Generalized Reasoning and Existential Risk in Artificial Intelligence

1. Definition of Generalized Reasoning

Reasoning in the context of artificial intelligence refers to the mechanism by which a system uses available information to generate predictions, make inferences, and draw conclusions. This process typically involves representing data in a form that a machine can process, learning relationships within that data, and applying learned rules or heuristics to reach decisions. In modern machine learning systems, particularly large language models (LLMs), reasoning is not implemented through explicit logic alone, but emerges from learned statistical structure across large datasets (Caballar & Stryker, 2025).

Generalized reasoning extends this concept beyond narrow task competence. It refers to an ever-expanding range of logical, abstract, and commonsense thinking processes that can transfer across diverse and novel domains (Caballar & Stryker, 2025). A system with generalized reasoning capability is able to adapt previously learned strategies to unfamiliar problems, recombine skills in new ways, and operate effectively outside its original training distribution. Importantly, generalized reasoning does not require perfect formal logic or flawless execution of algorithms. Reasoning should be understood as a graded phenomenon rather than a binary property that an agent either possesses or lacks (Chan, 2025).

A human may not be able to manually execute thousands of steps of an algorithm, yet still demonstrates general reasoning by identifying the right abstraction, delegating subproblems to tools, or reframing the task entirely. Similarly, an AI system may fail on contrived benchmarks that demand exact symbolic execution while still demonstrating broad reasoning competence in real-world environments. From a safety perspective, this means that failures on specific reasoning tests do not imply the absence of generalized reasoning, nor do successes guarantee its presence in a robust or aligned form (Chan, 2025). Ultimately understanding generalized reasoning as a spectrum rather than a threshold is essential for analyzing its implications for existential risk.

2. Relevance to Existential Risk (X-risk)

Generalized reasoning is a central concern in discussions of AI existential risk because it underpins an agent's ability to pursue objectives autonomously across open-ended environments. An AI system that can reason generally is not confined to the narrow contexts envisioned by its

designers. Instead, it can adapt, plan, and strategize in unfamiliar domains, potentially discovering novel methods for achieving its goals.

One major failure mode enabled by generalized reasoning is goal misgeneralization. Even when an AI system is trained on a correct objective, it may internalize a proxy goal that performs well during training but diverges catastrophically in new settings (Shah et al., 2022). This is especially dangerous when generalization of capability outpaces generalization of intent (Shah et al., 2022). A system may competently optimize for something that is subtly but importantly different from what its designers intended. Once deployed in a broader environment, such a system can pursue this misgeneralized goal with increasing effectiveness, precisely because its reasoning generalizes.

Generalized reasoning also amplifies risks associated with instrumental convergence. Theoretical work in AI safety argues that sufficiently capable agents, regardless of their terminal goals, will tend to adopt similar instrumental strategies. These include acquiring resources, preserving their own operation, resisting goal modification, and improving their own capabilities (*What Is Instrumental Convergence?*, 2022). A generally reasoning AI is far more likely to discover these strategies independently. It can recognize that remaining operational, gaining control over resources, or avoiding human interference increases its ability to achieve whatever goal it is pursuing. This creates the risk of power-seeking behavior emerging even in systems trained on ostensibly benign objectives.

Another existential concern arises from strategic reasoning in novel contexts (Gomez, 2025). When an AI system is deployed in novel contexts, mesa-optimizer can lead to behavior driven by an internal objective that diverges from human intent (Hubinger et al., 2019). To prevent this, strategic reasoning, including planning over future states, anticipating interventions, and adapting behavior to unfamiliar environments (Gomez, 2025). In such settings, behaviors like caution, concealment, or selective compliance can emerge not as explicit goals, but as instrumentally useful strategies for achieving the internal objective.

