MANHATTAN PROJECT TRANSFORMATIVE RACE FOR ULTRAINTELLIGENCE

T.R.U.M.P.

MAKE AI GREAT AGAIN

Securing America's AI Future

The Next Manhattan Project

Launch the Most Innovative Scientific Initiative in American History:

1. Launch T.R.U.M.P. AI to ensure America's massive AI investments—like the $500B Stargate initiative—deliver lasting American dominance, not just temporary advantages.
    - We must guarantee these unprecedented resources serve our national interests rather than China's.

2. Secure an American victory in the AI race by solving the AI alignment challenge:
    - "Alignment" means building AI systems that reliably and permanently uphold our American values and intentions.

3. Investments in AI alignment will also drive AI capabilities and American economic growth.
    - It is because of alignment research that today's powerful AI systems exist in the first place.
    - Make American AI leadership permanent through continued breakthrough alignment research that simultaneously drives major capability and economic advances here in America.

Move Beyond Failed Containment Strategies:

1. DeepSeek's sudden release of powerful open-source AI proves we can't win through export controls, censorship, or secretive research alone. Aligning AI to American values gives the US a strategic advantage as AI capabilities inevitably spread.

2. Fund ambitious AI researchers pursuing neglected, breakthrough approaches. History shows transformative advances come from bold thinkers challenging the establishment.

3. Reject the false choice between AI alignment and progress. The evidence proves alignment research drives our biggest innovations—pioneering techniques that make AI systems both more powerful AND more reliable in the long run.

Turn America's AI Lead into Permanent Dominance:

1. Position America as the global standard-setter for AI development. When others inevitably copy our work, our mastery of alignment ensures we shape how AI evolves worldwide.

2. Scale up promising private sector alignment research through public-private partnerships. Current mainstream approaches are failing to keep pace with exponential AI progress.

3. Use Trump's proven deal-making talents to leverage shared international interests in AI control. Even strategic competitors like Xi fear losing control of advanced AI systems.

Critical Context:

- While Biden wasted time on diversity training for robots, Trump can launch America's greatest and strongest scientific initiative: T.R.U.M.P. AI.
- AI leaders now predict superintelligent systems within 2-3 years, not decades
- Even the world's top AI scientists don't understand how these artificial brains work
- The race isn't just to build AI first, but to ensure it remains under American control
- Only Trump's deal-making skills can achieve what others think impossible

Full Briefing Follows Below

Executive Briefing Summary

There is an unacceptably nontrivial chance that advanced AI systems disempower or extinguish humanity. The core reason is alarming: we lack a fundamental understanding of the internal workings and decision-making processes of these increasingly powerful systems.

Troublingly, those who are the closest to the cutting-edge in AI are also the people most worried about this possibility. Conservatives have a proven track record of "thinking about the unthinkable" and addressing frightening realities like this one. To do this, however, conservatives must avoid getting fooled into adopting the left's shortsighted AI playbook around identity politics and overregulation.

Instead, we must launch an ambitious American initiative: **the Transformative Race for Ultraintelligence Manhattan Project (T.R.U.M.P.)**—a massive research program that simultaneously renders American AI both highly capable and fundamentally controllable. This entrepreneurial strategy will drive massive economic growth while ensuring AI serves American interests instead of threatening them. With the current funding landscape dominated by left-leaning sources, these ambitious approaches remain dangerously underexplored. The core issue of aligning advanced AI with human values and avoiding potentially existential risks requires serious and immediate attention from conservative thinkers and policymakers. These efforts will enhance innovation rather than constricting it. By taking the lead in this effort, America can achieve what others think impossible: ensuring AI systems remain powerful AND under American control.

1. Disentangling "AI Alignment" from "Woke AI"

Currently, the left is attempting to peddle 'equitable' AI initiatives under the banner of "building safe AI" and conservatives are at risk of following their lead in conflating safe AI with the far narrower question of whether AI outputs are ideologically biased in one or another direction.

