally against the original basic evolutionary argument, blunting its force as an objection to the cooperative argument. Using smooth statistical assumptions allows us to derive policy recommendations that increase the likelihood of long-term welfare provision at or above $W*$.

3.6 Synergy Skepticism

Objection. The crucial inequalities depend on a large value of γ – the synergies or increasing returns to scale of cooperation. Yet historical alliances rarely deliver truly super-additive gains.[12]

Response. AI and emerging technology are more likely than historical technologies to exhibit significant returns to scale and network effects. While, ultimately, this question is an empirical one, it is premature to rule out values of γ above 1.

4 Policy Recommendations

The possibility of cooperation in the anarchic game precipitates a set of policy recommendations. While these recommendations require more concreteness

[11] Nowak and May (1992); Nowak et al. (2004).
[12] Axelrod (1984).

to be actionable, they set a direction for further investigation.
Simply put:

1. If a state doesn't want to drift towards low-welfare institutional structures, it should cooperate with other similar states.

2. That cooperation should be based on providing a threshold level of welfare to its citizens.

3. States that care about human welfare should craft cooperative arrangements that exhibit increasing returns to scale.

The power of this argument, as demonstrated in section 2, is that cooperation need not take the form of coordinated joint governance or sacrifices of sovereignty. Military alliances, industrial partnerships, and trade agreements may suffice.

Recall that we modeled cooperation as an evolutionary stable strategy when:

$$\gamma g(W*) > (1 - \delta)g(0)$$

In greater detail, therefore, states should:

1. **Create better metrics for human welfare.** As MacInnes et al. and Kulveit et al. (2025) demonstrate, GDP levels and liberal democratic institutions are brittle proxies for civilian welfare in the face of the institutional possibilities opened up by advanced AI. If reliable cooperation is to be based on civilian welfare, then states must be able to measure the phenomenon of interest with greater, more robust accuracy.

2. **Investigate the signaling game.** Cooperation on the basis of welfare metrics would add a new signaling element to the anarchic game. More work remains to be done to understand the dynamics of this game and how verifiable claims about civilian welfare could be made.

3. **Balance welfare against growth.** As model in section 2, a coalition of states cooperating on the basis of welfare risks being out-competed if non-cooperative states are able to leverage automation to grow significantly faster.

4. **Boost the cooperative surplus.** The more cooperating states can exploit network effects and increasing returns to scale (raising γ), the more likely $W*$ is to be secured.

5. **Be willing to compete severely with low-welfare providing states.** The greater coalition members make the penalty that defectors pay for non-cooperation (raising δ), the more likely the cooperative evolutionary mechanism is to hold. States should consider swift military and economic intervention against states whose welfare provision drops below $W*$ to preempt the compounding gains that result from increased automation.

Conclusion

States drifting towards low-welfare providing, highly automated institutional structures is not as inevitable as first meets the eye. By precisely modeling the competitive interactions between states, this paper demonstrated that cooperation can sustain a threshold level of civilian welfare in the face of advanced automation.

This result is not a foregone conclusion. As emphasized in section 4, we must preemptively strengthen both our mechanisms of interstate cooperation and our metrics for civilian welfare; without those measures in place, we are indeed likely to drift away from institutional structures that provide for human flourishing.

References

Guive Assadi. Will humanity choose its future? 2023.

R. Axelrod. *The Evolution of Cooperation*. 1984.

James D. Fearon. Domestic political audiences and the escalation of international disputes. *American Political Science Review*, 1994.

James D. Fearon. Signaling foreign policy interests: Tying hands versus sinking costs. *The Journal of Conflict Resolution*, 1997.

Charles L. Glaser. Realists as optimists: Cooperation as self-help. *International Security*, 1994.

Peter Gourevitch. The second image reversed: The international sources of domestic politics. *International Organization*, 1978.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development, 2025. URL `https://arxiv.org/abs/2501.16946`.

Morgan MacInnes, Ben Garfinkel, and Allan Dafoe. Anarchy as architect: Competitive pressure, technology, and the internal structure of states. *International Studies Quarterly*, 68(4):sqae111, 2024.

J. Maynard Smith and G.R. Price. The logic of animal conflict. *Nature*, 1973.

J.J. Mearsheimer. *The Tragedy of Great Power Politics*. W.W. Norton Company, 2001.

M.A. Nowak and R. May. Evolutionary games and spatial chaos. *Nature*, 1992.

M.A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 2004.

K.N. Waltz. *Theory of International Politics*. Addison-Wesley Publishing Company, 1979.