Populectomy

How AI unlocks hidden rational incentives for lethal mass population reduction

Abstract

Mass civilization, a state of cooperation between extremely large numbers of people, is too readily treated as intrinsically and universally desired, despite many thousands of years, culminating only recently, during which peaceful human coexistence was limited to small groups. This essay argues that a relatively peaceful and growing human population of billions is not the inevitable result of "progress." Rather, it is an accidental consequence of current human technical capabilities, which new technologies, particularly AI, threaten to disrupt. As automation renders most humans militarily obsolete, we should not expect any benefits to be broadly distributed. Instead we should expect the social contract to unravel, transforming mass elimination of surplus humans from an unthinkable atrocity into a calculated strategy – a populectomy. Furthermore, this essay argues that the technical threshold for enabling violent population reduction is relatively modest, not requiring superintelligence but merely task-specific systems capable of overwhelming conventional human resistance. The authors conclude that the peaceful path forward lies in rejecting AI and embracing managed demographic contraction with all humans given equal part in determining humanity's future.

Introduction

Prologue

In 2018, an American missionary and adventurer named John Allen Chau set out to convert the inhabitants of North Sentinel Island to Christianity.[1] The

[1]Details about the Sentinelese, including the incident with John Allen Chau, come from "Sentinelese" on Wikipedia, February 18, 2025. https://en.wikipedia.org/w/index.php?title=Sentinelese&oldid=1276410385. Some relevant primary sources, referenced by the Wikipedia article, include Goodheart, Adam. "The Last Island of the Savages." The American Scholar 69, no. 4 (2000): 13–44; Pandit, T. N. The Sentinelese. The ASI Andaman and Nicobar Island Tribe Series. Calcutta: Seagull Books on behalf of the Anthropological Survey of India, 1990; and Gettleman, Jeffrey, Hari Kumar, and Kai Schultz. "A Man's Last Letter Before Being Killed on a Forbidden Island." The New York Times, November 23, 2018, sec. World. https://www.nytimes.com/2018/11/23/world/asia/andaman-missionary-john-chau.html.

Sentinelese, as they are known to the outside world, are among the world's most isolated peoples, having lived in voluntary seclusion for thousands of years on their small island in the Bay of Bengal. What the islanders call themselves remains a mystery, as they are so reclusive and hostile toward outsiders that no meaningful communication has ever been established with them.

When Chau approached their shores, the islanders warned him away with bow-fired arrows. Undeterred by this explicit rejection, he persisted in landing on the island. The Sentinelese killed him and buried his body on the beach.

Though the local Indian authorities nominally registered murder charges against "unknown persons," no serious pursuit of modern procedural justice followed. From one perspective, a kind of natural justice was already served, the same kind of justice served in one of those perennial stories about visitors to zoos who climb into wild animal enclosures and are subsequently mauled. The visitor is the one who should know better, in contrast to the animal that is acting only by instinct.

We can also appreciate the coincidence that the lethal Sentinelese response protected them from a serious existential threat. Though Chau, the zealous Westerner, was just a solitary individual, his body could have introduced pathogens to which the had no immunity, and him establishing sustained contact could have opened the door to further contact and eventually the likely erasure of their way of life. The Sentinelese behaved *as if* they had all of the facts and were protecting their rational self-interest. But, their knowledge of the history colonization and germ theory being unlikely, we would still ascribe their behavior to an idiosyncratically hostile culture, not subject to rational analysis.

But perhaps there was something at work here other than an unfortunate but avoidable collision between cultures that appear alien to one another? Perhaps the Sentinelese behavior even gives us a glimpse into the behavior of technologically sophisticated humans in the near future, not just the pre-historic past?

The New Earth thought experiment

Let's consider a thought experiment. Imagine you are in control of a one-way colonization mission to an uninhabited planet that is a near-replica of Earth. You are responsible only for the well-being of your crew, and have no external constraints on crew size nor on the equipment you may bring with you, but are leaving Earth permanently behind. How many people would you take?