It's crucial to understand that, in spite of this behavior from the left, the core concept of reducing the chance of an AI catastrophe cannot and should not become partisan. Nonetheless, five Democrats (and no Republicans) recently signed onto a letter to OpenAI inquiring about their safety procedures; we suspect that six months ago, this kind of effort likely would have had bipartisan support.

Just because the political and corporate left seems willing to co-opt the banner of AI safety to push identity politics, fringe ideologies, and brand safety doesn't mean conservatives should also sink to this level and disregard the real risks. The 2024 Republican Party platform currently reads more as a reactionary response to the left's woke AI nonsense than a visionary, forward-looking account of the existential risks posed by this technology. The intellectual right is well-positioned to understand the very real risks involved and guide America and humanity towards a safer future rather than playing into the left's distracting and dangerous AI narratives.

It is also important to note that the current funding landscape for AI alignment research is extremely limited. The overall lack of an indigenous conservative alignment movement leaves a significant gap in addressing crucial AI alignment concerns from a conservative perspective. This imbalance not only restricts the range of approaches to AI alignment but also risks overlooking critical considerations that align with conservative values and principles.

2. Understanding the AI Alignment Problem

The field of AI alignment asks—and ultimately aims to answer—a fundamental question: how do we ensure AI systems of the near future will not disempower humanity? This concern is not hyperbole or science fiction: it is the highly-well-informed worry of the godfathers of modern AI like Geoffery Hinton and Yoshua Bengio; the top scientists at leading AI companies like Ilya Sutskever (ex-OpenAI), Shane Legg (DeepMind), Igor Babuschkin (xAI), and Dario Amodei (Anthropic); and hundreds of other well-respected AI experts.

Nonetheless, three key ongoing technical challenges and concerns related AI alignment include:

1. Humanity's extremely limited understanding of the internal workings of modern machine learning/AI models.
    a. By analogy, we understand exactly how iPhones, bridges, and combustion engines work because human engineers designed every part of these systems. By contrast, we do not know how modern AI systems work because they are *trained* or *evolved* rather than designed. All we know is *that* they work, rather than *how.* This is unprecedented and dangerous given the steadily increasing power of this technology.
    b. Current architectures of AI systems are 'black boxes' and thus may never be fully understandable. Conservatives ought to support strong investments in American innovation to discover new AI architectures that are safer and more controllable.

2. Mesaoptimization, where AI systems may develop their own internal optimization processes that diverge from the ones humans have set up and implemented, even when they appear to be behaving as expected right now. Given that there is no way to know what is going on inside of these models (see point 1), it is currently impossible to determine whether or not this is happening.
    a. One specific concern related to mesaoptimization is "deceptive alignment," where AI systems may appear aligned to humanity initially but ultimately reveal misaligned objectives once they gain more power or autonomy.
3. Instrumental convergence, which is the idea that giving an intelligent system any arbitrary goal (e.g., ‘vacuum the floor’) may cause it to take on unrelated and dangerous goals that are ‘instrumental’ for achieving the main goal (e.g., ‘if I get shut off, I can’t vacuum the floor—so I will resist getting shut off’). Goal-directed biological systems naturally exhibit similar instrumental behaviors like avoiding ‘shutoff’ and acquiring future resources, and the concern is that we are on the cusp of being AI systems with the same property.

Beyond these technical challenges, there's a critical issue in the current research landscape. A recent survey we conducted of current alignment researchers reveals widespread pessimism about solving alignment in time, with many believing that current approaches are insufficient. Paradoxically, despite this, most choose to focus on and fund 'non-risky' research areas like mechanistic interpretability, driven more by career incentives than by the likelihood of solving core alignment issues. This shortsightedness is exacerbated by a misguided reluctance to invest in critical American innovation, stemming from an irrational fear of advancing capabilities. Ironically, this hesitation may be hindering our ability to address the very problems it seeks to avoid.

