

Thomas Gaskill

Cardinal Newman High School

Newman Oxford Scholars

12 November 2025

### Peace and AI: Stuart Russell

Stuart Russell has been working with AI for 45 years, and is widely respected for co-authoring *Artificial Intelligence: A Modern Approach* which has been used in 1,500 universities across 135 countries, and influenced decades of AI researchers (Stuart Russell). Despite being so involved in the development of AI, he has actively advocated for ensuring that AI systems are created in such a way that they lead to the benefit of humanity rather than its demise. His mission has largely focused on developing beneficial AI, which he calls human-compatible AI, and banning Lethal Autonomous Weapon Systems (LAWS). To those ends he has served as the Director of the Center for Human-Compatible AI (HCAI) at UC Berkeley, engaged in hundreds of media interviews and events, and given several presentations at the United Nations.

Stuart Russell works to align AI use and development with the preservation of peace, advancing human-compatible AI and encouraging an international ban on LAWS, so that human beings may live their lives in relative security. His efforts go beyond practical benefit, emphasizing the importance of forward-thinking, perseverance, and preemptive work toward a peaceful future during a time of rapid technological transformation.

Russell has been actively involved in changing sentiments on LAWS through media interviews, events, and public mobilization, but has faced significant challenges. LAWS according to the UN are, “weapons that, once activated, can select and engage targets without

further human intervention”(Russel, *Banning*). These weapons are not science fiction, as Stuart Russell explains in *Banning Lethal Autonomous Weapons: An Education*. Both the US and Russia, along with many other countries, have been opposed to a treaty to ban LAWS, and are interested in the technology. Russel emphasizes that currently, the technology exists to build these weapons. He explains mobile platforms that are suitable for LAWS, "quadcopters ranging from 3 centimeter to 1 meter in size; fixed-wing aircraft ranging from hobby-sized package delivery planes to full-sized missile-carrying drones; self-driving cars, trucks, and tanks; swarms of armed, unmanned boats” (Russel, *Banning*). According to a 2021 UN report, Kargu quadcopter drones, developed in Turkey, were used to hunt down and remotely engage retreating members of a faction in Libya in 2020 (Russell, *Banning*). Russell suggests that these weapons are a significant safety risk, and therefore must be stopped to ensure peace.

Russell clarifies that debating the issue of LAWS in the public realm is challenging because of pre-existing misconceptions. Particularly, that they are science fiction or incapable of significant destruction. In order to address this misconception, he worked to create a film grounded in near-current technology called *Slaughterbots* which premiered at a Convention on Certain Conventional Weapons (CCW) meeting in November 2017. Rather than humanoid robots, generally associated with killing, the robots were palm-sized drones capable of autonomously seeking targets by facial recognition and killing by exploding at close range. The video demonstrated that potentially in the present, or in the near future the technology will exist to cheaply produce autonomous weapons and will significantly lower the barrier to assassinations and mass killing without human intervention.

Russell has cautioned against building standard AI systems, even if for a benign purpose, if they have the potential to harm humans. As director of HCAI at UC Berkeley, he is able to

actively cultivate human-compatible AI. His work on human-compatible AI focuses on three principles: machines work to satisfy human preferences, the machine is uncertain about those preferences, and recognizes human behavior as the source of evidence for human preferences (Pomeroy). What these principles create is a system that will defer to humans whenever it is uncertain if an action it may take conflicts with their preference. On the contrary, standard single objective systems are much less likely to defer to humans. In order to illustrate this issue, he has used King Midas, a character from Greek mythology. Midas asked for everything he touched to be turned to gold. However, his food, drink, and family members were all turned to gold as a consequence of that single objective, and he died in starvation (Russel, *Principles*).

Russell's perseverance through slow tangible progress demonstrates the weight of problems which challenge peace, such as those created by AI. Russels advocacy, for over a decade, has raised awareness but not made significant progress in banning or regulating LAWS. He explains the challenge, "Efforts to ban AWS are at a standstill because the American and Russian governments, supported to some extent by Britain, Israel, and Australia, argue that a ban is unnecessary" (Russell, *Banning*). His choice to continue advocating against LAWS despite countries in opposition resembles the struggle for achieving peace that so many have experienced in pursuing their objective. Like him, those who work for peace do not need to be discouraged, but continue progressing toward their mission.

Russell shows that the requirements for peace evolve, and can transform rapidly. With this rapid transformation, some must be at the frontier to ensure peace. Russell's advocacy against harmful AI systems largely began in 2013, when he was contacted by Human Rights Watch (HRW) to support a campaign to ban killer robots (Stuart, *Banning*). His decision to move quickly, before AI gained momentum through the release of powerful LLMs like ChatGPT,

reveals that creating peace can require forward-thinking individuals that can see and respond to problems before others can. Considering this, society should not be dismissive of efforts to seek and ensure peace, simply because a problem is not yet at the point of causing harm. Russell exemplifies that society can learn to look ahead at technology, and recognize its true impact, rather than waiting until damage has been done.

Stuart Russell's efforts serve to ensure AI supports peace and safety as it rapidly develops. Through the challenge of convincing nations to agree on a ban of LAWS and encouraging companies to prioritize human-compatible AI reflects admirable resolve and dedication toward peace. He demonstrates how forward-thinking can be used to cultivate future peace by preemptive action, through his work at HCAI and his significant presence advocating for a ban on LAWS. While accomplishing his goals has proven challenging, his contributions to raising awareness among society and diplomats about AI concerns has been a significant and inspirational step toward peace.

## Works Cited

Russell, Stuart. "Banning Lethal Autonomous Weapons: An Education." *Issues in Science and*

*Technology*, vol. XXXVIII, no. 3, Spring 2022,

<https://issues.org/banning-lethal-autonomous-weapons-stuart-russell/>.

Russell, Stuart. "3 Principles for Creating Safer AI." *TED*, April 2017,

[https://www.ted.com/talks/stuart\\_russell\\_3\\_principles\\_for\\_creating\\_safer\\_ai](https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai).

Pomeroy, Robin, host. "The promises and perils of AI - Stuart Russell on Radio Davos." *Radio*

*Davos*, World Economic Forum, 6 Jan. 2022,

<https://www.weforum.org/stories/2022/01/artificial-intelligence-stuart-russell-radio-davos>

[L](https://www.weforum.org/stories/2022/01/artificial-intelligence-stuart-russell-radio-davos).

"Stuart Russell." *Research UC Berkeley*, UC Berkeley,

<https://vcresearch.berkeley.edu/faculty/stuart-russell>.