Existential AI risk: A Discussion of Jones (2025)

Judith Chevalier

September 2025

1 Introduction

The core takeaway of Jones (2025) seems incontrovertible: given our expectation of transformative AI within a short timeframe and given experts' forecasts of extinction-level risk from AI, society should be investing significantly to mitigate this threat. Calibrating a model using the federal Value of a Statistical Life (VSL), Jones concludes that a substantial fraction of GDP should be diverted to mitigating existential AI risk. Jones does not spell out what entity will implement this spending. A natural extension of Jones's representative agent model is to assume that each national government could implement such spending on behalf of its citizenry proportionately to the nation's GDP.

Jones bases his calibrations on the US federal value of a statistical life (VSL) and on COVID-19 spending. These benchmarks differ from the assumed nature of AI risk. First, neither COVID-19 nor typical VSL contexts involve extinction-level outcomes. To use these tools, Jones must assume that the cost of eliminating all human life equals the sum of the costs of losing each individual life, an assumption with precedent,[1] but that may not reflect actual demonstrated societal willingness to pay to reduce risk of total extinction. Second, COVID-19 mitigation and typical VSL applications (such as auto or airplane safety) involve considerable appropriability. For example, if we reduce consumption by staying home and inventing COVID vaccines, the people who stay home reduce their risk disproportionately, and the government investing in vaccines can prioritize their own citizens to receive them. While AI risk takes multiple forms, if an AI mishap can cause extinction, then a nation's investments in risk mitigation are also likely not fully appropriable by that nation. The US's historical willingness to spend to save some US lives is not necessarily indicative of willingness to spend to save all lives everywhere.

2 Asteroids

To explore these issues concretely, I consider spending to mitigate asteroid collision risk, a scenario briefly mentioned by Jones (2025). Asteroid risk parallels

[1] See Tännsjö (2023)

the nature of AI risk in key ways. First, like AI risk, an asteroid collision has the potential to cause planet-level extinction (as perhaps occurred for the dinosaurs). Asteroids also have the potential to cause a more localized catastrophe; collision with a smaller asteroid could destroy a city, though with potential widely significant climate and other effects.[2] For a so-called "city-killer" asteroid, forecasting the trajectory accurately enough to identify impact area may also be very difficult until very late.[3] Thus, as is likely the case for AI risk, investments by any nation in asteroid deflection are not fully appropriable to that nation's citizens.

In 2005, the US Congress passed a law requiring NASA to find and track at least 90% of all near-earth objects (NEOs) 140 meters or larger by 2020. Initially, this project was largely unfunded;[4] serious efforts to deflect an asteroid began in 2015 with the funding of the Double Asteroid Redirection Test (DART) and the creation of NASA's Planetary Defense Coordination Office. DART successfully demonstrated the deflection of the path of a (harmless) city-killer sized asteroid in a 2022 mission. The European Space Association's Hera mission that launched in 2024 will follow up on DART when it reaches the DART target asteroid in 2027.[5] NASA's and the ESA's efforts appear to constitute the vast majority of the planet's overall investments in mitigating asteroid risk.

The probability of substantial human destruction from an asteroid is roughly 1 in 20000 annually for a "local" mass casualty event or 1 in 700000 for an asteroid that NASA would characterize causing "possible collapse of civilization."[6] These probabilities are substantially lower than the estimate of existential AI risk in Jones (2025), though within the range of probabilities of AI extinction cited by individual experts.[7] Asteroid risk probabilities are always in flux; as individual asteroids are identified, their probability of colliding with the earth is individually assessed. The probability that a specific city-killer size asteroid, 2024 YR4, would collide with Earth peaked at 3% in February 2025.[8]

Another key difference from AI risk is that the technology to eliminate the risk is substantially better and better-defined than the current technology to mitigate AI risk. The ability of some national science agency to safely deflect an asteroid that is detected with multiple years of lead time may eventually reach nearly probability one, though it is not likely there yet. It is not clear what the probability of a successful deflection currently is, but it has increased dramatically over the last 10 years, given the success of the DART mission.[9] By my rough calculations, the US has spent less than $2bn on Planetary Defense

[2] See Titus et al. (2023) for a discussion of follow-on effects. AI causing localized disasters also seems completely plausible.

