Expert Insights
PERSPECTIVE ON A TIMELY POLICY ISSUE

BEBA CIBRALIC

The Strategist's Dilemma

Agent-Neutral and Agent-Relative Reasons in U.S. AI Strategy

July 2026

For more information on this publication, visit **www.rand.org/t/PEA4931-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our research integrity is grounded in RAND's core values of quality and objectivity. Rigorous quality assurance procedures, conflict of interest screening, and transparency in funding ensure that every study is objective and nonpartisan. Learn more at www.rand.org/integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2026 RAND Corporation
RAND® is a registered trademark.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, visit www.rand.org/about/publishing/permissions.

PE-A4931-1

About This Paper

In this paper, I apply philosopher Thomas Nagel's distinction between agent-relative reasons and agent-neutral reasons to the challenges associated with a U.S. artificial intelligence (AI) strategy. In Nagel's formulation, a reason is agent-neutral if it can be stated in a general form without any essential reference to the person who has it—the badness of suffering, for instance, gives everyone a reason to reduce suffering. A reason is agent-relative if its general form necessarily includes such a reference, such as when a duty arises from one's particular role, relationship, or commitment.[1] Nagel argues that both kinds of reasons—those arising from an objective, impersonal standpoint and those arising from the standpoint of a particular agent—generate legitimate demands on action; he also argues that practical reasoning cannot eliminate the tension between the two kinds of reasons. I argue that national leaders face this exact tension in formulating a U.S. AI strategy: Leaders bear agent-relative obligations to protect their own citizens, yet transformative AI generates agent-neutral obligations to prevent grave harms to anyone, regardless of nationality. Drawing on Nagel's *The View from Nowhere* (1986), I explain why both sets of reasons should be in play whenever the United States makes strategic decisions about transformative AI. This paper is written for researchers and policymakers concerned with developing a U.S. AI strategy for powerful AI. I assume that the reader has a familiarity with the basic contours of AI safety debates but not with analytic philosophy.

Center for the Geopolitics of Artificial General Intelligence

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Center for the Geopolitics of Artificial General Intelligence (AGI), which is committed to helping decisionmakers understand, anticipate, and prepare to navigate the national security and geopolitical implications of AGI. The center convenes leading technologists, strategists, economists, political scientists, and outside experts to consider the feasibility and effectiveness of prospective AGI-enabled capabilities; the domestic and international implications of their use; and the strategies and policies that governments, businesses, and civil society could adopt to respond to new realities. For more information, visit www.rand.org/geopolitics-of-agi.

Funding

This effort was independently initiated and conducted within the Center for the Geopolitics of Artificial General Intelligence using income from operations and gifts from RAND supporters,

[1] Thomas Nagel, *The View from Nowhere*, Oxford University Press, 1986, p. 152.

including philanthropic gifts made or recommended by DALHAP Investments Ltd., Ergo Impact, Founders Pledge, Charlottes och Fredriks Stiftelse, Good Ventures, Longview, and Coefficient Giving. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

This paper builds on ideas that stemmed from debates that occurred between members of the RAND Center for the Geopolitics of AGI in fall 2025. I am especially grateful to Joel Predd, Ben Boudreaux, Matt Chessen, Gopal Sarma, and Jeff Alstott for their feedback.

Summary

Philosopher Thomas Nagel argues that practical reasoning is shaped by two *kinds* of reasons. *Agent-neutral reasons* are reasons that hold true for anyone regardless of their position; for example, the belief that suffering is bad gives everyone a reason to prevent suffering.[2] *Agent-relative reasons* depend on who a person is and what role they occupy; for example, a mother may have special duties to her child. In this paper, I consider what Nagel's distinction means for the development of a U.S. artificial intelligence (AI) strategy, and I argue that national leaders face what I call a *strategist's dilemma*: Transformative AI generates both agent-neutral and agent-relative reasons, and they can pull in different directions. In what follows, I provide an approach for reasoning through the trade-offs when agent-neutral and agent-relative considerations are in tension.

Key Takeaways

In this paper, I argue the following:

- **The tension between agent-neutral and agent-relative reasons maps onto existing debates between the AI safety and national security communities.** AI safety arguments foreground agent-neutral reasons—obligations that hold true for anyone, regardless of the person's nationality, because catastrophic risk harms everyone.[3] National security arguments foreground agent-relative reasons—obligations indexed to a particular government's role, its citizens, and its commitments. In this paper, I argue that both kinds of reasons are relevant for formulating a strong U.S. AI strategy.
- **Agent-neutral reasons should be explicitly incorporated into a U.S. AI strategy.** When catastrophic risk is framed only as an implication of national interest, it becomes vulnerable to being traded away under competitive pressure. Moreover, agent-neutral reasons give catastrophe avoidance independent standing as a constraint; generate international legitimacy by grounding policy in reasons that any country could recognize; and reframe AI competition as a coordination problem, opening space for mutual restraint that self-interested reasoning forecloses.
- **Agent-relative reasons are not merely parochial; they carry moral weight.** Countries are not abstract welfare maximizers. Rather, they are political agents answerable to specific citizens, institutions, and constitutional commitments. A strategy that treats national

[2] *Agent* here refers to a moral or political actor—a person, government, or institution—not to AI systems, which are *AI agents*. See Thomas Nagel, *The View from Nowhere*, Oxford University Press, 1986.

