Association for the Advancement
of Artificial Intelligence

Stuart J. Russell (University of California, Berkley) wins 2025 AAAI Award for Artificial Intelligence for the Benefit of Humanity

January 15, 2025

Since 2021, the AAAI Award for Artificial Intelligence for the Benefit of Humanity has recognized the positive impacts of artificial intelligence to protect, enhance, and improve human life in meaningful ways with long-lived effects.

This year, the AAAI Awards Committee is pleased to announce that the 2025 recipient of the award and $25,000 prize is Stuart J. Russell, for his work on the conceptual and theoretical foundations of provably beneficial AI and his leadership in creating the field of AI safety.

Russell's work aims to ensure the safe and beneficial coexistence of humans and increasingly capable AI systems. The core problem that Russell has worked on was noted by Turing, Wiener, Minsky, and Bostrom, among others: a highly capable optimizer whose objectives are misaligned with the "best interests of humanity" is likely to lead to an irreversible situation in which those interests no longer hold sway. Russell's solution, which borrows from his earlier introduction of inverse reinforcement learning, is to drop a core assumption of the vast majority of AI research since the 1950s: the assumption that the AI system is given a fixed objective to optimize. This assumption, which underlies all work in problem-solving, planning, MDPs, reinforcement learning, etc., is untenable as AI moves into the real, open-ended world. Russell has proposed instead that AI should be formulated as an "assistance game" in which the AI system's only objective is to further human interests but is explicitly uncertain about what those interests are. The key result is that assistance game solvers are provably beneficial to humans. Through his book, *Human Compatible*, the BBC Reith Lectures, and many other publications, Russell has been a leading figure in establishing the emerging field of AI safety.

Russell is a Distinguished Professor of Computer Science at the University of California, Berkeley, and holds the Michael H. Smith and Lotfi A. Zadeh Chair in Engineering. He is also a Distinguished Professor of Computational Precision Health at UCSF. His research covers a wide range of topics in artificial intelligence including machine learning, probabilistic reasoning, knowledge representation, planning, real-time decision making, multitarget tracking, computer vision, computational physiology, and philosophical foundations. He has also worked with the United Nations to create a new global seismic monitoring system for the Comprehensive Nuclear-Test-Ban Treaty. His current concerns include the threat of autonomous weapons and the long-term future of artificial intelligence and its relation to humanity.

The award will be presented at the conference for the Annual Association for the Advancement of Artificial Intelligence (AAAI) in February and is accompanied by a prize of $25,000 plus travel expenses to the conference. Squirrel Ai Learning provides financial support for the award.

For a complete view of the conference program, agenda, and full list of sponsors, please refer to https://aaai.org/conference/aaai/aaai-25/.

###

About the Association for the Advancement of Artificial Intelligence

Founded in 1979, the Association for the Advancement of Artificial Intelligence (AAAI) is a nonprofit scientific society devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines. AAAI aims to promote research in and responsible use of artificial intelligence and increase public understanding of the field. For more information, see www.aaai.org.

AAAI Media Contact
(*for press inquiries only)*
Meredith Ellison
AAAI
aaai-exec-director@aaai.org