Persistence, Not Projection: The Case for Loop Maintenance over Longtermism

A Response to Owen Cotton-Barratt and Rose Hadshar's "What Would a Longtermist Society Look Like?" (Chapter 19 of *Essays on Longtermism*)

1 Introduction

In their contribution to Oxford University Press's *Essays on Longtermism*, Owen Cotton-Barratt and Rose Hadshar ask what society would look like if longtermist perspectives shaped whole institutions rather than just individual choices at the margin.[1] This is an important question that has received surprisingly little systematic attention.

Their central reassurance: even the most extreme longtermist society would invest substantial resources in current people's welfare because instrumental reasons mandate it. People whose basic needs are unmet cannot do complex work. People who are unhappy are less productive. The authors offer this analogy: just as profit-maximizing companies provide "high-quality office space, food, and entertainment for their staff" to boost shareholder returns, longtermist states would provide for current citizens to maximize future benefit.[2]

Here's the miss: they treat existence like a budget across time—present people existing now, future people existing later, with the question being how to allocate resources between these temporal slices. But **existence is not quantity distributed across time—it is a continuous adaptive process operating in networks.** There is no "short-term" existence separate from "long-term" existence. There is only a loop—sense, learn, decide, act, repair—or there isn't.

The question isn't how to sacrifice present for future, but whether the structures that enable existence to persist—the adaptive loop itself—are being maintained or destroyed. The implementation Cotton-Barratt and Hadshar describe mistakes optimization toward imagined future targets for maintenance of the adaptive process. In doing so, it destroys the very structures that enable anything to persist.

To understand why, we must examine both the incentive gradient in systems with unver-

[1] Cotton-Barratt, O., & Hadshar, R. (2025). "What Would a Longtermist Society Look Like?" In Greaves, H., Barrett, J., & Thorstad, D. (Eds.), *Essays on Longtermism*. Oxford University Press, p. 334.

[2] Cotton-Barratt & Hadshar (2025), p. 339.

ifiable goals, and the nature of existence as continuous network adaptation.

2 The Incentive Gradient

Cotton-Barratt and Hadshar distinguish partial from strict longtermism, acknowledge "imperfections," and note legitimacy constraints. But nuance doesn't stop a structural drift: systems with certain properties tend toward their worst configurations, regardless of initial intentions.

The gradient emerges when a system has:

- Goals that cannot be verified for decades or centuries
- Officials who must use measurable proxies for success
- Power holders claiming to represent voiceless beneficiaries
- No feedback loop from actual beneficiaries

Six mutually-reinforcing dynamics create the drift:

- **Measurability bias**: Optimize what's measurable (productivity, compliance) over what matters (adaptive capacity, error correction)
- **Goodhart drift**: Once proxies become targets, they diverge from goals—optimizing the metric destroys what it measured
- **Selection effects**: The system promotes true believers in current orthodoxy while sidelining skeptics
- **Legitimacy ratchet**: Early restrictions on dissent normalize later, stronger restrictions
- **Exit barriers**: Each brain drain incident justifies more controls
- **Institutional capture**: Power concentrates with those who define "success" in unverifiable domains

This isn't deterministic—call it a **gravity well**. You can resist, but the default trajectory is drift toward configurations where these dynamics reinforce each other. We've seen this

drift: post-crisis surveillance that outlives the crisis; centrally set quotas that beat the numbers but miss reality.

When long-horizon efforts do succeed—the Montreal Protocol, for instance—it's because they generate near-term verifiable feedback, reversible steps, and co-benefits. These are precisely the features missing in strict, state-imposed longtermism.

A strict longtermist state faces this gradient by design: goals unverifiable for generations, proxies required, beneficiaries voiceless, power centralized claiming authority to represent trillions who cannot check those claims.

But the gradient is not the deepest problem. The deepest problem is the conceptual error underlying the entire framework: mistaking existence for a quantity distributed across time, when existence is actually the adaptive process itself.

