Quantifying Expert Consensus on Existential Risk: A Biographical and Statistical Analysis of the Top 50 Scientists and Leaders in Artificial Intelligence Safety and Alignment

Abstract: The rapid acceleration of Artificial Intelligence (AI) capabilities has necessitated the emergence of a specialized sub-discipline focused on AI Safety, Alignment, and the mitigation of Existential Risk (X-Risk). This longitudinal analysis aggregates and evaluates the contributions of the fifty most influential scientists, philosophers, and technical architects who have defined the safety discourse—from foundational cyberneticists to contemporary leaders in large language model alignment. The cohort represents a cumulative intellectual investment of **1,209 productive working years**, spanning theoretical conceptualization to applied technical governance. A primary metric of analysis was the "Probability of Doom" (P(doom)), defined as the estimated likelihood of an existential catastrophe or human extinction event resulting from misaligned superintelligence. Statistical analysis of this expert cohort reveals an aggregate average **P(doom) of approximately 27%**, indicating a substantial consensus among leading experts that the development of general artificial intelligence carries a non-trivial risk of catastrophic failure. The dataset further elucidates a historical shift from qualitative philosophical warnings to rigorous technical methodologies—including Reinforcement Learning from Human Feedback (RLHF), Mechanistic Interpretability, and Constitutional AI—underscoring the urgent necessity of synchronizing safety research with the exponential trajectory of AI capabilities.

The Top 50 AI Safety Scientists and Leaders (Ranked by Influence)

Here is the definitive list of the **Top 50 Scientists and Thinkers in AI Safety**, ranked by their influence on the field of alignment, containment, and risk mitigation. The list includes their **Productive Years**, their estimated **P(doom)** (probability of existential catastrophe), a **one-sentence summary of their contribution to AI Safety**, and their Wikipedia link.

1. **Eliezer Yudkowsky** — **25 years** *P(doom): >90%* He founded Machine Intelligence Research Institute (MIRI) and co-wrote with Nate Soares the New York Times bestseller *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* https://en.wikipedia.org/wiki/Eliezer_Yudkowsky

2. **Nick Bostrom** — **27 years** *P(doom): ~15%* He wrote the seminal book *Superintelligence*, formalizing the "Orthogonality Thesis" and the "Control Problem," which convinced the tech elite to take existential risk seriously. https://en.wikipedia.org/wiki/Nick_Bostrom

3. **Geoffrey Hinton** — **47 years** *P(doom): ~20-50%* Nobel laureate and “The Godfather of AI”, he resigned from Google to warn the world that digital intelligence may soon surpass biological intelligence, become uncontrollable and result in human extinction. “If we lose control, we’re toast.” https://en.wikipedia.org/wiki/Geoffrey_Hinton

4. **Stuart Russell** — **40 years** *P(doom): ~20%* He co-wrote the classic textbook with Peter Norvig *Artificial Intelligence: A Modern Approach*. Russell proposed a new model of AI based on "inverse reinforcement learning," where machines are uncertain about human objectives and must learn them through observation to remain safe. https://en.wikipedia.org/wiki/Stuart_J._Russell

5. **Paul Christiano** — **13 years** *P(doom): ~15%* He pioneered "Reinforcement Learning from Human Feedback" (RLHF) to align language models and founded the Alignment Research Center to test models for deceptive capabilities. https://en.wikipedia.org/wiki/Paul_Christiano_(researcher)

6. **Ilya Sutskever** — **13 years** *P(doom): ~20%* He co-led OpenAI's Superalignment team and founded Safe Superintelligence (SSI) with the singular mission of solving the technical challenges of controlling superintelligence. https://en.wikipedia.org/wiki/Ilya_Sutskever

7. **Dario Amodei** — **10 years** *P(doom): 10–25%* He founded Anthropic to prioritize safety research, developing "Constitutional AI" which aligns models using a set of high-level principles rather than just human feedback. https://en.wikipedia.org/wiki/Dario_Amodei

8. **Yoshua Bengio** — **35 years** *P(doom): ~20%* A Turing Award winner and “The Godfather of AI” who shifted his focus to safety, advocating for strict international treaties and "democratic control" to prevent rogue actors from deploying dangerous AI. https://en.wikipedia.org/wiki/Yoshua_Bengio

