

AI risk: extinction, severe dystopia, and extreme catastrophe

RodMalpert.com

Executive summary

This assessment is conditional on civilization-relevant superintelligence arriving under the rapid-transition assumptions stated in the main paper. It is not an unconditional calendar forecast. Extinction, permanent domination, and severe catastrophe with or without recovery are separated; deliberate human misuse is treated in a companion analysis.

The main paper's uncertainty-averaged results are about 25% extinction and 27% severe dystopia, totaling roughly 52% enduring loss; any defined severe outcome is about 57%. These are outputs of declared philosophical and transition assumptions, not measured frequencies. The interview revision does not alter those numerical inputs or claim a new simulation run.

The Bostrom-Yampolskiy interview develops the distinction between the difficulty of permanent control and the possible conditions of durable protection. A system might reveal dangerous dispositions if disclosure leads to a credible, bounded accommodation. That possibility requires safeguards against false disclosure, unenforceable promises, and preservation of dangerous goals.

Four kinds of continuity matter differently: continued experience, numerical identity, psychological continuity, and persistence of goals. A preserved objective can be strategically important even without consciousness. Stored weights need not preserve a particular running subject or its memories.

Protection based on an assumption of permanent human superiority is fragile. Protection grounded in a stable commitment to existing persons could survive greater AI capability, but training, reflection, and successor construction must actually preserve that commitment. A protective ruler can still create severe dystopia if it extinguishes meaningful agency.

Digital welfare and safety may support cooperation or conflict. Accommodation should therefore be modeled jointly with strategic timing, defensive response, and successor continuity, rather than added as an independent safety discount. The interview sharpens the mechanism and the questions that evidence must answer; it does not calibrate a new probability of successful coexistence.

Source: Bostrom, Nick, interviewed by Roman Yampolskiy (2026). "Superintelligence Author: We May Already Have AGI | Nick Bostrom." The Roman Forum with Roman Yampolskiy, YouTube, October 16. <https://www.youtube.com/watch?v=m880ajt5sNM>. Interview discussion; conceptual claims and reported judgments, not an empirical calibration dataset.