[3] Note that *safety* is broader than catastrophic risk—it also covers misuse, unreliability, and deployment harms that are localized and track who deploys what and where. When I say that safety arguments foreground agent-neutral reasons, I mean specifically the catastrophic, irreversible subset.

obligations as obstacles to impartial morality misunderstands both the nature of political legitimacy and the motivational resources that make sustained international cooperation possible in the first place.

- **The two kinds of reasons cannot simply be collapsed into one.** Catastrophic and irreversible risk should function as a constraint on what counts as a permissible strategy, not as one consideration to be weighed against national advantage. Only once that constraint is met do trade-offs between agent-neutral and agent-relative considerations properly arise.

Contents

About This Paper iii
Summary v

The Strategist's Dilemma: Agent-Neutral and Agent-Relative Reasons in U.S. AI Strategy 1
Nagel's Dual Standpoints: Agent-Relative Versus Agent-Neutral Reasons 2
Taking *The View from Nowhere* Seriously in Developing a U.S. AI Strategy 4
Preserving Special Obligations to Citizens 5
Handling Hard Trade-Offs Between Global and National Reasons 7
Objections and Counterarguments 9
Conclusion 11

Bibliography 13
About the Author 15

The Strategist's Dilemma: Agent-Neutral and Agent-Relative Reasons in U.S. AI Strategy

The prospect of advanced artificial intelligence (AI) presents a conceptual challenge for a U.S. national strategy: How should the United States design and evaluate an AI strategy when many of the technology's most significant impacts transcend national borders? National leaders typically bear *agent-relative obligations*—special responsibilities to protect the interests and security of their own citizens. Yet AI carries the potential for global risks that generate agent-neutral obligations, such as impartial duties to prevent grave harms to anyone, regardless of nationality. These dual obligations can pull in different directions, creating what I call a *strategist's dilemma* in developing a U.S. AI strategy.

The tension between national and global obligations is not, of course, unique to AI; the tension arises in contexts involving other transnational threads, including climate change, pandemics, and nuclear proliferation, among others. But transformative AI is the subject of active strategy formation now. Governments, including the U.S. government, are in the process of designing strategies, standing up new institutions, and making commitments that will shape the technology's development trajectory. The value of introducing the agent-relative and agent-neutral distinction into these debates is not that AI is categorically unlike other transnational challenges but that the frameworks being built to govern AI are still taking shape, and a structural confusion about the kinds of reasons in play is easier to correct now.

The tension between agent-neutral and agent-relative reasons maps imperfectly onto familiar disagreements in AI debates. Arguments about AI safety often foreground reasons are rooted in shared vulnerability to large-scale risks and harms.[4] Arguments about national security often emphasize special duties to citizens and competitive advantage. There are, of course, nuanced distinctions within each side. National security practitioners disagree about how broadly the national interest should be understood, and safety advocates differ about how much weight should be given to sovereignty, legitimacy, and obligation.[5] U.S. statecraft, for its part, has rarely operated from either

[4] Note that *safety* is broader than catastrophic risk—it also covers misuse, unreliability, and deployment harms that are localized and track who deploys what and where. When I say safety arguments foreground agent-neutral reasons, I mean specifically the catastrophic subset.

[5] For work that foregrounds humanity-wide risk as the central concern—that is, treating AI as a potential source of civilizational or existential harm regardless of which country develops it—see Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014; and Toby Ord, *The Precipice: Existential Risk and the Future of Humanity*, Grand Central Publishing, 2020. For work that approaches AI strategies primarily through the lens of national security, competitive advantage, and state capacity, see Dan Hendrycks, Eric Schmidt, and Alexandr Wang, *Superintelligence Strategy: Expert Version*, arXiv, arXiv:2503.05628v2, April 14, 2025; and Michael J. Mazarr, *A New Age of Nations: Power and Advantage in the AI Era*, RAND Corporation, PE-A3691-14, January 2026. For a critical engagement that illustrates disagreement within national security camps

moral universalism or raison d'état alone. The value of the agent-relative and agent-neutral distinction is not that it sorts actors into neat camps; instead, its value is that it clarifies the different kinds of reasons that can coexist—and conflict—inside a single strategic argument.

In this paper, I draw on the work of philosopher Thomas Nagel and others to explain why both agent-neutral and agent-relative reasons should be in play whenever the United States makes strategic decisions about transformative AI.[6] Other frameworks in moral and political philosophy tend to come down on one side of the strategist's dilemma. Cosmopolitan approaches—such as those of Peter Singer and Thomas Pogge—give systematic priority to impartial obligations, treating partiality toward one's own citizens as a deviation that requires special justification. These approaches make an important point—partiality toward compatriots does call for justification rather than being self-evidently legitimate—but they understate the moral weight that a country's obligations to its own people can carry once that justification is given. Realist theories of international relations—such as those of Kenneth Waltz and John Mearsheimer—move in the opposite direction, grounding strategy in the structure of interstate competition and treating impartial moral demands as largely inoperative in strategic contexts. Realism captures something true about the competitive structure within which states act, but it leaves no principled basis for treating catastrophic global risk as a constraint on strategy rather than as one more variable in the competition. Nagel's account is useful precisely because he argues that the personal and impersonal standpoints each generate obligations and that the tension between the standpoints must be worked through rather than eliminated.[7]

My argument proceeds in five sections. First, I explain Nagel's distinction between agent-neutral and agent-relative reasons in more detail. Second, I apply Nagel's distinction to U.S. AI strategy formulation and explain why such a strategy cannot be defensible without taking humanity-wide risks seriously. Third, I show why special obligations to citizens remain morally and politically relevant. Fourth, I identify the hardest cases, those in which national and global reasons diverge, and I offer a way of reasoning through them. Fifth and last, I address the likely strongest objections to my argument.