[3] Wheeler et al. (2024) and Giorgini et al. (2008)

[4] Dreier (2019)

[5] See OHB SE (2024)

[6] For this assessment, see NASA Planetary Defense Strategy Action Plan Working Group (2023) and for a more comprehensive picture of the probability-asteroid size relationship see Tedesco (2025)

[7] See Growiec and Prettner (2025)

[8] Dunn (2025)

[9] See Science Media Centre (2025)

from 2015 to 2024.[10] The spending for 2025 is roughly $300m.[11] A plausible estimate is that the risk of an asteroid collision fatal to the US or a substantial part of it is $2.4x10^{-6}$ per year absent mitigation efforts.[12] A back-of-the envelope guess is that the risk of devastating effects of such asteroids have been permanently halved by PDCO efforts (with some ongoing maintenance expenditures particularly for surveillance required).

Suppose that the technology that was developed for mitigating asteroid risk exhibited constant returns to scale. We could then calculate the value of life that rationalizes the scale of the investment that was chosen. Given these assumptions and Jones's construct of a 10-year horizon, the probability that an asteroid would wipe out the US population during the 10 years following DART was reduced by $1.2x10^{-5}$ for a generously estimated total cost of $3 billion. Given a US population of 340 million, this suggests that the implicit value of a US life used to scale the investment was a 10-year value of around $0.75 million. This is very low relative to the regulatory value of a statistical life and the benchmark suggested by Jones for this exercise of $7.5 million. It is possible that the technology is worse than constant returns to scale, or that DART was more successful than forecast ex ante, or that ongoing monitoring costs should figure into the calculations; these adjustments would raise the implied value of a statistical life. Even considering these factors, spending is likely low relative to standard VSL estimates.

This low spending exists even though the public is generally in favor of asteroid defense. A 2023 Pew survey found that only 9% of the adults surveyed believed that NASA should not undertake asteroid defense and 60% percent thought it should be NASA's highest priority.[13] Therefore, although Jones's calculations may accurately reflect normative spending benchmarks if extinction is valued as the sum of individual lives, actual US expenditure on this existential risk is far below these benchmarks.

International willingness to spend on existential risk appears even more limited. Spending on the ESA's Hera project totaled roughly $398 million over multiple years.[14] Given that both Europe and the US each cover approximately 2% of the earth's surface, the underlying risk to Europe and to the US of a city-killer asteroid strike is similar. European GDP per capita is about half that of the US and VSL estimates in the literature are correspondingly about half as large for Europe as for the US.[15] If we extrapolate from the HERA cost and call EU spending (generously) $193 million per year, EU spending is currently about half of US spending per capita. Thus, at current spending levels, Europe and the US are similarly underspending what we might expect from VSL benchmarks. Although US and EU spending seem small when we consider US

[10] See Dreier (2019)

[11] National Aeronautics and Space Administration (2025)

[12] This is the sum of the probability of an extinction level asteroid plus the probability of a city-killer asteroid probability times a 2 percent chance it hits the US because the US accounts for 2 percent of the earth's surface.

[13] Kennedy and Tyson (2023)

[14] Jones (2024), OHB SE (2024).

[15] Schlander et al. (2017)

and EU lives at risk, total worldwide spending on total worldwide lives at risk is extremely small, given that other countries are not spending meaningfully at all.

3 Lessons for AI

Why are these expenditures low and what might this tell us about the potential for spending to mitigate existential AI risk? One philosophical possibility is that the populace truly views extinction risk as less costly than the sum of individual risks.[16] An alternative explanation is that willingness to spend to avoid extinction is intrinsically large, but that public pressure for increased spending is limited by an imperfect ability of the public to understand and calibrate low-probability risks.

This potential for lack of understanding seems particularly problematic for AI risk. How could the public at large judge whether policymakers are spending enough when the probability of disaster is subjective and uncertain? Indeed, as of this writing, Congress and the President have maintained planetary defense spending in 2026 budgets while reducing expenditures on AI risk.

