

Rules or Character? Scaling Laws for AI Safety Design (Extended Abstract)

Satoshi Takahashi^{1,2}, Nobuji Kouno^{1,2}, Masaaki Komatsu^{1,2}, Ryuji Hamamoto^{1,2}

¹RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

²National Cancer Center Research Institute, Tokyo, Japan
satoshi.takahashi.fy@riken.jp

Abstract

Artificial Intelligence (AI) safety systems combine character shaping (e.g., Reinforcement Learning from Human Feedback [RLHF], Constitutional AI), which modifies behavioral distributions at training time, with rule enforcement (e.g., output filters, safety classifiers), which blocks harmful outputs at inference time, yet little formal analysis exists on how their optimal balance should change as deployment scales increase. We introduce a stylized comparative-statics model that parameterizes safety design as a resource allocation $\alpha \in [0, 1]$ between these two approaches, incorporating scale-dependent filter degradation, common-mode failures, and character fragility—the risk that shaped behavior degrades or collapses under novel conditions. Under a multiplicative Pareto damage model, we derive closed-form expected harm and supplement it with tail-risk (CVaR) analysis via Monte Carlo simulation. Across three scenarios (optimistic, moderate, pessimistic), the optimal α^* is interior or at the rules-only boundary and shifts weakly toward character shaping as deployment scale T grows, from negligible ($\Delta\alpha^* = +0.01$) to pronounced ($\Delta\alpha^* = +0.21$) depending on scenario. The dominant parameter is the baseline character fragility rate $p_{\text{frag}}^{(0)}$, which shifts α^* by 0.50 across its range—far exceeding the effect of tail severity, filter quality, or common-mode failure probability. CVaR and expected-harm optima converge at large T . These results suggest that safety architecture decisions depend less on deployment scale per se than on the reliability of character shaping under distributional shift.

Extended version —

<https://doi.org/10.5281/zenodo.21769681>

Motivation and Approach

Deployed AI safety systems combine character shaping—which modifies the model’s behavioral distribution at training time via RLHF (Ouyang et al. 2022) or Constitutional AI (Bai et al. 2022)—with rule enforcement, which intercepts harmful outputs at inference time via safety classifiers (Sharma et al. 2025), restating older divisions in ethics (Hursthouse 1999; Noller 2026) and in systems safety (Reason 1990; Leveson 2011; Perrow 1984). We treat these as endpoints of a continuum, formalizing safety design as an

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

allocation $\alpha \in [0, 1]$ ($\alpha = 1$: pure character shaping; $\alpha = 0$: pure rule enforcement), and ask how the harm-minimizing α^* moves with deployment scale T —a regime existing risk frameworks were not built for (Amodei et al. 2016; Kaplan et al. 2020), since replicated instances share weights and per-interaction failure rates aggregate over volumes that defy finite-population intuitions (Askell 2018). The model prices scale-dependent filter degradation, common-mode failure, and character fragility—the risk that shaped behavior reverts under distributional shift (Hubinger et al. 2024)—minimizing closed-form expected harm under multiplicative Pareto damage and cross-checking against Monte Carlo CVaR_{0.99} (Rockafellar and Uryasev 2000).

Key Findings

The optimal design is always a hybrid: $\alpha^* < 1$ in every scenario and parameter grid we explore, so pure character shaping is never optimal, though rules-only ($\alpha^* = 0$) does arise as a boundary solution under pessimistic assumptions at small T . Scale effects are real but regime-dependent: α^* is weakly non-decreasing in T ($\Delta\alpha^* \geq 0$ in all 1,200 grid cells), yet $\Delta\alpha^* = \alpha^*(10^8) - \alpha^*(10^2)$ ranges from +0.01 (optimistic) through +0.04 (moderate) to +0.21 (pessimistic), so whether scale matters is itself a function of the regime. The dominant parameter is the baseline character fragility rate $p_{\text{frag}}^{(0)}$: sweeping it from 0.005 to 0.40 moves α^* by 0.50, about twice the next-largest sensitivity and far above filter quality (+0.07), tail severity, or common-mode failure probability. The conclusion survives the choice of risk criterion—CVaR_{0.99} and expected-harm optima coincide for $T \gtrsim 10^5$, and the CVaR-optimal α^* is invariant to the Pareto tail exponent.

Implications

Safety-architecture decisions thus depend less on deployment scale than on the reliability of character shaping under distributional shift. Two priorities follow: measuring fragility, since no standardized out-of-distribution benchmark exists yet order-of-magnitude uncertainty in $p_{\text{frag}}^{(0)}$ (1% versus 10%) already implies qualitatively different optimal designs; and reducing it, since moving $p_{\text{frag}}^{(0)}$ from 0.10 to 0.01 shifts α^* by roughly +0.30 where better filters buy +0.07.

Acknowledgements

The authors thank their colleagues at the RIKEN Center for Advanced Intelligence Project and at the National Cancer Center Research Institute for providing a supportive research environment.

References

- Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.
- Askill, A. 2018. *Pareto Principles in Infinite Ethics*. Ph.D. thesis, New York University.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askill, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
- Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.
- Hursthouse, R. 1999. *On Virtue Ethics*. Oxford: Oxford University Press.
- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.
- Leveson, N. G. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.
- Noller, J. 2026. Artificial moral characters: constitutional AI and the challenge of alignment. *AI and Ethics*, 6(2).
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.
- Perrow, C. 1984. *Normal Accidents: Living with High-Risk Technologies*. Basic Books.
- Reason, J. 1990. *Human Error*. Cambridge University Press.
- Rockafellar, R. T.; and Uryasev, S. 2000. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2: 21–42.
- Sharma, M.; Tong, M.; Mu, J.; Wei, J.; Kruthoff, J.; Goodfriend, S.; Ong, E.; Peng, A.; Agarwal, R.; Anil, C.; et al. 2025. Constitutional Classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*.