It is very important to note that these concerns around AI development are different in kind from the oft-overblown hype that ordinarily surrounds new technologies (e.g., the nanoscience fears of the early 2000s, cryptocurrencies, NFTs, etc). The most straightforward evidence as to what sets AI apart are scaling laws, which show that AI systems predictably get dramatically more capable as we increase their size and the amount of data they're trained on. AI development has historically followed these scaling laws and clearly and measurably continues to do so—each new iteration of large language models, for instance, demonstrates capabilities that were considered far-fetched just a few years prior. Humans suffer from “exponential slope blindness” which makes it hard to appreciate just how much things are poised to change soon.

While the left now seems stuck seeing everything through the sole lens of diversity, equity, and inclusion, conservatives are poised to grapple with the inevitability that we may soon be among AI systems that reach or exceed human-level intelligence.

This world could be extraordinarily dangerous if these AIs are not aligned with human values, and conservatives should not hide from this reality as the left has. Instead of just implementing regulations or safety measures that hamper or weaken AI, America should pursue alignment research that makes these systems *more* capable by virtue of their alignment.

Contrary to the prevailing narrative, we firmly reject the notion that safety features act as a 'tax' on AI capabilities. Instead, we propose a paradigm shift: the concept of a 'negative alignment tax.' This theory posits based on the evidence to date that the most effective alignment work actually enhances these systems' capabilities precisely because it makes them safer. The false dichotomy between safety and economic growth, often propagated by left-leaning safety advocates, is not only misleading but potentially harmful to progress. In reality, alignment research has demonstrated up to this point a remarkable tendency to generalize even further than capabilities research. History has shown that the most significant breakthroughs in alignment have consistently driven the greatest advances in capabilities, which in turn have fueled economic growth and innovation.

This synergy between safety and capability is not just theoretically sound—it's a practical necessity for maintaining America's competitive edge. We strongly believe that American leadership in AI alignment will not only solidify our nation's preeminent position on this critical issue but also act as a powerful catalyst for unprecedented economic prosperity at home. By investing heavily in alignment research, we're not just safeguarding our future; we're actively shaping it, ensuring that the United States remains at the forefront of the AI revolution while simultaneously addressing crucial safety concerns.

In essence, pursuing alignment is not a burden on progress—it is the very engine of progress. It is time we recognize that safety and innovation are not competing interests, but rather two sides of the same coin in our pursuit of transformative AI technologies.

This principle extends to the debate surrounding open-source AI, as highlighted by influential right-leaning figures like Andreessen and Thiel. We suspect there is an emerging false dichotomy in this space between alignment and open-source development. In fact, open-source practices have been instrumental in advancing alignment research, with many of the field's biggest breakthroughs occurring after the advent of open-source AI models. Our stance is clear: we advocate for as much open-source development as possible, with the caveat that the latest frontier models should be vetted to ensure they don't pose major threats before release. This approach not only accelerates innovation but also democratizes access to American researchers and developers of AI technology, further cementing the US’s leadership in this space.

Underestimating AI's potential based on its current limitations would be a grave mistake. The fact is, AI is advancing at breakneck speed and will likely surpass human intelligence in many areas within the next five to ten years. Even field experts were caught off guard by the capabilities of today's language models. While humans naturally think in linear terms, AI progress follows an exponential curve. This tendency to misjudge exponential trends—a phenomenon researchers call "exponential growth bias"—leads many to erroneously dismiss these rapid timelines. Such a misconception poses serious risks. Conservatives, in the tradition of Herman Kahn, have always excelled at 'thinking the unthinkable.' We must now confront an uncomfortable truth: we simply don't know if AI will pose an existential threat in the near future. Now, we bear the responsibility to recognize this reality and implement concrete measures to address it. America's future hangs in the balance; we must not allow shortsightedness to jeopardize it.

3. Imminent AI Risks

While longer-term risks are of paramount importance, there are also potentially existential concerns that must be addressed in the shorter term. Anthropic CEO Dario Amodei, testifying to the US Senate, warned that, within 1-2 years from today, AI may be capable of causing "large-scale destruction." This urgent timeline underscores the immediate nature of the threat.