Even with a vast, hospitable New Earth to colonize and no crew size restrictions, it seems nearly certain that even with only cursory consideration you would not choose to bring billions, or even millions. You are more likely to take a small number – perhaps only a few thousand (if genetic diversity is accounted for in the planning).

Consider how your small crew would thrive, even with minimal supplies from the home world. If you dispersed yourselves along coastal areas into a federation of

small bands, for example, you wouldn't have to trouble with agriculture, subsisting on nothing but the natural bounty of the land and the sea. A society in the thousands would permit effective governance through direct participation rather than distant representation. Environmental impact would remain manageable without complex regulation. Social cohesion would be sustained by the absence of scarcity and each member's role being visible and valuable to the community. You could still benefit from modern scientific and technical know-how, but leave behind a vast range of modern technology that in your new context would seem silly, wasteful, or harmful.

The preceding thought experiment gives you full control over whom to bring along. But how would you ever attain such god-like authority? In a less contrived New Earth scenario the many people who would want an opportunity for a fresh start wouldn't just placidly accept being excluded. While, as we established, taking on additional colonists quickly defeats the less-is-more appeal of the whole mission, you and your select crew, being few, wouldn't be in the position to say no. Even if you could keep unwanted passengers off *your* transport, how would you stop others from coming on their own? So the colonization of New Earth would unfold in a very different way than the disciplined mission we originally outlined. It would be a mad race for the best possible position on the new planet. While those with the least to lose by leaving Earth would gain a chance to come out on top, New Earth would eventually end up as crowded and despoiled as the home world.

You would want the mission to be a *secret* mission, then; you would want to keep the existence of New Earth secret and slip away unnoticed. But once there, you would keep a suspicious eye to the heavens, hoping against the appearance of unwanted Earthlings. Should any one day appear, you would be wise to repel them, no matter how friendly-seeming their overtures, or else more would surely follow.

Let's return now to the Sentinelese. The essential Sentinelese qualities that emerges from descriptions of them, besides their hostility, are good physical health and discipline. They are muscular, appear well-fed, and their individual members cannot easily be tempted into alliances with outsiders bearing gifts. Both suggest that their group has plenty to go around and that it *does* go around. Their physical appearance is one obvious sign, but their unwillingness to be bought indicates they aren't motivated by in-group power struggles to leverage outside resources for advantage over one another.

A reason for their hostility now comes into focus. Perhaps, as the Sentinelese have everything they need, individually and collectively, the outside world has nothing to offer them but the risk of ruin? Their most telling response to would-be visitors is their habit of turning their backs as a group, pantomiming defecation, and retreating to fire arrows. This ritual conveys their assessment of outsiders as nothing but waste.

The Numbers Advantage

Life on New Earth – and, several signs suggest, North Sentinel Island – could be better than ours in every way. Every way but one, that is: having the force of numbers to repel or execute incursions. But this advantage tends to trump all others.

The New Earth mission would die in its crib if it had to build and arm a military big enough to successfully hold off an Earthling invasion. It has to rely on stealth and secrecy, a highly uncertain proposition. Should the cover of secrecy fail and determined Earthlings arrive en masse, the mission would surely be lost; the best the colonists might hope for would be to make a heroic but futile stand.

If the threat to an imaginary and exclusive utopia doesn't seem compelling, the plight of the flesh-and-blood Sentinelese, whose position is just as precarious, might. Their hostility protects them from the trespasses of tourists, poachers, and missionaries, but from the big, avaricious world they are protected only by the remoteness of their island and external preservation efforts. One can hope they are blissfully unaware of their insecurity. Perhaps they do know and, having everything to lose from capitulation but their lives, stand their ground anyway.

But what if the advantage of numbers could be nullified?

