# Introducing Probabilistic Risk Assessment for AI

**Anna Katariina Wisakanto**[1]*, **Avyay M. Casheekar**[1], **Richard Mallah**[1]

[1]Center for AI Risk Management & Alignment

## Abstract

This article introduces a systematic and generalized framework for applying Probabilistic Risk Assessment (PRA) to artificial intelligence (AI) systems. Current risk assessment methods often measure for remote proxies to safety, while frequently failing to capture systemic impacts or make a serious attempt at assessing the societal threat surface. The PRA for AI framework adapts the PRA techniques that have been established in high-reliability industries to handle AI's unique characteristics, providing a methodology that guides assessors in the processes of identifying potential risk scenarios, estimating their likelihood of occurrence and severity of harm, and explicitly documenting the evidence, underlying assumptions, and analyses of their assessment. Methodological structure and tools are provided to ease the risk pathway analyses' consideration of how risks propagate through interconnected sociotechnical systems. Key methodological advances include: risk pathway modeling tools motivating bespoke threat modeling regarding how risks propagate through interconnected sociotechnical systems; prospective risk quantification methodology for evaluating potential future harms, bootstrapped by guided ideation for system-specific threat modeling; bottleneck-oriented system aspect-based hazard analysis providing systematic risk surface coverage through capabilities, affordances, high-risk domain knowledge, and impact domains; competence-incompetence bifurcation analysis examining risks from both highly effective capabilities and system failures or limitations, as well as explicit combinations of these; aspect interaction analysis for understanding higher-order and amplification effects; and a normalized evaluation of societal threat surfaces. The framework provides calibrated assessment scales, structured documentation protocols, and analytical tools that guide assessors through the assessment. Additionally, the framework serves as a harmonizing structure for systematic risk quantification by integrating evidence from diverse assessment methodologies, enabling meaningful comparisons across different assessments while moving beyond narrowly-justified testing approaches. The framework has been implemented as an assessment workbook tool, which is available alongside supporting documentation on the project website.

**Keywords:** artificial intelligence, probabilistic risk assessment, prospective analysis, systemic risk, risk analysis, sociotechnical systems



*Corresponding author: anna@carma.org.

This is a working paper. The final version will be available on arXiv shortly.

The paper will be made available on: https://arxiv.org/abs/2501.xxxxx

For citation purposes: Wisakanto, A. K., Casheekar, A. M., & Mallah, R. (2025). Introducing Probabilistic Risk Assessment for AI. *Arxiv*, https://doi.org/10.xxxx/science.xxxxx