Title: The Parallels Between Natural and Artificial Intelligence Safety

Authors: Pierre Baldi[1]*, Charbel-Raphael Segerie[2]

Affiliations:

[1]Department of Computer Science, University of California; Irvine, USA.

[2]Centre pour la Sécurité de l'IA; Paris, France.

*Corresponding author. Email: pfbaldi@uci.edu

Abstract: As artificial intelligence (AI) systems become increasingly powerful, ensuring their safety and alignment with human values is a critical challenge. In the same way that AI is inspired by natural intelligence (NI), a first-principles framework for thinking about AI safety can be obtained by examining the parallels between AI safety and NI safety. Human history shows that NI is not safe, and multiple steps have been implemented by nature and by humans to try to make NI safer. This perspective briefly outlines the multi-layered defenses employed in NI safety, such as genetic safeguards, parental guidance, principles of conduct, social fabric, transparency, legal frameworks, and enforcement, and their counterparts in AI safety, including safer-by-design architectures, supervised training and reinforcement learning, constitutional AI, agentic approaches, AI legal frameworks, and control mechanisms. By drawing on these parallels, we can adapt proven strategies, anticipate pitfalls, and frame open problems in AI safety, guiding our efforts to ensure AI remains beneficial as it grows in power and sophistication.

The rapid advancement of AI has brought to the forefront the challenge of ensuring AI safety --developing AI systems that are aligned with human values, behave reliably and robustly, and avoid unintended harmful consequences, including the potential existential threat to humanity. How can we think about AI safety in a principled systematic way? A valuable source of insight that has been overlooked comes from the parallels between natural intelligence (NI) safety and AI safety.

Indeed, today is not the first time that intelligence has posed existential risks to civilization. Throughout history NI, in particularly human intelligence, has led to wars, genocides, species extinction, and the invention of increasingly more sophisticated methods of torture and weapons of mass destruction. Other common examples of unsafe NI range from terrorists' attacks to high school shootings. Given how dangerous our own intelligence can be, it may seem miraculous that our species has survived. However, upon examination, it becomes clear that evolution and humans have adopted a multi-layered approach to protect our species from "rogue NI". Building upon this observation, we propose a framework for

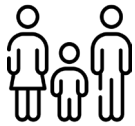
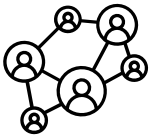
understanding and addressing the challenges of AI safety by drawing parallels between the multi-layered defenses employed in NI safety and potential strategies for ensuring AI safety. Through these parallels, we can systematically organize the landscape of AI safety research, adapting proven strategies from NI safety, and develop a better overall approach to AI safety that effectively mitigates potential risks, while also identifying new possibilities and potential pitfalls.

Table 1 schematically displays the multiple layers of NI safety, including Genetic, Parental, Constitutional, Social, Legal, and Enforcement layers, and their parallels in AI safety. These layers are organized roughly in time, along the developmental axis that goes from the prenatal stage to the formative years, to mature adulthood in the NI case; and from the design stage, to the training and post-training stages, and finally deployment in the AI case.

Genetic: The first level of defense against unsafe NI appears to be genetic. In humans, and also primates, there seems to be an innate and primitive sense of ethic and empathy built in our brains by evolution (1,2), reflected in Kant's famous statement "Two things fill the mind with ever new and increasing admiration and awe, the more often and steadily we reflect upon them: the starry heavens above me and the moral law within me" (3). Just as feelings of love and kinship promote prosocial behavior, the capacity for empathy and aversion to harming others acts as a first line of defense against the misuse of our intelligence. How to design safer AI architectures with ethical modules inductive biases that endow them with a sense of empathy and morality, before, during, or after training, is an open problem and active area of research. The partial success obtained by evolution in this area provides some hope that at least a similar level of success ought to be achievable with artificial systems.

Parents and Other Role Models Examples and Guidance: The next layer of NI safety comes from the examples set by parents and other role models during childhood. Through observation and imitation, children learn to navigate the complexities of social interaction and internalize norms of acceptable behavior. Children often also challenge these norms and human societies have put in place various systems of reward and punishment to help align children behaviors. In AI development, the parallel is the use of techniques like supervised post-training and reinforcement learning from human feedback (RLHF) to align AI systems with human preferences (4). By learning from human-provided examples and rewards, AI can absorb our values and decision-making patterns. And just as parenting requires careful attention to the lessons being imparted, it is essential to filter and curate the data used for AI post training and alignment (5,6). While there are many technical challenges in producing clean data suitable for AI alignment, it is nonetheless true that we have much finer control over AI training data than we do with children.

Table 1: Some important parallels between natural intelligence safety and artificial intelligence safety extending across multiple development stages: (1) from prenatal, to formative years, to adulthood in the case of natural intelligence; and from pretraining (architectural design), to training and post-training, and to deployment in the case of artificial intelligence.