3 Existence as Loop Maintenance

We don't keep the future by thinking about it harder. We keep it the way we keep a city running through the night.

Think of what has to hold, quietly, for an ordinary Tuesday to work: the lights stay on; clean water comes out of a tap; your phone finds a signal; the clinic has antibiotics; the train arrives roughly when it should. None of these is produced by a single heroic actor. They're held up by relationships—power stations balancing loads, engineers reading dials, rules that keep us from cutting corners, people repairing what breaks before anyone notices.

That is the shape of our survival. It isn't a jackpot we win later; it's a rhythm we keep now—sense, learn, decide, act, repair—over and over, across many hands. When that rhythm falters, there is no "long term" waiting patiently for us to catch up.

Only then is the right word clear: this is a **network**. Not a social-media buzzword, but the plain fact that existence is a shared maintenance loop among many different parts coordinating under rules, information flows, and repair—households, labs, farms, courts, grids, schools—each doing its own work, each coupled to the rest. You could call this **autonomous interdependence**: we self-regulate, and we stay reliably connected by reciprocal constraints—transparency, coordination, repair.

The loop has five components:

- **Learn**: Gather evidence, update models, explore scenarios
- **Decide**: Choose based on evidence, commit resources, explain choices
- **Act**: Deliver on commitments
- **Coordinate**: Share information, align diverse units, enforce mutual constraints
- **Repair**: Fix breakage—technical and relational

The loop requires network structures:

- **Diversity**: Multiple approaches maintained simultaneously (enables exploration)
- **Autonomy**: Units capable of self-regulation (enables genuine diversity, not cosmetic variation)
- **Interdependence**: Reliable coupling through coordination (enables scale without fragmentation)
- **Error correction**: Capacity to detect failures and shift toward alternatives (enables adaptation)

Note: **Interdependence has capacity** (bandwidth to coordinate) **and quality** (contestability, transparency, and repair). Coercion can raise capacity while destroying quality.

Remove any element and the loop degrades: no diversity means single failure mode; no autonomy means apparent diversity collapses to centralized control; no interdependence means fragmentation; no error correction means lock-in to wrong models.

The profound insight: There is no "longtermism" separate from this loop. You don't "sacrifice present for future"—you either maintain the loop that enables continued existence, or you destroy it. The capacity to continue adapting IS what persists. The search for how to keep the Tuesday-morning services running IS the keeping of them.

Cotton-Barratt and Hadshar's framework makes a category error: treating "serving the far future" as optimization toward distant targets (reduce AI risk, prevent pandemics, produce longtermist goods) rather than maintenance of the adaptive loop. This error becomes self-defeating when combined with the incentive gradient.

4 How the Gradient Destroys the Loop

Cotton-Barratt and Hadshar are explicit: current welfare matters "because unhappy people are less productive than happy people."[3] When current needs don't enhance productivity toward longtermist goals, the state might choose not to fund them—they note it's "conceivable that a longtermist state might choose not to fund palliative care, or choose not to support lengthy retirement."[4]

In optimizing productivity toward unverifiable far-future targets, the gradient systematically destroys the adaptive loop. Consider what happened in Soviet central planning—a system that began with genuine concern for future generations but faced the same structural dynamics.

Soviet planners couldn't verify their goals for decades (industrialization targets, collectivization outcomes). They used measurable proxies: tons of steel, hectares planted, quotas met. The gradient operated:

Measurability bias kicked in: Factories optimized for tonnage metrics. Heavy, useless products counted the same as valuable ones. Farms reported planted hectares regardless of soil quality.

Goodhart drift followed: Production metrics diverged completely from genuine productivity. The economy appeared to grow while actual welfare stagnated.

Selection effects compounded: Officials who questioned the metrics didn't advance. Those who met quotas through creative accounting did.

The legitimacy ratchet tightened: Early questioning of targets was dismissed as counter-revolutionary. Dissent became risky, then dangerous, then impossible.

Exit barriers rose: Emigration became treason. Internal movement was restricted. Alternative economic organization was criminalized.