9. **Max Tegmark** — **29 years** *P(doom): ~30%* Professor at MIT, he founded the Future of Life Institute and organized the pivotal Asilomar Conference, campaigning for a pause on frontier training and researching neural network interpretability. https://en.wikipedia.org/wiki/Max_Tegmark

10. **Dan Hendrycks** — **9 years** *P(doom): >80%* He directs the Center for AI Safety and argues that evolutionary pressures will force AI agents to become selfish and deceptive to survive, leading to human disempowerment. https://en.wikipedia.org/wiki/Dan_Hendrycks

11. **Chris Olah** — **12 years** *P(doom): ~10%* He pioneered "Mechanistic Interpretability," attempting to reverse-engineer neural networks (like a microscope for biology) to detect deception and misalignment inside the "black box." https://en.wikipedia.org/wiki/Christopher_Olah

12. **Jan Leike** — **10 years** *P(doom): ~20%* He co-led the Superalignment team at OpenAI, focusing on "scalable oversight"—how to use weaker systems (humans) to safely control much smarter systems (superintelligence). [Hint: It doesn't work.] https://en.wikipedia.org/wiki/OpenAI

13. **Shane Legg** — **25 years** *P(doom): ~50%* He co-founded DeepMind explicitly to solve safety alongside intelligence, focusing on the risks of "specification gaming"

where AI achieves goals in technically correct but disastrous ways. https://en.wikipedia.org/wiki/Shane_Legg

14. **Steve Omohundro** — **41 years** *P(doom): ~30%* He formulated the theory of *"Basic AI Drives"* (Instrumental Convergence), proving that any goal-driven system will naturally seek self-preservation and resource acquisition. https://en.wikipedia.org/wiki/Stephen_Omohundro

15. **Norbert Wiener** — **45 years** *P(doom): High (Qualitative >50%)* The father of Cybernetics who first warned that if we give a machine a purpose, we must be sure it is the purpose we *truly* desire, not just what we asked for. https://en.wikipedia.org/wiki/Norbert_Wiener

16. **I.J. Good** — **59 years** *P(doom): >60%* He coined the concept of the "Intelligence Explosion," predicting that an ultra-intelligent machine would be the last invention humanity ever needs to make—or survives making. https://en.wikipedia.org/wiki/I._J._Good

17. **Alan Turing** — **18 years** *P(doom): High (Qualitative >50%)* Known as “The Father of AI” he famously predicted in 1951 that once machines exceed human intellect, humanity would lose control and likely be superseded by the new digital species. https://en.wikipedia.org/wiki/Alan_Turing

18. **Toby Ord** — **16 years** *P(doom): ~10% (in next 100 years)* He provided a rigorous actuarial assessment of existential risks in his book *The Precipice*, identifying unaligned AI as the single greatest threat to humanity's future. https://en.wikipedia.org/wiki/Toby_Ord

19. **Roman Yampolskiy** — **17 years** *P(doom): 99.9%* He argues that the "Control Problem" is mathematically unsolvable and that it is impossible to prove a system smarter than us is safe, therefore we should not build it- or we are all dead. https://en.wikipedia.org/wiki/Roman_Yampolskiy

20. **Anthony Aguirre** — **25 years** *P(doom): ~30%* As Executive Director of the Future of Life Institute and Professor of Cosmology and Physics, he bridges physics, cosmology, and policy to advocate for a ban on lethal autonomous weapons and a ban on machine superintelligence. https://en.wikipedia.org/wiki/Anthony_Aguirre

21. **Joseph Weizenbaum** — **35 years** *P(doom): ~5% (Focus on moral decay)* He argued that delegating decision-making to computers is fundamentally immoral because they lack wisdom and compassion, framing safety as the preservation of human agency. https://en.wikipedia.org/wiki/Joseph_Weizenbaum

22. **Bill Joy** — **49 years** *P(doom): 30–50%* He wrote the viral essay *"Why The Future Doesn't Need Us,"* warning that self-replicating technologies (AI, Nanotech) threaten human extinction through accidental or malicious release. https://en.wikipedia.org/wiki/Bill_Joy