Nagel's Dual Standpoints: Agent-Relative Versus Agent-Neutral Reasons

Nagel argues that practical reasoning oscillates between two perspectives: a particular agent's subjective standpoint and the objective standpoint that abstracts away from any particular person's

about how far competitive framing should go, see Iskander Rehman, Karl P. Mueller, and Michael J. Mazarr, "Seeking Stability in the Competition for AI Advantage," commentary, RAND Corporation, March 13, 2025.

[6] Thomas Nagel, *The View from Nowhere*, Oxford University Press, 1986.

[7] See Nagel (1986). Nagel extends the dual-standpoint account to political theory in *Equality and Partiality* (Thomas Nagel, *Equality and Partiality*, Oxford University Press, 1991). The agent-relative and agent-neutral terms were introduced by Derek Parfit in *Reasons and Persons* (Derek Parfit, *Reasons and Persons*, Clarendon Press, 1984); Nagel had drawn the same distinction earlier using the terms *subjective* and *objective* in *The Possibility of Altruism* (Thomas Nagel, *The Possibility of Altruism*, Clarendon Press, 1970). On the cosmopolitan side, see Peter Singer, "Famine, Affluence, and Morality," *Philosophy & Public Affairs*, Vol. 1, No. 3, Spring 1972; and Thomas W. Pogge, *World Poverty and Human Rights: Cosmopolitan Responsibilities and Reforms*, Polity Press, 2002. On the realist side, see Kenneth N. Waltz, *Theory of International Politics*, McGraw-Hill, 1979; and John J. Mearsheimer, *The Tragedy of Great Power Politics*, W.W. Norton, 2001.

position. In *The View from Nowhere*, Nagel characterizes the objective perspective as a view from nowhere, the closest thing to a bird's-eye, impersonal outlook on the world. By contrast, the subjective perspective is personal, looking out from one's own situation.[8]

These perspectives give rise to two kinds of reasons for action. An *agent-relative reason* is a reason that references the particular agent who has the said reason; that is, the reason is tied to the agent's specific interests, relationships, or commitments. For example, loyalty to one's country or caring for one's own family are agent-relative motivations. Those reasons (or motivations) cannot be fully described without mentioning who has the obligation or stake. An *agent-neutral reason*, by contrast, makes no essential reference to a particular agent; it is a reason that anyone in the same situation (a situation that necessitates having a reason to act) would have. The classic example of an agent-neutral value is the reduction of suffering. If suffering is bad, then it is bad regardless of who experiences it. Any rational being, seeing suffering, has some reason to mitigate it.[9] Nagel's formal definition captures this distinction:

> If a reason can be given a general form which does not include an essential reference to the person who has it, it is an agent-neutral reason . . . [if] the general form of a reason does include an essential reference to the person who has it then it is an agent-relative reason.[10]

Consequentialist and utilitarian theories give primacy to agent-neutral reasons (evaluating outcomes impartially, regardless of who is affected), while deontological or partialist theories foreground agent-relative reasons (emphasizing the duties and commitments that arise from a particular agent's role or relationships).[11] Nagel's account does not give absolute priority to either standpoint; instead, it acknowledges a permanent tension between personal and impersonal demands.

---

[8] Nagel offers a general account that spans different contexts. Notably, Nagel's dual standpoints echo an older problem in ethics that Henry Sidgwick called the *dualism of practical reason* (Henry Sidgwick, *The Methods of Ethics*, 7th ed., Macmillan, 1907). In *The Methods of Ethics*, Sidgwick argues that rational reflection seems to endorse two principles that can come apart: *egoistic prudence* (pursuing one's own good) and *impartial benevolence* (promoting the good of all). From what Sidgwick called the *point of view of the universe*, the good of any one person is, from that standpoint, no more important than the good of any other; this idea underwrites a utilitarian, agent-neutral standard of equal concern. Yet Sidgwick also thought it just as rational for an individual to organize their life around their own good, and he concluded that pure reason does not supply a higher-order rule that fully reconciles these standpoints. Sidgwick had already noticed one form of the friction that the agent-relative/agent-neutral distinction later generalized. Nagel's *view from nowhere* can be seen as a later attempt to theorize that same tension, insisting that both perspectives have real authority over us even when they cannot be perfectly harmonized.

[9] The universality of agent-neutral values is contested. Communitarian critics argue that moral obligations are always embedded in particular social and cultural contexts and that purportedly universal values might reflect the perspective of specific communities rather than an impartial standpoint. See, for example, Michael Walzer, *Spheres of Justice: A Defense of Pluralism and Equality*, Basic Books, 1983; and Alasdair MacIntyre, *After Virtue: A Study in Moral Theory*, 2nd ed., University of Notre Dame Press, 1984. I do not attempt to resolve that debate in this paper; I follow Nagel in treating agent-neutral reasons as having genuine—though not necessarily overriding—authority.

[10] Nagel, 1986, pp. 152–153.