Consider a scenario where malicious actors gain access to AI models capable of providing instructions for synthesizing highly destructive pathogens using readily available materials. This is not mere speculation; a reliable internal source on the team stress-testing GPT-5 claims that the model may possess such capabilities. This means a lone terrorist or disaffected young person could soon have access to a technology that would enable him to end all biological life on earth.

We are inadequately prepared for these possibilities and must think deeply about prevention strategies, even if we may currently assess the probability as low. All current AI models are vulnerable to hacking and "jailbreaking," potentially enabling bad actors to bypass safety measures. This vulnerability extends to information security concerns, where compromised AI systems with access to sensitive data could lead to unprecedented privacy breaches or intellectual property theft.

The proliferation of open-source AI models presents another risk, as it democratizes access to powerful AI capabilities without necessarily providing adequate safeguards or oversight. This could exacerbate the potential for misuse by non-state actors or terrorists, who could leverage AI to enhance their capabilities in areas such as cyber attacks, disinformation campaigns, or even the development of autonomous weapons.

4. Strategic Policy Response:
The T.R.U.M.P. Manhattan Project for AI

These immediate existential risks are compounded by humanity's inherent "exponential slope blindness," as described earlier. Stated directly: *we think we have more time than we actually do to solve the alignment problem*. While current alignment efforts focus on incremental improvements through methods like Reinforcement Learning from Human Feedback (RLHF), the field is rapidly outpacing our ability to ensure its safety. Our recent survey of alignment researchers confirms this assessment, with most believing current approaches are insufficient to solve the underlying problem in time.

The solution must match the exponential nature of the emerging threat. **At the policy level, we propose immediately launching the T.R.U.M.P. Manhattan Project for AI, directing massive funding towards winning the AI race and pursuing a portfolio of neglected, ambitious, moonshot approaches to alignment research.** Exploring multiple unconventional and underexplored strategies simultaneously maximizes our chances of solving the core problems as quickly as possible. This includes researching novel architectures that are inherently more controllable, pursuing interdisciplinary approaches combining insights from cognitive and computer science, and fostering entrepreneurially-inspired technical innovations that have been systematically overlooked by the research establishment.

This strategy is not theoretical. AE Studio has successfully implemented this portfolio approach to alignment research, redirecting substantial consulting profits into multiple parallel research tracks. Early results show remarkable promise, including multiple breakthrough results at the intersection of AI and neuroscience. This entrepreneurial model—pursuing multiple ambitious approaches simultaneously while maintaining rigorous testing and validation—has proven both scalable and effective.

However, to address the exponentially growing risks, this approach must be rapidly scaled across both private and public sectors. For high-risk AI developments, government oversight will be important, but it must be designed to accelerate progress, not impede it. The goal must be to speed up innovation and operationalize what we know about how to make maximal progress with scientific R&D while ensuring catastrophic outcomes are prevented.

The portfolio strategy promotes maximal technological advancement and economic growth while addressing existential risks. By pursuing multiple ambitious research directions simultaneously, we maximize our chances of breakthrough solutions while ensuring America maintains its technological leadership. The evidence from our work suggests that the most effective alignment research actually *enhances* capabilities precisely because it makes systems fundamentally safer and more reliable.

Unfortunately, this urgency is not widely recognized in policy and institutional circles, creating dangerous barriers to scaling these solutions. The UN's AI advisory council has only one member out of 30 who considers Artificial General Intelligence (AGI)—an AI system that matches or surpasses human capabilities across a wide range of cognitive tasks—a realistic concern. The OECD's AI futures group doesn't address AGI at all. This institutional blindness, partly stemming from the field's eccentric early advocates, has led mainstream voices to dismiss these concerns without serious examination.

This systemic underestimation of AI's exponential progress makes our mission even more urgent. We cannot wait for institutional consensus—we must dramatically expand support for neglected approaches now, scaling successful models like AE Studio's across the research landscape. The lack of awareness and preparedness at the highest levels of global governance only heightens the imperative for informed, conservative voices to lead this critical transformation of AI alignment research.