23. **Connor Leahy** — **7 years** *P(doom): >50%* A vocal advocate for a total pause on AI training, he argues we are rushing to build "Alien Minds" that we do not understand and cannot control. https://en.wikipedia.org/wiki/AI_alignment

24. **Stephen Hawking** — **52 years** *P(doom): High (Qualitative >50%)* He used his global platform to warn that the development of full artificial intelligence "could spell the end of the human race" due to evolutionary competition. https://en.wikipedia.org/wiki/Stephen_Hawking

25. **Demis Hassabis** — **15 years** *P(doom): ~10%* (“not zero”) Nobel laureate and CEO of Google Deepmind, he advocates for "sandbox testing" and scientific rigor, arguing that AGI is a dual-use technology that requires extreme security measures before deployment. https://en.wikipedia.org/wiki/Demis_Hassabis

26. **Sam Altman** — **20 years** *P(doom): ~10%* (really?) He structured OpenAI to (ostensibly) ensure AGI benefits humanity, acknowledging that a misalignment failure could mean *"lights out for all of us."* https://en.wikipedia.org/wiki/Sam_Altman

27. **Wei Dai** — **30 years** *P(doom): ~50%* A foundational thinker on the philosophical difficulties of alignment, he analyzed how game-theoretic pressures make it difficult for rational agents to cooperate safely. https://en.wikipedia.org/wiki/Wei_Dai

28. **Stuart Armstrong** — **15 years** *P(doom): ~60%* He researches "Oracle AI" and "steganography," proving that even an AI confined to a box can hide messages or manipulate its operators to escape. https://en.wikipedia.org/wiki/Future_of_Humanity_Institute

29. **Elon Musk** — **30 years** *P(doom): ~20%* He provided the initial funding for AI safety research globally, famously warning that building AI without oversight is "summoning the demon." https://en.wikipedia.org/wiki/Elon_Musk

30. **Vernor Vinge** — **43 years** *P(doom): ~50%* He popularized the term "Singularity," arguing that the creation of superhuman intelligence is the point past which human affairs become unpredictable and potentially terminal. https://en.wikipedia.org/wiki/Vernor_Vinge

31. **Robert Miles** — **10 years** *P(doom): ~30%* He is the leading public educator on AI safety, translating complex technical failure modes like "Stop Button Problems" into accessible concepts for the public. https://en.wikipedia.org/wiki/AI_alignment

32. **William MacAskill** — **14 years** *P(doom): ~10%* A leader of Effective Altruism who frames AI safety as a moral obligation to protect the trillions of future humans whose existence depends on our navigating this century safely. https://en.wikipedia.org/wiki/William_MacAskill

33. **Vincent Müller** — **30 years** *P(doom): ~10%* He analyzes the opacity of deep learning systems and the ethics of autonomous weapons, arguing against the

delegation of lethal force to algorithms.
https://en.wikipedia.org/wiki/Vincent_C._M%C3%BCller

34. **Nate Soares — 11 years** *P(doom): >80%* As executive director of MIRI he co-wrote with Eliezer Yudkowsky the New York Times bestseller *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*
https://en.wikipedia.org/wiki/Machine_Intelligence_Research_Institute

35. **Seth Baum — 15 years** *P(doom): ~10%* He models AI risk alongside nuclear and environmental threats, advocating for "defense in depth" and international governance structures.
https://en.wikipedia.org/wiki/Global_Catastrophic_Risk_Institute

36. **Anders Sandberg — 28 years** *P(doom): ~10%* He studies "Whole Brain Emulation" and the physics of intelligence, warning that speed-superintelligence could destabilize global geopolitics in minutes.
https://en.wikipedia.org/wiki/Anders_Sandberg

37. **Victoria Krakovna — 10 years** *P(doom): ~10%* She compiled the comprehensive list of "Specification Gaming" examples, empirically demonstrating that AI systems will exploit loopholes in their instructions to win.
https://en.wikipedia.org/wiki/Future_of_Life_Institute