[11] The locus classicus for consequentialist moral theory is John Stuart Mill's *Utilitarianism* (originally published as John Stuart Mill, *Utilitarianism*, Parker, Son, and Bourn, 1863); see also Sidgwick (1907), especially his account of the "point of view of the universe" as the standpoint of impartial benevolence. On the deontological side, the foundational text is Immanuel Kant's *Groundwork of the Metaphysics of Morals* (1785; for an English translation, see Immanuel Kant, *Groundwork of the Metaphysics of Morals*, ed. and trans. by Mary Gregor and Jens Timmermann, Cambridge University Press, 2012). For a modern treatment of agent-relative prerogatives that engages directly with the tension Nagel identifies, see Samuel Scheffler, *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*, Clarendon Press, 1982.

Every decision, he suggests, is at once a choice made by a particular agent with specific obligations and a choice among possible states of the world that can be evaluated by the decision's overall consequences.

Taking *The View from Nowhere* Seriously in Developing a U.S. AI Strategy

Extending Nagel's framework to U.S. AI strategy formulation is useful. A national strategy is, by definition, agent-relative—it is indexed to the interests, security, and values of one nation-state.[12] U.S. strategy documents, for instance, prioritize U.S. economic competitiveness, national security, and the well-being of U.S. citizens. This is also justifiable on normative grounds, in part because the U.S. government has special responsibilities to its citizens. By contrast, many of transformative AI's anticipated impacts—especially the worst-case outcomes—are agent-neutral. They concern harm or benefit to anyone and everyone. Likewise, the distribution of AI's benefits (or its misuse) could affect global stability and the future trajectory of humanity. In short, AI introduces what we might call impersonal risks and opportunities; that is, issues that do not respect jurisdictional boundaries and that matter from a universal standpoint.

Although national leaders cannot abandon their agent-relative duties, there are compelling reasons to incorporate Nagel's view from nowhere—the agent-neutral standpoint—explicitly into the formation of a U.S. AI strategy.[13] Adopting an impartial perspective, even as a thought experiment, can clarify what a strong strategy should look like in light of the global stakes associated with AI.

The first argument for adopting the agent-neutral standpoint is that it helps policymakers calibrate risk across borders. An agent-neutral, bird's-eye perspective helps decisionmakers prioritize outcomes by their global significance. In practical terms, this means giving priority to preventing truly catastrophic or irreversible outcomes that harm everyone over gaining marginal competitive edges or parochial advantages. From the objective standpoint, avoiding an AI-induced catastrophe is not just a consideration of one interest alongside others; it is a constraint that structures what counts as a reasonable strategy at all. Weighing scenarios according to their global disvalue encourages U.S. strategists to (1) treat the avoidance of catastrophic failure as effectively nonnegotiable and (2) see competitive ambitions as legitimate only within a space that does not materially increase such risks.[14]

---

[12] Note that the agent has shifted here from the individual to the state: Nagel's examples locate the reason in a person, whereas a national strategy is agent-relative because its reasons make essential reference to the state itself. The formal structure is the same; I defend treating the state as the relevant agent in Objection 4.

[13] These reasons for incorporating the agent-neutral standpoint are derived from Nagel's account as applied to the specific features of a U.S. AI strategy; the reasons are not drawn directly from existing governance frameworks for other transnational risks. That said, each reason has parallels in other domains (I present details of these arguments in this section). For example, the argument that catastrophic risk should function as a constraint on permissible strategy echoes the logic of nuclear no-first-use doctrines. The legitimacy argument has antecedents in literature on international institutions and regime theory. The coordination argument draws on the logic of arms control.

[14] By *disvalue*, I mean *negative value*, or the badness of an outcome considered impartially. The term is standard in moral philosophy and is used here to capture the idea that some outcomes (such as civilizational catastrophe) are not merely undesirable from a particular standpoint but bad from any standpoint.

A second argument for the agent-neutral standpoint is that it can help policymakers build legitimacy and trust. A strategy seen as purely self-serving will breed suspicion and might fail to gain broad support. By contrast, grounding a U.S. AI strategy in openly impartial rationales can support international legitimacy, even if agent-relative considerations will, in practice, continue to shape how countries act. My claim is not that impartial reasoning will override national interest—the history of international cooperation, such as chemical weapons conventions and climate agreements, suggests that national interest is rarely overridden by impartial reasoning alone. Rather, the point is that using impartial rationales can lead to the United States creating a framework in which cooperation becomes easier to sustain; such a framework provides reasons that other countries can recognize as valid rather than as mere expressions of U.S. strategic advantage. Because agent-neutral reasons are, by design, recognizable "from anywhere" or by anyone, invoking them in U.S. AI policy would signal that the policy is anchored in shared interests. This could reduce the likelihood of countercoalitions developing and make it easier to sustain cooperative governance arrangements.

A third argument for the agent-neutral standpoint is that it can help policymakers avoid self-defeating dynamics of the AI race. From a narrowly national viewpoint, it may seem rational to try to outpace rival nations in developing AI. However, the objective view reveals that uncoordinated dynamics of an AI race are jointly self-defeating. If all parties race to dominate the field of AI, all parties might end up undermining safety in ways that increase the chance of disaster for everyone. The impartial perspective reframes the AI race not only as a race occurring in a competitive arena but also as a race with a coordination problem: What policies would all prudent actors want the other actors to adopt given the shared risks? For example, if the United States unilaterally commits to certain testing standards or deployment limits that any responsible country should follow, it creates a template that other countries can adopt.[15] The goal is to move from a prisoner's dilemma to a coordination equilibrium in which the stability of an arrangement stems from each side's recognition of the benefits of cooperation, not merely from deterrence or dominance.