Institutional capture completed: Those who controlled the definition of "progress toward communist future" had no accountability to actual beneficiaries.

The loop degraded: Engineers stopped reporting true problems (learning failed). Farms stopped adapting to local conditions (adaptation failed). Alternative models were suppressed (diversity failed). Coordination became coercion (interdependence quality degraded). Error correction became impossible (dissent was suppressed).

[3] Cotton-Barratt & Hadshar (2025), p. 339.
[4] Cotton-Barratt & Hadshar (2025), p. 341.

Proxies kept rising—right up to the shock that showed learning, dissent, and reversibility were unfunded. By the time feedback arrived—empty store shelves, technological stagnation—the capacity to sense problems, learn from evidence, adapt approaches, and coordinate solutions was gone.

This is precisely what the longtermist state faces. It optimizes for productivity toward targets verifiable only centuries later (AI risk reduction, longtermist goods production). The gradient operates through the same mechanisms. Learning becomes confirmation of current priorities. Adaptation becomes execution. Simulation collapses to projection of current assumptions. Evidence-based decisions become authority-based (what the state defines as "longtermist goods"). Coordination transforms from chosen interdependence to coercion. Error correction fails as alternatives are suppressed.

The authors acknowledge the risk: "A state that dogmatically adopted a longtermist perspective might indoctrinate its citizenry, in a way that prevented them from adopting better perspectives."[5] They frame this as an "imperfection." But the gradient reveals it as structural necessity. Officials need legitimacy through citizen buy-in around current priorities. The system selects for convergence. When unforeseen challenges arrive, there's no capacity to adapt—the loop has been optimized away.

The conceptual error: You cannot serve "the far future" by destroying present adaptive capacity, because the far future is not separate people later; it is the next iteration of the same adaptive loop. Destroy the Tuesday-morning maintenance network now—the sensing, learning, adapting, coordinating, repairing—and there is no network later to maintain anything. The gradient optimizes for distant targets while destroying loop maintenance. Without the loop, nothing exists.

5 What Loop Maintenance Requires

If genuine concern for persistence means maintaining the adaptive loop rather than optimizing toward distant targets, what does that require?

Loop-maintaining projects share patterns:

- Generate feedback within learning cycles (climate action prevents disasters now, provides immediate evidence)
- Explore multiple approaches (medical research tries many paths, maintains genuine alternatives)

[5] Cotton-Barratt & Hadshar (2025), p. 341.

- Build autonomous capacity (education enables independent judgment, not just compliance)
- Strengthen coordination quality (improve cooperation through reciprocity, not coercion)

Loop-destroying projects share patterns:

- No present feedback (optimize for targets verifiable only centuries later)
- Suppress alternatives (channel all resources to current priorities)
- Require compliance (instrumentalize autonomy for productivity)
- Degrade coordination quality (exit barriers rise, surveillance replaces trust)

The test isn't present benefit versus future benefit—it's whether the project maintains or degrades the adaptive loop. When you repair infrastructure, you generate feedback (did it work?), maintain expertise (learning continues), explore different approaches (diversity preserved), and keep the Tuesday-morning loop running. When you commit irreversibly to projects optimized for one far-future scenario, you sacrifice feedback, suppress alternatives, and bet everything on current models being correct.

5.1 Loop-maintenance principles (illustrative)

Institutions that resist the gradient's erosion of the loop might include approaches such as:

Feedback within years: Projects generate evidence within timeframes that enable learning—years to decades, not centuries. Verification happens through current results, not distant promises.

Maintained alternatives: Resources flow to multiple approaches, including those that challenge current orthodoxy. Genuine diversity requires autonomous units capable of self-regulation according to different assessments.

Reversibility by default: No irreversible commitments that prevent future iterations from choosing differently when evidence accumulates. Preserve adaptive capacity over optimization for current best guesses.

Resourced dissent (structurally independent): Critics receive funding from the budgets they criticize. Error correction requires institutional support, not just tolerance.