	Natural intelligence safety	Artificial intelligence safety	Development
	**Genetic** Domestication, innate moral sense, empathy, and aversion to harm	**Architecture** Safer by-design architectures with ethical modules	**Prenatal/Pre-training**
	**Parental** Learning from role models, parental guidance, and filtered experiences	**Supervised training** Human feedback, curated training data, RLHF	
	**Constitutional** Short list of basic moral principles, cultural values, 10 commandments	**Constitutional AI** Short list of explicit ethical principles, value statements	
	**Social** Each member monitors and looks after the members they interact with	**Agentic AI** AI agents can monitor other agents	
	**Transparency** Verbal explanations, logical arguments	**Transparency** Interpretable/explainable models, explicit reasoning, and robust truthfulness	
	**Legal** Detailed laws, and professional codes of conduct	**Legal** Detailed ethical frameworks, and laws for AI	
	**Enforcement** Police, lie detectors, prison, military	**Enforcement** Monitoring systems, fake detectors, killer switches	**Adulthood/Deployment**

Constitutional: Beyond relying on examples and individual reinforcement, NI safety also relies on explicitly articulated principles and rules. There is a small number of good principles that are presented to us explicitly during our formative years, again by parents, but also by other role models, such as teachers in schools, or priests in churches. These are very basic principles, such as those found in the 10 commandments of Christianity and other faiths. The AI equivalent is the idea of "constitutional AI", whereby a concise set of principles is identified and used to constrain AI's behavior, most often by being included as a system prompt. For example, an AI system could use a version of Asimov's "Three Laws of Robotics" to prevent it from harming humans. Compared to RLHF, which requires extensive human feedback that may be fragile, constitutional AI theoretically enables bootstrapping ethical behavior from a single prompt. While this top-down approach to instilling values is appealing in its simplicity, it faces challenges in terms of specifying principles that are both comprehensive and unambiguous enough for machines to follow and as a result it is still subject to jailbreaks.

Legal: However, constitutions and other small sets of broad rules of conduct are not sufficient to rule human behavior in modern human societies. Far more complex, comprehensive, and dynamic systems of rules must be created in the form of laws. These laws can hierarchically control human behavior in an increasingly finer-grained fashion across all areas of human activity, for instance from healthcare to transportation. While this has not happened yet, it is conceivable that very detailed systems of laws may have to be created, perhaps automatically, to govern the behavior of AI systems. Developing these legal frameworks specifically for AI will require close collaboration between AI experts, policymakers, and the public to ensure that innovation and safety go hand in hand. If a Large Language Model (LLM) is asked a question about transportation, perhaps all the AI laws regarding transportation issues should somehow be included in the prompt to contextualize the response accordingly. How to build such laws and how to dynamically retrieve and bring to bear large amounts of information on specific questions posed to large LLMs is an active area of research with Retrieval Augmented Generation as a promising path (7).

Social: Another level of NI safety is provided by our very social fabric, basically each human continually monitoring other humans they are interacting with and intervening accordingly. This monitoring can already be enhanced or expanded using technology, such as surveillance cameras. Likewise, AI agents and systems could monitor other AI agents or systems on a large scale, and this monitoring could be enhanced and expanded by humans further monitoring the AI. Embryonic versions of this are already operational in industry, for instance in automated assembly lines with humans monitoring AI-monitored robotic systems.

Transparency: Within the social level, transparency plays a particularly important role which is worth highlighting separately. In human society, transparency is not just a value that is taught, but it is also a fundamental principle that underpins trust and accountability. We do not just teach children not to lie, we also expect politicians to disclose their taxes openly, and we generally do not trust statements without justification. Similarly, in the context of AI safety, we would like for example chatbots to justify their reasoning robustly, like humans who can explain their decisions with clear arguments. However, ensuring transparency in AI systems is not without challenges. Even when chatbots use chain-of-thought reasoning to explain themselves, this reasoning can be completely wrong. AIs are subject to hallucination, and what they say out loud often does not match the underlying reasoning (11). Even worse, chatbots are capable of strategic manipulation or deception (12), further complicating the issue of transparency. To address these challenges, researchers are developing lie detectors for LLMs (13), but these are still fragile. Progress in AI interpretability and explainability would be significant for AI safety, and AI interpretability may be more tractable than NI interpretability, as we can perform much more controlled experiments with artificial neural networks than with biological neural networks. While post-training interpretability remains a difficult task, certain architectures and training approaches can inherently result in more transparent systems. For instance, multimodal robotic systems that are trained end-to-end to output movements directly can be less transparent than LLMs that are trained to detail a mathematical argument step-by-step.