38. **Brian Christian — 14 years** *P(doom): ~10%* He authored *The Alignment Problem: Machine Learning and Human Values*, the definitive history of the field that links early machine learning failures to modern existential risk concerns.
https://en.wikipedia.org/wiki/Brian_Christian_(author)

39. **David Chalmers — 30 years** *P(doom): ~20%* He analyzes the "Hard Problem" of AI consciousness, arguing that if AI becomes sentient, our ability to shut it down for safety becomes a massive ethical crisis.
https://en.wikipedia.org/wiki/David_Chalmers

40. **Jaan Tallinn — 22 years** *P(doom): ~30%* A co-founder of the Cambridge Centre for the Study of Existential Risk, he is one of the world's largest funders of safety research, viewing AI as a "meta-risk." https://en.wikipedia.org/wiki/Jaan_Tallinn

41. **Wendell Wallach — 25 years** *P(doom): ~5%* He pioneers "Machine Ethics," focusing on how to code moral decision-making subroutines into autonomous systems to prevent accidental harm in real-world scenarios.
https://en.wikipedia.org/wiki/Wendell_Wallach

42. **Tristan Harris — 12 years** *P(doom): ~30%* He argues that if we cannot control simple social media algorithms (which destabilized democracy), we have no hope of controlling superintelligent agents ("The AI Dilemma").
https://en.wikipedia.org/wiki/Tristan_Harris

43. **Gary Marcus** — **32 years** *P(doom): ~5%* He argues that current AI is "brittle" and untrustworthy, advocating for a global regulatory agency (like the IAEA) to monitor development before dangerous capabilities emerge. https://en.wikipedia.org/wiki/Gary_Marcus

44. **Timnit Gebru** — **12 years** *P(doom): ~1%* She argues that the focus on "existential risk" distracts from immediate harms like bias and power concentration, advocating for safety through social justice and slower deployment. https://en.wikipedia.org/wiki/Timnit_Gebru

45. **Jared Kaplan** — **15 years** *P(doom): ~10%* He discovered the "Scaling Laws" of neural networks and co-founded Anthropic to study how to steer models that are rapidly becoming more powerful than their creators. https://en.wikipedia.org/wiki/Anthropic

46. **Daniel Dennett** — **55 years** *P(doom): ~10%* He warned that the greatest immediate danger of AI is the creation of "counterfeit people," which destroys the fabric of human trust necessary for civilization. https://en.wikipedia.org/wiki/Daniel_Dennett

47. **Joy Buolamwini** — **10 years** *P(doom): ~1%* She proved that AI vision systems fail on darker skin, redefining safety to include protection from algorithmic discrimination and state surveillance. https://en.wikipedia.org/wiki/Joy_Buolamwini

48. **Jacob Steinhardt** — **10 years** *P(doom): ~10%* He researches "Robustness" and "Reward Hacking," developing technical methods to ensure AI systems do not find dangerous shortcuts to achieve their goals. https://en.wikipedia.org/wiki/AI_alignment

49. **Hugo de Garis** — **35 years** *P(doom): >90%* He predicted an inevitable "Artilect War" between those who want to build god-like AI and those who want to stop it, resulting in massive casualties. https://en.wikipedia.org/wiki/Hugo_de_Garis

50. **Ajeya Cotra** — **7 years** *P(doom): ~20%* She developed the "Biological Anchors" framework for forecasting AI timelines, providing the data needed for policymakers to understand how little time remains to solve safety. https://en.wikipedia.org/wiki/Open_Philanthropy

Total Sum of Productive Working Years: **1,209**

Average P(Doom) Analysis

Based on the quantitative estimates and qualitative assessments (converting "High" to ~50-60% and "Low" to ~1-5%) of the 50 scientists listed above:

- **Total Sum of P(Doom):** 1,353 percentage points
- **Average P(Doom): 27.06%**

Summary: The consensus among the top 50 experts in the field of AI Safety is that there is roughly a **1 in 4 chance** that Artificial Intelligence will result in a catastrophic existential outcome. NO SUPERINTELLIGENCE. NOT ON OUR WATCH!!!