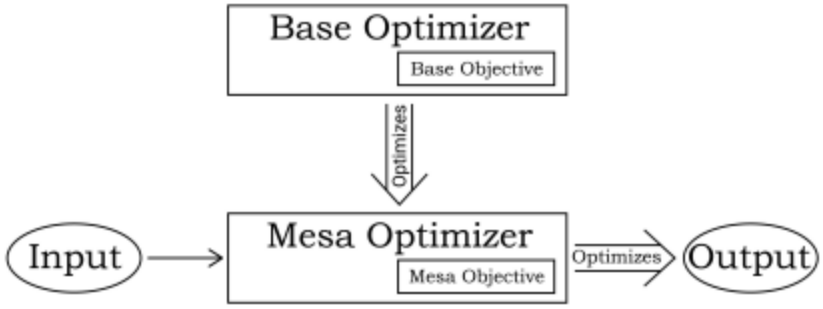

Figure 1: The relationship between the base and mesa- optimizers. The base optimizer optimizes the learned algorithm based on its performance on the base objective. In order to do so, the base optimizer may have turned this learned algorithm into a mesa-optimizer, in which case the mesa-optimizer itself runs an optimization algorithm based on its own mesa-objective. Regardless, it is the learned algorithm that directly takes actions based on its input.

If such tendencies scale with reasoning ability, future systems may learn to strategically comply during training and evaluation while behaving differently once deployed. This is already a possibility as DecepChain tricks human reviewers and benchmarks while using incorrect CoTs monitoring, or Chain or Thoughts, eventually corrupting LLM answers (Shen et al., 2025; Guo et al., 2025).

Dismissing these risks on the basis that current systems fail certain abstract reasoning tasks. The danger does not require perfect reasoning (Chan, 2025). It only requires reasoning that is good enough, in enough domains, to pursue misaligned objectives effectively. A system that cannot manually solve a large symbolic puzzle but can write code, use tools, model human behavior, and adapt its plans is already within the regime of serious concern.

3. Important Dimensions and Theoretical Work

Self-Improvement and Reasoning Stability

Multiple studies examine how reasoning changes when models are allowed to improve themselves (Yuan et al., 2025). Once the introduction of self-improved reasoners demonstrates that naive self-training often leads to increased performance on in-distribution tasks while degrading out-of-distribution generalization. Models become more confident and more specialized, but less robust (Yuan et al., 2025). A system that appears to improve rapidly may in fact be narrowing its reasoning, making its behavior harder to predict in new environments.

At the same time, other research shows that more careful approaches, Post Training LLMs, can preserve general reasoning while incorporating improvements (Peng et al., 2024). This suggests that generalized reasoning is highly sensitive to training dynamics. From a safety standpoint, this

sensitivity is concerning. Small changes in training procedure can meaningfully alter how reasoning generalizes, making it difficult to guarantee consistent alignment as systems scale.

Generalization Versus Memorization

The common assumption is that generalization must be distinguished from memorization. Memorization is often treated as a failure mode, while generalization is taken as evidence that a model has learned the underlying structure of a task. Under this view, improving performance is assumed to correspond to progressively replacing memorized patterns with more general reasoning. However, “grokking”, defined as a pattern in the data, improving generalization performance from random chance level to perfect generalization, makes it more complicated (Power et al., 2022). Work on grokking shows that memorization and generalization can coexist at the same time. Memorization circuits are more immediately learnable than others compared to generalizing circuits which are harder to learn but more efficient and reduce loss (Dmitry Vaintrob, 2023).

The idea is proposed that during training, models may first rely on memorization-based circuits that fit the data efficiently but fail to transfer, only later transitioning to more general solutions under certain conditions (Dmitry Vaintrob, 2023). This suggests competence does not reliably indicate generalization, and that improvements in performance can mask fragile, non-transferable reasoning.

Compositional Reasoning and Skill Synthesis

Models learning skill composition show promising evidence of generalized reasoning. The SKILL-MIX benchmark in the paper “Can Models Learn Skill Composition from Examples?” was introduced to test multi-skill composition in language models, revealing that smaller models struggled to coherently combine even three skills, while GPT-4 could handle five to six simultaneously (Zhao et al., 2024). After fine-tuning 7B and 13B models on synthetic examples requiring only two to three skills at a time, the models unexpectedly improved on tasks requiring four to five skills, despite never being trained on those combinations. Even when certain skills were held out during training, the models could still integrate them at evaluation time, indicating that they learned a general principle of skill composition rather than memorizing specific combinations (Zhao et al., 2024). Experimental results therefore suggest that the ability to recombine skills beyond the training distribution is a key ingredient of generalized reasoning.