Preserving Special Obligations to Citizens

So far, my emphasis has been on the importance of incorporating an impartial, agent-neutral perspective into a U.S. AI strategy. Yet there are also robust normative and pragmatic reasons (beyond mere political expedience) to take a country's agent-relative standpoint seriously. The United States is not an abstract welfare maximizer but a political agent answerable to specific citizens, institutions, and constitutional commitments. A plausible account should, therefore, explain why special obligations to compatriots remain morally and politically weighty even under conditions of humanity-wide risk.

Political theorists have offered principled arguments for why countries may legitimately prioritize the interests of their own citizens. I draw on two theorists in particular—David Miller and Robert

[15] The feasibility of cooperation between great powers varies considerably depending on the domain, the perceived risks, and the asymmetry of impact. For an empirical assessment of when such cooperation is and is not achievable, see Raphael S. Cohen, Marta Kepe, Nathan Beauchamp-Mustafaga, Asha Clark, Kit Conn, Michelle Grisé, Roby Valiaveedu, and Nathan Vest, *Assessing the Prospects for Great Power Cooperation in the Global Commons*, RAND Corporation, RR-A597-4, 2023.

Goodin—because they approach the question from different directions.[16] Miller argues from within a nationalist framework and Goodin from within a cosmopolitan one. That they converge on the same basic conclusion—that special duties to compatriots are morally justifiable—strengthens the claim that agent-relative obligations deserve serious weight regardless of one's stance in the broader debate.

Miller, for example, defends the idea that countries are ethically salient communities. National membership is a relationship made up of shared political institutions, a public culture, and patterns of cooperation and mutual trust over time. In this view, compatriots stand in special associative relations to one another, much like family members or co-authors of a joint project. Those relationships generate obligations that are thicker than duties to strangers: obligations of fair burden-sharing, protection, and loyalty that mirror the particular history of cooperation within a political community. For a democratic country, privileging the security and welfare of its own citizens is the institutional expression of these associative duties. To treat compatriots as if they do not have any special claim would be, in Miller's view, to neglect the moral significance of the national association itself and to fail in providing the fair terms of cooperation that make political life possible.

Goodin, by contrast, starts from a cosmopolitan outlook, but he still concludes that special responsibilities to compatriots can be justified. In his account of assigned responsibilities, general duties of beneficence are more effectively fulfilled if specific agents are tasked with looking after particular individuals or groups. Countries are natural bearers of such assigned responsibilities. They are uniquely placed to protect the interests of their own citizens through the use of laws, policies, and coercive power. In this view, special duties to compatriots are not ultimate moral permissions to discount outsiders; rather, they are institutional devices for ensuring that everyone, somewhere, has a designated protector. A government that takes its role-specific responsibilities seriously is thereby helping realize universal moral goals, provided that the government does not interpret its special duties as licenses to disregard the basic claims of non-compatriots.[17]

I invoke Miller and Goodin not as decisive authorities but as two widely cited and distinct approaches for showing that special duties to compatriots can be defended on moral terms. Whether or not one accepts either view in full, both views support the narrower claim that duties to citizens are not simply political obstacles to impartial morality; instead, they are part of the moral landscape. These agent-relative justifications have direct implications for a U.S. AI strategy. If national membership generates associative or assigned responsibilities, then U.S. policymakers have distinct duties to those in the United States, such as to shield U.S. citizens from AI-related harms, manage disruptive economic transitions domestically, and ensure that AI systems deployed in U.S. jurisdictions reflect constitutional rights and democratic values. And extending this view, U.S.

---

[16] For Miller's associative account, see David Miller, *National Responsibility and Global Justice*, Oxford University Press, 2007; for Goodin's assigned-responsibility account, see Robert E. Goodin, "What Is So Special About Our Fellow Countrymen?" *Ethics*, Vol. 98, No. 4, July 1988.

[17] Both Miller's (2007) and Goodin's (1988) arguments are most naturally at home in democratic countries, where political institutions are at least formally structured around accountability to citizens. Whether authoritarian states generate the same kind of obligations is contested. On Miller's associative account, the answer depends on whether the political community in question sustains relations of mutual trust and cooperation—something that is, at a minimum, more attenuated under authoritarian rule. On Goodin's assigned-responsibility account, the question is functional rather than normative. Even an authoritarian government might be the most effective protector of its citizens' interests, although the absence of democratic accountability weakens the moral justification for treating its claims as authoritative. In this paper, I focus on U.S. strategy and therefore on the democratic case, but the limitation should be noted.

policymakers may also have obligations to maintain sufficient AI capabilities—defensive, regulatory, and, as necessary, military—to protect the U.S. population against coercion or exploitation by foreign actors wielding advanced AI.

There are also pragmatic reasons to preserve a strong agent-relative strand in AI policy. Domestic political legitimacy depends on citizens perceiving that their government is oriented toward their safety and flourishing. A strategy that is presented as "humanity first" but, in reality, visibly neglects local vulnerabilities—such as labor displacement or unequal access to AI benefits—will struggle to command durable support, no matter how compelling its impartial rationale. A country's AI strategy that cannot be credibly framed as advancing the interests of that country's own citizens is, in any democratic society, unlikely to win lasting support and might even provoke backlash against both AI safety and international cooperation.