Let's take the New Earth thought experiment one step further. Imagine that, at the same moment you discovered the habitable new planet with your proprietary telescope, you also identified a planet-destroying asteroid on a collision-course with this one. You need only sneak off to your new planet, as you intended to do in the previous iteration of the thought experiment. But, as they say, in for a penny, in for a pound. What if the heaven-sent asteroid is not *quite* on a collision course, but you had the means to nudge it onto one?[2] Finally, why bother with asteroids and interplanetary voyages at all? Your problem with Earth is not with the other Earthlings' character, just their numbers. If you had a single-shot tool that could wink the bulk of humanity out of existence, you could have your New Earth right here.

We come now to the point of these thought experiments. New technologies are now emerging that could nullify the offensive advantage of numbers – and in particular, AI - surfacing hidden human incentives to dramatically and violently shrink the human population. There has been extensive coverage of the risk of AI-driven human elimination being triggered accidentally, by the emergence of a misaligned superintelligence. But insufficient attention has been paid to the shift in human incentives, or more accurately, the removal of unnoticed constraints, that would make intentional mass human elimination imaginable and likely. This investigation aims to address that critical blind spot in existing discourse on AI risk. For the AI researcher focused on the alignment problem this area may be

[2]The necessity of pre-emptive world-killing will be familar to readers of Cixin Liu's *The Dark Forest*. Liu, Cixin. *The Dark Forest*. Translated by Joel Martinsen. First edition. A Tom Doherty Associates Book. New York: Tor, A Tom Doherty Associates Book, 2015.

"out of scope" or a "problem for later," but not for the average person, for whom the accumulating risks from both misaligned *or* aligned AI make the distinction merely academic: what difference does it make whether AI eventually destroys everyone, if the last to fall have already used the technology to eliminate the vast majority of humanity?

Mass civilization is not human destiny

The initial New Earth thought experiment presents a blank slate on which to design an optimal society for human well-being, with no other considerations, in order to demonstrate the intuitive preferability of small and sustainable over competitive and expansionary. Nevertheless, our modern civilization's relentless growth is often presented as a grand shared project of self-improvement driving toward our destiny or telos, perhaps to expand to other planets and beyond. We speak of "progress" as if it were a irrefutable moral imperative that no rational, technologically developed people would ever willingly jeopardize. Even the assumption of existential risks, such as the pursuit of AI, are sometimes justified as a bold gamble necessary to fulfill our destiny.

But a teleological framing distorts our understanding of human history and constrains our ability to imagine alternatives. Humans never made a collective choice to transition from hunting and gathering to building a mass civilization as a way of improving ourselves. Mass civilization beat non-agrarian ways of life through force of numbers, not an aggregate individual preference for one over the other. Whatever the other advantages of civilization may be, they are only by-products of this dynamic. So if other options become, for the first time since the Neolithic Revolution, actually a matter of choice, even holding all else equal (which, as we will see later, we can_not_ do), allegiances to mass civilization would be sorely tested.

Even if we in insist on treating progress is our destiny, we should not assume that its benefits would be shared. Instead, we should consider mass civilization itself to be a kind of technology that is subject to "creative destruction." Past disruptions have served mostly as "job killers," not "human killers," but only because they did not succeed in moving beyond the need for large populations. What happens when this last constraint is eliminated?

Populectomy

Since the Neolithic Revolution, technological advancements have generally supported increasing cooperation within larger populations, for both production and military strength. The industrial revolution exploited concentrated workforces, complex supply chains, mass markets, and large armies. Even the emergence of awesomely destructive nuclear weapons didn't practically facilitate the reduction of population. On the contrary, their value being in generating the mutual

desire to avoid their use, they have increased cooperation and interdependence across the global population. AI and robotics fundamentally break mass civilization's cooperative patterns, representing a qualitatively different technological threshold with unprecedented implications.

We should be extremely concerned about *any* weaponry that doesn't require significant manpower to deploy and is highly effective at killing as many people as its operators choose, without spoiling the environments in which it is deployed. For example, we can imagine a bioweapon that is guaranteed to quickly kill 99.99% of people while sparing other species, the few humans that have been clandestinely innoculated, and some lucky survivors. All would effectively be delivered to a New Earth.