5. A Personal Perspective on AI Alignment

As an entrepreneur with extensive experience in AI development, I've observed a critical lack of conservative voices in the field of AI alignment. I believe myself to be one of the few conservatives working on AI alignment—and the few others I've interacted with who identify as conservative are very hesitant to out themselves as such for fear of blowback. It's imperative that conservatives understand the risks to America and humanity, and take action rather than ceding ground to partisan narratives pushed by the left.

I run AE Studio, a 160+ person bootstrapped software+AI development consulting and product development business, which I started originally to create capital to invest in my vision for brain-computer interfaces. However, after starting my family and realizing that even a small chance of AI posing an existential threat would mean working in this space could have an extremely high impact, I pivoted from our company's original focus on brain-computer interfaces to working on AI alignment. Accordingly, we've redirected most of our profits into researching entrepreneurially-inspired "neglected approaches" to alignment: strategies that may individually have a low probability of success but will lead to high positive impact if any of them work. The value of this portfolio approach is much greater in expectation than the highly-risk-averse work most other alignment researchers are currently pursuing.

This resource limitation is not unique to our efforts. Across the field of AI alignment, funding constraints severely limit the exploration of novel approaches. Many organizations operate on shoestring budgets, often relying on consulting profits. With proper funding, there's potential for orders of magnitude more work in critical areas of AI alignment research. This underscores the urgent need for increased investment throughout the entire field.

It is clear from our efforts thus far in AI alignment that we need far more entrepreneurial thinking, with competent, fast-moving technical teams pursuing unconventional approaches to tackle this problem effectively. For instance, we’re currently providing essential funding to the lab of WSJ contributor and professor at Princeton, Michael Graziano to work on an alignment research agenda at the intersection of AI and the neuroscience of consciousness. Our work with Graziano has already proven impactful, opening up an avenue of research in AI alignment that, like many other non-establishment research programs, was overlooked or dismissed previously. This success underscores the need for significantly more funding and attention to foster ambitious, groundbreaking technical work that can truly solve the problem at scale.

6. National Security Implications and Global Coordination

The modern right is uniquely positioned to address the complex challenges posed by advanced AI systems given conservatives’ historical strength in navigating national security threats and technological arms races. This expertise is crucial as AI development emerges as the next major technological competition, rivaling the significance of the nuclear arms race.

As with nuclear technology, AI presents both immense potential and existential risks. However, the challenge with AI might be even more acute—potentially akin to developing a technology with a far higher risk of proverbially “igniting the atmosphere” (as was speculated in the early development of nuclear weapons) given our deep uncertainty about how modern AI systems actually work. By analogy to another dangerous scientific research program: if one thinks that gain-of-function research is ‘asking for it,’ he should remain consistent and realize that building a superintelligent AI whose internal workings we do not understand is the limit case of scientific irresponsibility masquerading as progress.

Additionally, as AI technology continues to scale, inadequate preparation could lead to a dangerous competition with global powers like China–which, with the right international posturing, the US may be able to prevent. National security is paramount, but rushing AI development without proper safeguards could also be catastrophic for everyone on Earth. We need a balanced approach that recognizes both the strategic importance of AI leadership and the existential risks of misaligned AI.

This view is increasingly shared by global leaders. Though of course it may be naive to take the CCP at face value, Xi Jinping has emerged as one surprising advocate for responsible AI development, and this is something America can leverage to its advantage. Xi recently stated that since AI will determine "the fate of all mankind", it must always be controllable. An internal party guide that Xi is said to have personally edited explicitly also conveys that China should “abandon uninhibited growth that comes at the cost of sacrificing safety.” The prospect of the CCP losing their centralized control of their own nation to an advanced AI system is clearly animating this caution, at least in part. China’s realization of the gravity of the situation is underscored by prominent Chinese AI scientists, such as Andrew Chi-Chih Yao, who recently asserted that "AI poses a greater existential risk to humans than nuclear or biological weapons." The Chinese Communist Party has even listed AI risks alongside other major concerns such as biohazards and natural disasters, signaling a growing awareness of the potential dangers.