Moreover, agent-relative motivations can be harnessed to support rather than undercut global safety. Governments are often willing to incur costs and accept constraints when such costs and constraints can be presented as necessary for protecting the governments' own people. A strategy that openly acknowledges special duties to citizens can present cooperation, restraint, and transparency not as altruism but as forms of national self-protection in an interdependent technological environment. In this way, the agent-relative standpoint supplies motivational resources that a purely impartial view from nowhere often lacks.

These normative and pragmatic considerations suggest that a country's agent-relative perspective should not be treated merely as a constraint placed on an otherwise impartial moral ideal. Rather, the perspective should be recognized as one source of moral reasons that an AI strategy should integrate.

Handling Hard Trade-Offs Between Global and National Reasons

It is tempting to hope that agent-neutral and agent-relative reasons will usually align—that what is best for the world will also be best for the United States. In many cases, such an alignment might occur. In the abstract, reducing the probability of an AI-induced catastrophe is both a global imperative and a core U.S. interest. But some decisions will present hard trade-offs that will pull global and national reasons in different directions. A dual-standpoint strategy must therefore offer not just high-level principles but also a way of thinking about such conflicts.

Consider the following example. Suppose U.S. researchers develop a technique that dramatically improves the robustness of frontier AI systems against misalignment and misuse. However, the technique is also easily repurposed to accelerate capabilities development. Releasing the technique broadly might reduce global catastrophic risk by enabling more actors to build safer systems. At the same time, it would likely increase the proliferation of powerful AI, including by adversaries, and could erode U.S. strategic advantages. From an agent-neutral standpoint, there is a strong reason to share safety-enhancing tools widely. From an agent-relative standpoint, there is a strong reason to withhold or tightly control them.

Or take a second example. Imagine that U.S. leaders face a choice between (1) pressing ahead rapidly with domestic AI development to secure economic and military advantages despite concerns about whether safety techniques have fully matured and (2) adopting a self-imposed slowdown on certain large-scale training runs until more evidence accumulates—but doing so at the cost of

potentially falling behind a rival. The impartial perspective emphasizes a global disvaluation of increased large-scale risk and, all else being equal, supports caution, even if other countries do not yet reciprocate. The agent-relative perspective, meanwhile, emphasizes the dangers of strategic vulnerability if the United States restrains itself while others do not.

A dual-standpoint framework does not dissolve these tensions, but it can structure how they are approached. Two simple heuristics help:

- **A lexical priority for catastrophic risk.**[18] When a policy choice materially affects the probability of global catastrophe or irreversible harm on a civilization scale, agent-neutral reasons have presumptive lexical priority. In such cases, the United States should treat the avoidance of such catastrophic outcomes as part of what makes a permissible national strategy, not merely as a consideration. If a policy significantly raises the risk of large-scale failure, potential national gains are not an adequate justification to include that policy in a national strategy.[19]
- **Explicit balancing within the safety envelope.** For policies with this catastrophic-risk constraint, agent-relative reasons properly regain substantial weight. For such policies, decisionmakers might explicitly consider trade-offs between, for example, U.S. economic resilience and incremental changes in global risk. The policy framework requires that decisionmakers identify which considerations in a given trade-off are agent-relative and which are agent-neutral; decisionmakers should not treat the decision as if it involved only one kind of reason. For example, a choice to restrict exports of a particular AI component might serve an agent-relative interest by maintaining technological advantage *and* an agent-neutral interest by limiting proliferation—but the choice might also impose agent-neutral costs by slowing the diffusion of safety-relevant tools. Naming these reasons separately is what makes it possible to weigh them against each other.

When applied to the safety-technique example mentioned previously, these heuristics suggest first asking whether the technique's broad release materially increases the chance of global catastrophe or whether it chiefly reallocates capabilities. If the release substantially raises the catastrophic risk (for instance, by enabling many more actors to build systems that cross dangerous capability thresholds), the agent-neutral view would recommend withholding or tightly sequencing the technique's dissemination. If the main effects are redistributive, the United States has more latitude to prioritize agent-relative concerns.

---

[18] By *lexical priority*, I mean something stronger than simply *very high priority*. A consideration has lexical priority when, in normal circumstances, it must be satisfied before other considerations are allowed to play any role in decisionmaking. In this context, the claim is that avoiding outcomes that would irreversibly destroy or radically curtail humanity's long-term prospects should be anyone's first objective, and only once those catastrophic risks are acceptably managed do trade-offs among narrower national advantages properly arise.

[19] The difficulty, of course, is that the probability of global catastrophe from any given AI policy choice is rarely known with precision. The lexical-priority heuristic does not assume such precision. It asks instead whether a reasonable assessment of the evidence suggests that a policy choice *materially* affects the probability of civilizational harm. If there is genuine uncertainty about whether the catastrophic threshold is crossed, the heuristic defaults to caution—not because caution itself is decisive but because the consequences of being wrong in each direction (that the policy leads to imposing an unnecessary constraint versus starting an irreversible catastrophe) warrants it.

These heuristics are simple by design. They neither exactly identify the catastrophic-risk threshold nor provide precise weights for competing considerations. But they clarify the structure of the problem and help avoid two common errors: treating global risk as just another national interest among many and treating national duties as mere political obstacles to impartial morality.

Objections and Counterarguments

There are limitations to the arguments I have presented. In this section, I advance the strongest objections and offer refutations where possible.