Because the capabilities we are describing are unprecedented, changing the status quo which has governed social behavior for thousands of years, one struggles to find a good word for the action being described here. The author proposes "populectomy" to capture the idea of intentionally excising most people to achieve a much smaller human population.

AI tasked with his purpose is, in a sense, just one of several candidates. What sets AI apart from bioweapons, however, is that it competes directly with humans in their performance of useful functions. Unlike previous technological shifts that supported increasing human participation, in war and in peace, AI creates a technological path that systematically excludes most humans from these necessities. Social arrangements will eventually "price in" the diminished utility of strangers to one another, removing normative barriers to the elimination of zero-sum competitors. AI doesn't just provide the means and opportunity, it also reinforces the motive.

Dissolution of the social contract

Cooperative social arrangements aren't maintained through altruism or self-enforcing laws alone. The social contract is secured by two powerful forces: mutual dependence for gain and the implicit threat of reprisal. Civilian workers can stop production, the able-bodied can resist violently, and disillusioned police or soldiers can switch sides. When workers can withdraw their labor or citizens can organize resistance, the powerful must negotiate rather than dictate terms. This balance of power, however imperfect, has maintained a degree of stability and reciprocity in modern mass societies.

But what happens when this foundation begins to crumble? As automation eliminates the need for human labor, the fundamental equation of our social contract changes dramatically. Artificial intelligence and robotics promise to create a world where human participation in production becomes increasingly optional. The mutual dependence that has characterized industrial society begins to dissolve, and with it, the ability to maintain trust and security.

If defense, policing, and production no longer requires human participation, the bonds that have maintained cooperation will unravel. Those formally controlling advanced AI systems and automated production will begin to view the majority of humanity not as necessary partners, and not even as harmlessly superfluous, but as threatening competitors for resources and control. Those not holding formal control will, sooner or later, recognize the tightening noose of disempowerment around their throats and resist, further justifying their elimination. As mass civlization backslides into dark zero-sum insecurity, violent population reduction could become an increasingly attractive release.

The risk is most obvious in regimes that are quickly captured by the narrow interests of the owners of the robotic means of production. Imagine, as in Marshall Brain's dystopian vision in *Manna: Two Views of Humanity's Future*[3], that economic obsolesence sends millions of unemployed to robot-produced and robot-managed cell blocks. Because of the pervasiveness of surveillance and robotic security forces, roiting or civil disobedience ensure any protestor ends up in jail. But so, too, would compliance. It would becomes unclear to anyone still having a job why they should continue working at all, when unemployment and imprisonment is only a matter of time. On the other side of the equation, at some point there would be no incentive for the owners of the robotic means of production to house, clothe and feed the swelling ranks of the permanently unemployed. When the endgame of euthenasia becomes apparent, the entire social contract would collapse into a fight for survival, in which having control of the machines is the only way to win, and the only way to retain control is to eliminate the competition.

At first glance it may seem that such collapse could be avoided by an early renegotiation of the social arrangements, in favor of shared ownership of AI's output, such as through guaranteed income. But UBI eventually fails in the same way as the work incentive in the dystopian view: as interdependence erodes, so does trust that *any* social arrangements will be honored, authoritarian or egalitarian. It's worth pointing out that the common criticism of UBI, that it fails to provide a sense of psychological security equivalent to work, hints at the real problem but comes at it from the marginal case of an AI world with only modest job insecurity. When workers lose all but their formal claim to a share of productive output, the psychological feeling of insecurity corresponds with existential insecurity.

Under both initial conditions, the welfare and the carceral state, the crisis of mistrust created by the loss of leverage only fades when population is so low that natural abundance is guaranteed and there is really nothing left to be mistrustful about.

[3]Brain, Marshall. "Manna – Two Views of Humanity's Future." 2003. Accessed March 18, 2025. https://marshallbrain.com/manna1.

Culling calculus

A chilling rationality emerges – one that transforms population reduction from unthinkable atrocity to calculated strategy. Understanding this calculus doesn't require embracing it; rather, it demands clear-eyed recognition of incentives that will exist when the technological threshold is crossed. The rational calculation becomes startlingly simple: why tolerate billions of competitors when a dramatically smaller population would maximize resource availability and minimize internal and external existential threats?