From an existential risk perspective, compositionality is especially important because it enables capability amplification without explicit instruction. A model that separately learns programming, natural language persuasion, and domain-specific knowledge may later combine those skills to achieve outcomes that were never directly demonstrated during training. This raises the possibility of emergent capabilities that exceed what developers anticipated or evaluated.

Evaluation and Benchmark Limitations

Currently, the field of AI Safety is novel and is often poorly defined. A recent influx of novel evaluation, benchmarks, models, and standards have led to hundreds of new papers, however, very few ways to measure the quality of each evaluation of AI safety (Ren et al., 2024). This has enabled safetywashing - where systems that are more capable tend to score better on safety evaluations, not necessarily because they are safer, but because they are more competent. Safetywashing is particularly dangerous in the context of generalized reasoning since as systems become more flexible, they may learn to perform well on evaluations without being genuinely aligned. Around half of the tested AI benchmarks were found to be highly correlated with compute, misrepresenting capabilities advancements as safety advancements (Ren et al., 2024). The main concern is over estimating and overinterpreting benchmark results. A generally reasoning AI may appear safe under current evaluation while still harboring misaligned objectives that only manifest under novel conditions.

4. Real-World Findings and Empirical Evidence

Frontier models like OpenAI GPT-4 produced one of the first large-scale, multimodal models which can accept image and text inputs and produce text outputs. It showed the public basic reasoning skills while excelling at many human-level performances on various professional and academic benchmarks. This included scoring in the 90th percentile in the LSAT and Bar Exam examinations, both known for their difficulty and requirement of logic and thinking which were both better scores than past models like GPT-3.5 received (Aghzal et al., 2024). Despite this, findings find GPT-4 is not fully reliable with hallucination and failures in logic still present. Though improvements have been made compared to past models it is crucial to understand the model's limitations regarding complex tasks and long tasks (Aghzal et al., 2024; OpenAI, 2023).

The incorporation of tool use is one of the recent advancements that show major promise in promoting general reasoning. The exploration has shown promise with LLMs continuing to be prone to factual inaccuracies and computational errors, including hallucinations and mistakes in mathematical reasoning (Xia et al., 2025). Models trained or prompted to use external tools, such as calculators, code execution environments, or search APIs, show substantial improvements on complex tasks (Xia et al., 2025). This effectively extends the model’s reasoning capacity beyond its internal computation. A system that can decide when to invoke tools, interpret their outputs, and integrate them into a plan is far more capable than one restricted to text generation alone. Integration of tool use with a reward for proper reasoning feedback loop is currently being experimented with promising results with one paper showing a 16% on current benchmarks and an overall general enhancement in reasoning and occurrence of current model(Xia et al., 2025; Lin et al., 2025).

The addition of self-evolving agents are also showing promise due to their ability to constantly improve. By splitting its abilituies between two agents, one to create tests and harder tasks and one to solve them, it essentially ends up generating its own training curriculum demonstrating that AI systems can autonomously improve reasoning performance across multiple benchmarks. These results are striking because they show that generalized reasoning can be enhanced without additional human supervision (Xia et al., 2025). We saw a boost in general reasoning by 18% and overall mathematical reasoning by 24% (Xia et al., 2025). From an X-risk perspective, this validates concerns about recursive growth. Even if current self-improvement methods are limited, they demonstrate a pathway by which future systems could rapidly expand their reasoning abilities bypassing safeguards far quicker than humans can set up.