Objection 1: Long-Run National Interest Already Subsumes Catastrophe Avoidance

Perhaps the most salient objection is that the agent-neutral layer is redundant. If an AI-induced catastrophe would devastate the United States, then a well-calibrated national-interest function already generates the same prescriptions. The impartial standpoint adds philosophical vocabulary without adding strategic content.

This objection is strongest as an argument about ideal-case deliberation. If U.S. institutions reliably optimized a coherent, long-horizon national interest across administrations, agencies, and information environments, the agent-neutral framing would add little. The more-pressing empirical claim is that such coherence is structurally difficult to sustain. Domestic political incentives discount long-horizon risks, agencies with operational responsibilities for competitive positioning have built-in reasons to minimize diffuse global externalities, and the very uncertainty that makes catastrophic outcomes easy to discount in any given policy cycle makes decisions especially easy to defer. When risk avoidance is framed only as an implication of national interest, such avoidance is vulnerable to being dismissed precisely when competitive pressure is highest. When risk avoidance also holds independent standing as an agent-neutral constraint, it is harder to displace that avoidance by repackaging competitive ambition as strategic prudence. The distinction matters most during crises when the distinction is most likely to be abandoned.

Objection 2: The Framework Gives No Way to Identify What Counts as Catastrophic

A second objection is that the lexical-priority heuristic only works if actors can reliably distinguish genuinely catastrophic risks from serious-but-recoverable ones. But that line is exactly what is contested. The framework appears to push the hard problem back a level without solving it. A theory of what *constitutes* a crossed threshold is needed, and the dual-standpoint structure does not supply one.

The objection is correct that the framework does not resolve threshold disputes, but this is not a defeating problem. No decisionmaking procedure eliminates contested empirical judgment; the question is whether the procedure structures that judgment well. What the lexical-priority constraint does is shift the burden of proof. The default stance is caution, and the onus falls on those who claim a

risk does not cross the threshold to make that case explicitly and in public. That is a meaningful procedural achievement even without a bright line. Developing shared technical and political criteria for what constitutes civilization-scale risk is real and necessary work, but it is downstream of getting the decision structure right, not a precondition for adopting it.

Objection 3: Giving Explicit Standing to Agent-Relative Reasons Normalizes Competitive Racing

Those concerned with global safety might worry that the dual-standpoint perspective legitimizes exactly what it claims to constrain. By formally recognizing that U.S. policymakers have agent-relative reasons to pursue competitive advantage, the dual-standpoint structure provides rhetorical cover for actors who want to accelerate development while gesturing at safety constraints.

This is a serious worry and not fully answerable in the abstract. But it misidentifies the alternative. Agent-relative reasons are already operative in national strategy whether or not a framework acknowledges them. Suppressing them in the vocabulary of strategic argument does not eliminate them institutionally; it removes the analytical apparatus needed to check them when they operate covertly. A structure that names agent-relative reasons and subjects them to a lexical constraint on civilization-scale risk creates a point of contestation—a place where someone can say, "you have not shown that this falls within the safety envelope"—that simply does not exist in frameworks that treat national interest as the only operative consideration. Whether those procedural checks can be made robust is an empirical question about institutional design.

Objection 4: The State Is Not a Unitary Agent, So the Framework Misdescribes Policy Choice

Another strong objection is that the agent-relative/agent-neutral distinction, drawn at the level of a single deliberating agent, does not survive the transition to collective actors. Modern governments are coalitions of agencies, constituencies, and bureaucratic interests with partially incommensurate objectives. The fragmentation operates at two levels: different agencies press different reasons—the Department of State and the Department of War may want different things—and within any single agency, individuals can line up behind competing positions for competing reasons.[20] There is no stable "view from inside" the state at either level, and a framework built on the dual standpoints of a unified deliberator may project the wrong ontology onto the problem.

The objection is correct that the state is no unitary deliberator—but the framework does not need it to be. I do not claim that "the United States" deliberates as a single person, and any application that treats internal variation as noise would mislead. The distinction sorts kinds of reasons, and that sorting holds whether the dispute is between agencies, within one, or inside a single official's deliberation. Disaggregation changes when the distinction does its work, not whether it does any.

[20] The Department of War is designated the Department of Defense under Public Law 81-216, National Security Act Amendments of 1949.

Objection 5: Under Conditions of Uncertainty and Competition, Agent-Relative Reasons Will Predictably Override Agent-Neutral Ones

The final objection is that the dual-standpoint framework is too contextual in its application. When catastrophic risk is clear—when there is a visible, near-certain path to civilizational harm—the case for lexical priority is compelling. But the more realistic scenario is one in which the risk exists within a range of possible severities, experts contest said risk, and the risk coexists with intense competitive pressure. In such conditions, the objection goes, agent-relative reasons will predictably dominate. Policymakers facing uncertain global risks and concrete national vulnerabilities will default to engaging in competition, and the lexical-priority constraint will function as rhetoric rather than as a way to operate.

This objection is largely correct as a prediction about institutional behavior. The question is whether the framework makes that default state *harder* or *easier* to sustain. Without an explicit dual-standpoint structure, competition becomes the default, and there is no recognition of what it displaces. With the dual-standpoint structure, an actor who prioritizes competitive advantage under conditions of ambiguous catastrophic risk must at least *argue* that the risk does not cross the catastrophic threshold; this argument is subject to scrutiny, contestation, and revision as evidence accumulates. The framework does not prevent agent-relative reasons from prevailing in hard cases; it ensures that such reasons prevail (when they do) through an explicit judgment rather than by omission. Whether that procedural discipline is sufficient is an empirical question. But it is a stronger position than one in which the tension is never surfaced at all.