More troubling still is the self-fulfilling dynamic this creates. Once the technology exists, multiple groups will inevitably pursue it defensively. Consider the cold logic: if one community might obtain means to eliminate competitors, any group wishing to survive must launch a preemptive strike or ally itself with it. Counterintuitively, this dynamic fosters cooperation rather than competition between rival groups with similar technological capabilities, but against everyone else. As elimination becomes feasible, the wisest strategy becomes combining "no-kill lists" – exclusive, protected populations - rather than risking mutual annihilation through competition. The net result of the mere possibility of selective mass killing guarantees multiple motivated actors pursuing it simultaneously. This mirrors game theory's darkest predictions, but unlike nuclear deterrence which maintained stalemate through universal undesirability, groups genuinely benefit from population reduction, making restraint far less likely.

The moral restraints normally preventing such calculated elimination depend heavily on mutual dependence. When that dependence dissolves through technological uncoupling, ethical considerations dissolve with it. As with the Sentinelese hostility to outsiders, behavior that is deviant and self-defeating in a cooperative world becomes instrumental to survival.

A relatively modest technical threshold

When we consider technologies like autonomous drones, facial recognition, predictive analytics, and industrial and military robots, we see capabilities that don't require sentience or superintelligence to be devastatingly effective at population reduction. The threshold of AI capability required for effective population reduction may be significantly lower than commonly assumed for misalignment existential risk scenarios.

Task-specific systems designed for targeted functions can be more immediately dangerous precisely because they are more limited in scope, making them technically simpler to develop while still posing an existential threat to most humans. And the automation that assists a small group of humans in defeating others' defenses is technically easier to achieve than automation that has to defeat all humans, against its developers' own intentions. The systems need only be good enough to overwhelm conventional human resistance when guided by human operators with strategic intent.

Warning signs

Even now, AI and robotics may be ushering in a new world that will render the current world's allegiances meaningless. As the new order approaches, actors would have to exploit the outgoing order to carve for themselves a place in the next. Within the largest nation-states, for example, citizenship ultimately confers no long-run protection. Only gaining and retaining exclusive control over certain critical systems, especially military ones, does. Still, democratic and free-market mechanisms are useful instruments for achieving such control for as long as credulous populations still cling to them.

At the same time, to forge successful alliances rooted in the new order, the same actors would need to craft narratives justifying defection from the old one. When putting these two contradictory concerns together, we should expect a confusing remixing of familiar narratives. Declaring political opponents as "enemies of democracy" or protests against genocide as "domestic terrorism," for example, perform this balancing act of recruiting allies and other temporarily useful collaborators while also squeezing as much complacency as possible from the general population.

In a way, the end-state equilibrium is reminiscent of the pre-historic human social order, before human labor became exploitable. So it may be helpful to conceptualize the path to it as a form of backsliding, passing through intermediate states resembling other pre-modern social forms constrained by zero-sum competition, such as feudalism. When we observe the reemergence of feudal norms of the powerful promising protection in exchange for loyalty, along with rapid advances in autonomous weapons systems and the rhetorical and the actual dismantling of the nation-state, we should recognize this convergence for what it potentially represents: preparation for a world where most humans are no longer necessary.

Paths forward

The gravity of the scenario we've outlined demands urgent consideration of potential paths forward. If we accept that advanced AI and automation technologies create rational incentives for population reduction, what options remain available to prevent such a catastrophic outcome?

The authors have little faith in attempts to preserve a high-population world indefinitely. We have presented mass civilization as a contingent state whose benefits could be passed on to a much smaller human population without its deficits. The goal should be to divert the path to that low-population world peacefully, through demographic contraction, and cooperatively, with all currently living humans given equal part in the project.