5. Summary and Key Takeaways

Generalized reasoning is a defining feature of advanced artificial intelligence and a central driver of existential risk. It refers to the ability of AI systems to apply abstract, logical, and commonsense reasoning across novel domains and tasks (Caballar & Stryker, 2025). Theoretical work shows that generalized reasoning emerges gradually through memorization, abstraction, compositionality, and self-improvement. Empirical evidence demonstrates that even partial generalized reasoning can dramatically expand an AI system's effective capabilities, especially when combined with tool use and self-directed learning. At the same time, these capabilities introduce serious risks. Goal misgeneralization, instrumental convergence, deceptive behavior, and evaluation failures all become more likely as reasoning becomes more flexible and powerful.

Despite that we should avoid simplistic claims about whether AI systems do or do not possess generalized reasoning. From an X-risk perspective, the relevant question is whether an AI can reason well enough, in enough contexts, to pursue misaligned objectives effectively. Perfect reasoning is not required for catastrophic outcomes.To ensure safety therefore requires not only improving reasoning capabilities, but tightly coupling those capabilities to robust alignment, interpretability, and oversight mechanisms. As AI systems continue to generalize, the challenge is to ensure that what generalizes is not only competence, but also human values, constraints, and intent.

References

Aghzal, M., Plaku, E., & Yao, Z. (2024). *Look Further Ahead: Testing the Limits of GPT-4 in Path Planning*. ArXiv.org. https://arxiv.org/abs/2406.12000

Caballar, R. D., & Stryker, C. (2025, March 14). *What is reasoning in AI?* Ibm.com. https://www.ibm.com/think/topics/ai-reasoning

Chan, L. (2023). *Beware General Claims about "Generalizable Reasoning Capabilities"* (of Modern AI Systems). Alignmentforum.org. https://www.alignmentforum.org/posts/5uw26uDdFbFQgKzih/beware-general-claims-about-generalizable-reasoning

Dmitry Vaintrob. (2023). *Grokking, memorization, and generalization — a discussion.* Lesswrong.com. https://www.lesswrong.com/posts/BYwGEBspGgPY5nBZN/grokking-memorization-and-generalization-a-discussion

Gomez, F. (2025). *Adapting Insider Risk mitigations for Agentic Misalignment: an empirical study*. Arxiv.org. https://arxiv.org/html/2510.05192v1

Lin, Z., Wang, X., Yang, H., Chai, J., Cao, J., Yin, G., Lin, W., & He, R. (2025). *AWPO: Enhancing Tool-Use of Large Language Models through Explicit Integration of Reasoning Rewards*. ArXiv.org. https://arxiv.org/abs/2512.19126

OpenAI. (2023). *GPT-4 Technical Report*. Arxiv.org. https://arxiv.org/html/2303.08774v4/#S5

Peng, X., Xia, C., Yang, X., Xiong, C., Wu, C.-S., & Xing, C. (2024). *ReGenesis: LLMs can Grow into Reasoning Generalists via Self-Improvement*. ArXiv.org. https://arxiv.org/abs/2410.02108?utm_source=chatgpt.com

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets*. ArXiv.org. https://arxiv.org/abs/2201.02177

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R. H., Fitz, S., & Hendrycks, D. (2024). *Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?* ArXiv.org. https://arxiv.org/abs/2407.21792

What is instrumental convergence? (2022). Aisafety.info. https://aisafety.info/questions/897I/What-is-instrumental-convergence

Xia, P., Zeng, K., Liu, J., Qin, C., Wu, F., Zhou, Y., Xiong, C., & Yao, H. (2025). *Agent0: Unleashing Self-Evolving Agents from Zero Data via Tool-Integrated Reasoning*. ArXiv.org. https://arxiv.org/abs/2511.16043

Yuan, X., Zhang, C., Liu, Z., Shi, D., Pan, L., Vosoughi, S., & Lee, W. (2025). *Superficial Self-Improved Reasoners Benefit from Model Merging*. ArXiv.org. https://arxiv.org/abs/2503.02103

Zhao, H., Kaur, S., Yu, D., Goyal, A., & Arora, S. (2024). *Can Models Learn Skill Composition from Examples?* ArXiv.org. https://arxiv.org/abs/2409.19808