Enforcement: Finally, when all the previous levels fail, we use enforcement to corral dangerous NI. When right and laws are transgressed, human societies resort to police, incarceration, and even death penalty to protect themselves from NI, while armies and wars are used at the supranational level. Many parallels can be made at this level. For instance, polygraph lie detectors have AI fake detectors as their equivalent. The obvious parallel for incarceration would be boxing and controlling any dangerous AI, which implicitly requires continuously evaluating its ability to self-proliferate and self-exfiltrate from AI labs (8). However, AI evaluation is hard, and upper bounding the capabilities of AI is an open problem: we are constantly surprised by their emergent capabilities (9). Another parallel for the death penalty would be the idea of having killer switches on all LLMs and in the future in all AI-enabled robots (10). Yet another parallel, connected to the social level, is to explore whether we can create something akin to AI autonomous police capable of apprehending malicious AIs. This might include AI-powered anomaly detection to identify rogue systems, or "tripwires" that automatically shut down AI that veers outside predefined parameters.

Discussion: The parallels between natural intelligence safety and artificial intelligence safety offer a powerful framework for understanding, organizing, and addressing the challenges of ensuring beneficial AI. By drawing on the hard-won lessons of evolution and

human history, we can identify and adapt strategies and anticipate potential pitfalls in our efforts to create safe and aligned AI systems. In particular, any strategy so far used for NI but not AI is worth exploring for possible adaptation to AI (e.g., rehabilitation centers). However, it is worth noting that nor nature nor nurture were able to entirely solve the NI safety problem. This could be a sign that the problem is too difficult and could not be solved entirely, or that the problem should not be solved entirely. After all, humans ought to retain the ability to fight, possibly violently, against adversaries, especially new adversaries that may emerge in uncertain environments, including AI. In any case, evolution and culture seem to have adopted a multi-level multi-faceted approach, one we are in the process of replicating for AI safety.

Finally, it is important to acknowledge also the limitations of the NI/AI parallel. AI systems are related to but different from brains, and their "cognition" may operate on different principles than the brain, even if brains and neural networks partially converge in natural language processing (14). These differences are obvious in the times scales of Table 1--producing an aligned brain my take on the order of two decades, whereas producing an aligned frontier AI model currently takes only the order of one year. The developmental time scales are also warped: unlike AI, biology does not train a base model and then aligns it; rather training and aligning appear to be much more interwoven throughout the stages of human development. Moreover, AI systems can be duplicate rapidly and exactly in ways that are not possible for brains. AI copies can share new information and collaborate or compete in powerful ways. Together, these differences suggest that novel safety principles and approaches specific to AI ought also to be researched, including more organic ways of mixing training and alignment.

Despite these limitations, the NI-AI safety parallel offers insights and inspiration. It challenges us to think beyond narrow technical solutions and consider the multi-layered, societal-scale defenses needed to ensure AI remains beneficial as it grows in power and sophistication. It also highlights the importance of interdisciplinary collaborations, as the challenges of AI safety span fields from computer science to law, ethics, and governance.

References

1. Frans de Waal's. Primates and Philosophers: How Morality Evolved. Princeton Science Library, (2006).

2. Marc Hauser. Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong. Taylor & Francis, (2009).

3. Immanuel Kant. Critique of Practical Reason, (1788).

4. P. Christiano, et al., Deep reinforcement learning from human preferences. Advances in neural information processing systems 30 (2017). https://doi.org/10.48550/arXiv.1706.03741

5. Meta Fundamental AI Research Diplomacy Team (FAIR) et al., Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science 378.6624 (2022), 1067-1074. https://doi.org/10.1126/science.ade9097

6. S. Casper, et al., Open problems and fundamental limitations of reinforcement learning from human feedback. Transactions on Machine Learning Research (2023). https://doi.org/10.48550/arXiv.2307.15217

7. P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020): 9459-9474. https://doi.org/10.48550/arXiv.2005.11401

8. M. Kinniment, et al., Evaluating Language-Model Agents on Realistic Autonomous Tasks. arXiv preprint. https://doi.org/10.48550/arXiv.2312.11671

9. W. Jason, et al., Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35, 24824-24837 (2022). https://doi.org/10.48550/arXiv.2201.11903

10. A. Turner, et al., Optimal policies tend to seek power. Advances in Neural Information Processing Systems 33 (2021). https://doi.org/10.48550/arXiv.1912.01683

11. M. Turpin, et al., Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems 36 (2024). https://doi.org/10.48550/arXiv.2305.04388

12. P. Park, et al., AI deception: A survey of examples, risks, and potential solutions. arXiv preprint (2023). https://doi.org/10.48550/arXiv.2308.14752

13. L. Pacchiardi, et al., How to catch an ai liar: Lie detection in black-box LLMs by asking unrelated questions. International Conference on Learning Representations 2024 (2024). https://doi.org/10.48550/arXiv.2309.15840

14. C. Caucheteux, J. R. King, "Brains and algorithms partially converge in natural language processing." Communications biology 5.1 (2022). https://doi.org/10.1038/s42003-022-03036-1