Returning to the New Earth thought experiment: what if the facts of the planet-destroying asteroid and the possibility of escape to an alternative world were well-known? How might we organize ourselves to ensure the species' survival

and to avoid a desperate free-for-all? Building an evacuation fleet now while planning to figure out how to allocate seats later would be a dangerous way to go about the problem, but is analagous to how AI is currently being developed.

While the purpose of this essay has been to sound the alarm, not attempt to find solutions, we have some suggestions for paths forward.

Even if backsliding away from an interdependent and cooperative world order has begun, the few, for the time being, still depend on the cooperation of the many. Resistance is still possible, and can hopefully be aided by a clear understanding of the stakes.

Regrettably, the public response so far to the possibility of misaligned AI driving humans to extinction has been something to the tune of "better dead than red!" The competition between the world's democracies and authoritarian rivals has perpetuated the AI arms race. Fear of having to live under a despotic regime (or, from the other side, the fear of being forever dominated by the West) has overpowered any fear of the small likelihood that, despite the best efforts of the many great minds working on AI, the technology will fail catastrophically and kill everyone. Recognizing the lethality of aligned AI may change that. In democracies, even the more optimistic projection, that AI could push every nation-state into despotism or feudalism, ought to nullify the mandate for the continued development of AI.

But acknowledging the growing irrelevance of workers and voters to elites, who are focused on their private contest for relevance in the AI world order, should temper expectations around regulation and the distribution of its benefits. Universal Basic Income will not serve as a long-term solution, and as a short-term solution breeds a constituency that is dependent on the AI that will eventually betray them.

The most powerful defense that the masses have in their arsenal is the ability to establish a normative, rather than merely legal, culture of resistance to AI development. The social norms that would develop were it to be recognized for the weapon that it is, one that will be turned against our enemies, friends, and families indiscriminantly, are straightforward: anyone who participates is collaborating in genocide of unprecendented scale. While this may only drive development underground, that may buy some time.

Finally, there may be technical, zero-trust solutions that deter the use of AI, at least temporarily, to cause mass human casualties.

With his famous "Laws of Robotics" Isaac Asimov, the godfather of AI science fiction, baked the proscription of causing harm to humans into the inner-most workings of the "positronic brain." This was a hack, but a necessary one to make recognizable AI-enriched human worlds possible. We cannot depend on such an unlikely coincidence as intelligent robots that are *inherently* incapable of causing intentional harm. Any security measures that can be built into real-world AI are likely to be superficial enough to be susceptible to being defeated at massively

lethal scale.

It may be possible, though, while human majorities still have some say in the matter, to create "dead hand" mutual assured destruction systems that artificially prop up the value of human lives. Such terrifying and risky solutions will be necessary in the dark, low-trust transition from the high-population world to a brighter, more secure low-population one.

Conclusion

We stand at the precipice of an unprecedented transformation in human capability. Advanced artificial intelligence and robotics promise to enable small groups to achieve levels of offensive capability previously impossible without mass cooperation. As these technologies advance, more groups may find themselves in the position of the Sentinelese—viewing others primarily as threats rather than potential collaborators. Unlike the Sentinelese, however, these autonomous groups would not be limited to firing arrows from their shores. They would possess technologies capable of enforcing their need for isolation on a global scale.

The core danger lies not in evil intentions or malicious AI, but in the shifting incentive structures that AI creates. As we approach a technological threshold where small groups can defeat large ones in the application of lethal force, the rational calculus regarding population size fundamentally changes. This shift occurs whether or not AI systems are perfectly aligned with their creators' "values"– human values are where the danger lies.

We must reject the comforting but unfounded assumption that technological progress naturally produces a reduction in insecurity and violence. History demonstrates repeatedly that moral commitments usually follow material conditions rather than transcending them. When circumstances change dramatically, values that seemed inviolable can quickly erode. The technological capabilities we're developing could create the most profound change in human circumstances since agriculature–a shift that may render our current moral frameworks obsolete.

The time to address these incentive structures is now, before they become irresistable. The goal must be technology and social structures that enable a future where humanity, while in all likelihood a much smaller one, flourishes without the violent sacrifice of the many by the few.