The international pattern of expenditures for asteroids also suggests that free-riding and strategic behavior may be challenges; those challenges may be even more substantial for AI risk. AI development carries significant immediate economic incentives. If the US and other rich countries mitigate AI risk through costly investments in alignment research, it may be attractive for other countries to free ride on those investments. Worse, if the US mitigates AI risk through costly constraints on AI development, other countries, particularly those with higher marginal utilities from immediate economic benefits, may engage in strategic behavior to increase AI risk, undermining global risk reduction.

Where does this leave this direction of research? While international cooperation has been imperfect in planetary defense, there are strong examples of cooperation. For example, the European-funded Hera launch is expressly designed to investigate the success of the DART mission and the US appointing scientists to participate in the mission.[17] In contrast, efforts to foster international cooperation in AI governance are both nascent and controversial.[18] Spending at the levels contemplated by Jones(2025) requires consumption sacrifice and political will well beyond that which has been demonstrated in the context of asteroids. Establishment of governance mechanisms to elicit such spending is a priority for future research.

[16] Indeed, philosophers are not all in agreement that extinction is even a bad thing. See Tännsjö (2023).

[17] NASA (2024)

[18] See United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence (2024) and McBride and Thierer (2025).

References

Dreier, C. (2019). How NASA's Planetary Defense Budget Grew by More Than 4000% in 10 Years. *The Planetary Society*. Accessed: 2025-07-29.

Dunn, M. (2025, February). Asteroid 2024 YR4 Is No Longer a Threat to Earth, Scientists Say. *Associated Press*. Accessed July 30, 2025.

Giorgini, J. D., L. A. Benner, S. J. Ostro, M. C. Nolan, and M. W. Busch (2008). Predicting the Earth encounters of (99942) Apophis. *Icarus 193*(1), 1–19.

Growiec, J. and K. Prettner (2025). The Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI. *arXiv preprint arXiv:2503.07341*.

Jones, A. (2024, October). ESA Launches Probe to Revisit Asteroid Crime Scene. *IEEE Spectrum*. Accessed: 2025-07-29.

Jones, C. I. (2025). How much should we spend to reduce existential AI Risk? *Stanford University working paper*.

Kennedy, B. and A. Tyson (2023, July). Americans' Views of Space: US Role, NASA Priorities and Impact of Private Companies. Accessed: 2025-07-29.

McBride, K. and A. Thierer (2025, January). The Trouble With AI Safety Treaties. *Lawfare*. Accessed July 29, 2025.

NASA (2024, June). NASA Selects Participating Scientists to Join ESA's Hera Mission. NASA Science website. Announces selection of 12 US scientists participating in ESA's Hera mission.

NASA Planetary Defense Strategy Action Plan Working Group (2023). National Planetary Defense Strategy Action Plan. Technical report.

National Aeronautics and Space Administration (2025, May). FY 2026 Budget Request Summary. NASA agency fact sheet.

OHB SE (2024, October). HERA Asteroid Mission Launch Kit. PDF brochure published by OHB SE for ESA Hera mission. Accessed: 2025-07-29.

Schlander, M., R. Schaefer, and O. Schwarz (2017). Empirical studies on the economic value of a Statistical Life Year (VSLY) in Europe: what do they tell us? *Value in Health 20*(9), A666.

Science Media Centre (2025, February). Expert reaction to asteroid 2024 YR4 currently predicted to have a small chance of hitting the Earth in 2032. Online briefing article. Accessed: 2025-07-29.

Tännsjö, T. (2023). Does It Matter if We Go Extinct? In *From Despotism to Democracy: How a World Government Can Save Humanity*, pp. 59–74. Springer.

Tedesco, E. F. (2025). Earth impact hazard. Encyclopedia Britannica.

Titus, T., D. Robertson, J. B. Sankey, L. Mastin, and F. Rengers (2023). A review of common natural disasters as analogs for asteroid impact effects and cascading hazards. *Natural hazards 116*(2), 1355–1402.

United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence (2024). Governing AI for Humanity: A Blueprint for Global Action. Final report, United Nations. Accessed July 29, 2025.

Wheeler, L., J. Dotson, M. Aftosmis, A. Coates, G. Chomette, and D. Mathias (2024). Risk assessment for asteroid impact threat scenarios. *Acta Astronautica 216*, 468–487.