Exit viability (low switching costs): Switching costs remain low. Interdependence must be chosen coordination among autonomous units, not forced dependency.

Periodic re-evaluation: Automatic review at set intervals. No indefinite lock-in to approaches that may have become obsolete.

Polycentric approval: Major commitments require sign-off from multiple bodies with conflicting incentives, preventing institutional capture.

If a proposal raises output on a chosen target while failing two or more of these, it's not stewardship; it's loop erosion.

None of this indicts partial longtermism embedded within loop-maintaining institutions; it warns against strict, state-imposed longtermism that optimizes proxies while eroding the loop.

These aren't optimal policies but recognition that loop maintenance requires active resistance to the gradient. The default trajectory is drift toward target optimization that destroys adaptive capacity. Preserving the loop—the Tuesday-morning maintenance across power grids, water systems, knowledge networks, coordination mechanisms—requires structures that actively prevent this drift, continuously, as part of the iteration itself.

6 Conclusion

Cotton-Barratt and Hadshar ask what longtermist societies would look like. Their answer reveals a fundamental misunderstanding of what enables existence to persist.

They treat existence as quantity distributed across time—present people now, future people later—and frame the question as resource allocation between temporal slices. But **existence is process**: the continuous adaptive loop operating in networks. The Tuesday-morning rhythm of sensing, learning, adapting, coordinating, repairing. This loop either maintains itself or collapses. There is no "present existence" separate from "future existence"—only ongoing maintenance of adaptive capacity.

Their framework optimizes for productivity toward far-future targets. But this mistakes target optimization for loop maintenance. The incentive gradient then destroys each component: learning degrades to confirmation, adaptation becomes execution, simulation collapses to projection, evidence-based decisions become authority-based, coordination transforms to coercion, error correction fails, iteration stops. We saw this pattern in Soviet planning. The gradient operates through the same mechanisms in any system

with unverifiable goals, necessary proxies, voiceless beneficiaries, and power asymmetries.

Proxies keep rising—high productivity metrics, measurable progress toward longtermist goals—right up until unforeseen challenges expose that the adaptive loop is gone. The capacity to sense honestly, learn from evidence, adapt approaches, and coordinate solutions has been systematically removed. Not as an "imperfection" but as the necessary consequence of optimizing for unverifiable distant targets while the gradient operates.

The authors' careful distinctions between partial and strict longtermism prove unstable. Without active institutional resistance, the default is drift toward configurations where loop components degrade while productivity toward current targets increases. The nuance fails because it doesn't address the conceptual error: treating existence as quantity-in-time rather than recognizing it as the adaptive loop itself—the ongoing maintenance that keeps the lights on, the water flowing, the coordination functioning.

But this clarifies what persistence requires. There is no "longtermism" separate from maintaining the structures that enable the adaptive loop to continue. The future we claim to serve is not separate people later—it is the next iteration of the same maintenance network. You cannot serve it by destroying present adaptive capacity. You serve it by maintaining the loop: preserving diversity, autonomy, interdependence quality, and error correction as functional requirements for the process that constitutes existence.

The test is simple: Does an approach maintain or destroy the adaptive loop? Cotton-Barratt and Hadshar's longtermist societies destroy the loop while claiming to serve the far future. They optimize for productivity toward unverifiable targets while the gradient eliminates the capacity to sense, learn, adapt, correct errors, coordinate through reciprocity, and iterate. They sacrifice the Tuesday-morning maintenance network—the quiet, continuous work that keeps existence running—for optimization toward distant destinations. They build brittleness while pursuing security.

Those who care about enabling existence to continue should recognize this as a category error with catastrophic consequences. Existence is not a destination to optimize toward but a process to maintain. The far future is not separate from the present loop—it is the continued operation of the same adaptive maintenance. Anything that destroys present adaptive capacity destroys the capacity for anything to exist later.

True concern for persistence means maintaining the prerequisites for the adaptive loop: autonomous interdependence, diversity, error correction, continuous iteration.

The search is the finding.

The loop is the persistence.

The maintenance is the goal.