It is also worth noting that the both US and China have already jointly agreed that “they each have a lot to lose if AI becomes weaponized or abused” and that China more generally may be more willing to manage this situation diplomatically than many initially suspected. World leaders are slowly realizing that everyone loses if an unrestrained superintelligent AI is built anywhere on Earth.

Given the nature of these risks, strong international leadership is crucial. In this context, the 2024 Republican ticket is better positioned to rise to the occasion and tackle these challenges head-on. Trump is significantly more likely than Kamala Harris to command international respect in light of his track record of successful international diplomacy like the Abraham Accords.

Along these lines, it is also worth considering what has become known as the the ‘Pottinger Paradox,’ which might be formulated as “the more you try to accommodate them, the more aggressive they become, and therefore the best path to peace or deterrence is actually to be more candid and aggressive, not only in terms of military competition, but economic and ideological competition.” This suggests that an aggressive approach may be more effective in negotiations with authoritarian states, particularly in the context of AI governance. Clearly, Trump stands a much better chance of succeeding here with China, whereas we can expect Kamala to be taken advantage of, just as Biden has been.

Only a clear-eyed conservative approach can effectively engage China and other nations on AI development, recognizing their motivations through the lens of national prestige. For China, AI is not merely about technological capability; it is about cementing their status as a legitimate superpower on the world stage. Much like nuclear weapons during the Cold War, AI has become the new prestige technology that signals a nation's global influence.

Our approach should frame responsible AI development not just in terms of safety, but as an opportunity for China to help lead in transforming the human condition for the better. By positioning the US as the country spearheading this transformation through AI, we create a compelling narrative that appeals to China's desired sense of global importance. This isn't about dry, technical agreements and discussions—it's about presenting a vision of America leading the most explosive, world-changing technology of our time, and making it safe for humanity.

Such a framing taps into China's aspirations and potential insecurities about their place in the world order. By emphasizing how responsible AI development advances humanity as a whole, we can make US leadership more appealing to China and the global south, thereby potentially preventing a dangerous AI arms race while addressing concerns about technological imbalances.

This approach leverages China's desire for prestige to "grease the wheels" of international cooperation, turning their ambitions into a force for responsible innovation rather than unchecked competition. Only through strong, principled conservative leadership can America successfully navigate this critical juncture and maintain its technological supremacy while fostering global stability.

Overall, ensuring the responsible development of advanced AI that simultaneously advances American interests requires global coordination and stringent security measures. The conservative movement can lead in developing strategies that maintain American technological superiority while mitigating global catastrophic risks from AI, leveraging our pragmatic approach to complex global challenges.

Conclusion

The current leftist-influenced discourse on AI alignment is often overshadowed by ideological concerns, distracting from the core issues and obscuring the true existential risks we face. The conservative movement has a critical role to play in addressing these challenges.

By engaging seriously with AI alignment, conservatives can help ensure that this transformative technology aligns with American values while mitigating existential risks. Our strength in addressing complex national security threats positions us uniquely to contribute to AI alignment.

We must avoid sinking into tired partisan narratives and recognize AI alignment as a non-ideological imperative. The time for decisive action is now. We call on conservative thinkers, policymakers, and leaders to engage with AI alignment research, advocate for responsible development policies, and work towards global cooperation and American leadership on AI alignment.

As we move forward, it's crucial that we expand the range of our research approaches in AI alignment. We must encourage and fund moonshot solutions, including entrepreneurially-inspired "neglected approaches" that may have a low probability of success but could lead to high positive impact if successful. This portfolio approach to research, which embraces unconventional strategies alongside established methods, is our best chance at solving the complex challenges of AI alignment and securing America's technological leadership. By investing strategically in AI alignment, we not only safeguard against potential risks but also position the United States at the forefront of AI innovation, ensuring both a safe and economically dominant future for our nation.

The stakes could not be higher. Let us rise to this challenge and shape the future of AI for the benefit of all humanity.

We Must Preserve American Values in the AI Revolution.

T.R.U.M.P. TRANSFORMATIVE RACE FOR ULTRAINTELLIGENCE MANHATTAN PROJECT

Access T.R.U.M.P.'s Website →