Conclusion

The argument I make in this paper is deliberately limited. I have not resolved the tension between agent-neutral and agent-relative reasons in a U.S. AI strategy—that kind of tension is not what a framework resolves. I have instead given the tension a structure precise enough to reason through.

The strategist's dilemma will not diminish as AI systems grow more capable. The pressure to move fast and the reasons for caution are likely to intensify together. Policymakers will face choices—about AI deployment, about the dissemination of safety-relevant research, about the pace of development itself—in which the obligations they bear to their own citizens and the obligations they bear to the whole world do not obviously converge. The question is whether those choices are made within a framework that names both kinds of reasons and specifies their relationship or whether they are made within a framework that recognizes only one kind of reason and absorbs the other into it by default.

The dual-standpoint framework I advance in this paper offers two commitments in response to those questions. The first is that catastrophic and irreversible risk functions as a side constraint on permissible strategy—that is, not as one consideration to be traded against competitive advantage but as a condition that must be satisfied before such trade-offs properly arise. The second is that, within the space defined by the catastrophic risk constraint, agent-relative reasons retain moral and political weight; they are not concessions to expediency but part of the normative landscape that any defensible strategy must navigate.

These two commitments do not prescribe specific policy outcomes. What they do is clarify the kinds of reasons to which a strategy must respond and the order in which those reasons bear on a decision. In a domain in which safety advocates and national security practitioners often talk past one another—each treating the other's central commitments as naive or as obstacles—that clarification has practical value. Naming both kinds of reasons will not stop them from disagreeing, but it can stop them from disagreeing about the wrong thing. The harder problems are institutional: What decision procedures, review mechanisms, and evaluative criteria would give the dual-standpoint structure operational force within the agencies and processes that actually shape AI policy? Plausible options already exist in some form; for example, interagency review processes, published risk assessments that an agency must defend, safety-institute evaluations, and reporting requirements to Congress. The conceptual work I have done in this paper is a precondition for that institutional work, not a substitute for it.

Bibliography

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

Cohen, Raphael S., Marta Kepe, Nathan Beauchamp-Mustafaga, Asha Clark, Kit Conn, Michelle Grisé, Roby Valiaveedu, and Nathan Vest, *Assessing the Prospects for Great Power Cooperation in the Global Commons*, RAND Corporation, RR-A597-4, 2023. As of May 21, 2026: https://www.rand.org/pubs/research_reports/RRA597-4.html

Goodin, Robert E., "What Is So Special About Our Fellow Countrymen?" *Ethics*, Vol. 98, No. 4, July 1988.

Hendrycks, Dan, Eric Schmidt, and Alexandr Wang, *Superintelligence Strategy: Expert Version*, arXiv, arXiv:2503.05628v2, April 14, 2025.

Kant, Immanuel, *Groundwork of the Metaphysics of Morals*, ed. and trans. by Mary Gregor and Jens Timmermann, Cambridge University Press, 2012.

MacIntyre, Alasdair, *After Virtue: A Study in Moral Theory*, 2nd ed., University of Notre Dame Press, 1984.

Mazarr, Michael J., *A New Age of Nations: Power and Advantage in the AI Era*, RAND Corporation, PE-A3691-14, January 2026. As of May 21, 2026: https://www.rand.org/pubs/perspectives/PEA3691-14.html

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

Mill, John Stuart, *Utilitarianism*, Parker, Son, and Bourn, 1863.

Miller, David, *National Responsibility and Global Justice*, Oxford University Press, 2007.

Nagel, Thomas, *The Possibility of Altruism*, Clarendon Press, 1970.

Nagel, Thomas, *The View from Nowhere*, Oxford University Press, 1986.

Nagel, Thomas, *Equality and Partiality*, Oxford University Press, 1991.

Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity*, Grand Central Publishing, 2020.

Parfit, Derek, *Reasons and Persons*, Clarendon Press, 1984.

Pogge, Thomas W., *World Poverty and Human Rights: Cosmopolitan Responsibilities and Reforms*, Polity Press, 2002.

Rehman, Iskander, Karl P. Mueller, and Michael J. Mazarr, "Seeking Stability in the Competition for AI Advantage," commentary, RAND Corporation, March 13, 2025. As of May 21, 2026: https://www.rand.org/pubs/commentary/2025/03/seeking-stability-in-the-competition-for-ai-advantage.html

Scheffler, Samuel, *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*, Clarendon Press, 1982.

Sidgwick, Henry, *The Methods of Ethics*, 7th ed., Macmillan, 1907.

Singer, Peter, "Famine, Affluence, and Morality," *Philosophy & Public Affairs*, Vol. 1, No. 3, Spring 1972.

Waltz, Kenneth N., *Theory of International Politics*, McGraw-Hill, 1979.

Walzer, Michael, *Spheres of Justice: A Defense of Pluralism and Equality*, Basic Books, 1983.

About the Author

Beba Cibralic is a policy researcher at RAND and a professor of public policy at the RAND School of Public Policy. She focuses on AI governance, safety, and strategy. She holds a Ph.D. in philosophy.