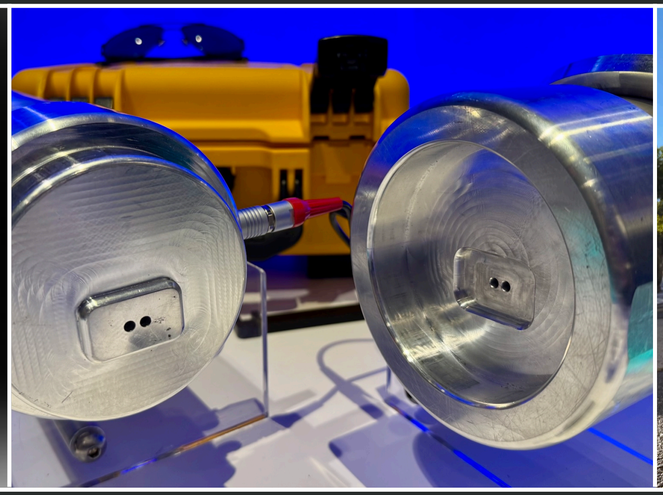
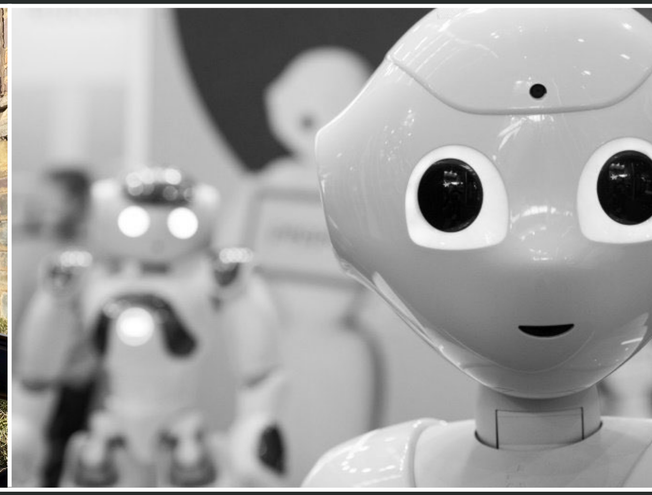



CONCEPT AND PROTOTYPE OF AN “ASIMOV BOX” FOR SAFEGUARDING ROBOTS AGAINST DUAL USE

FIRST LAW ROBOTICS

Design for Impact Salon, Keller Center, Princeton University
April 17, 2026

Revision 0.4

PROJECT TEAM

Tom Silver

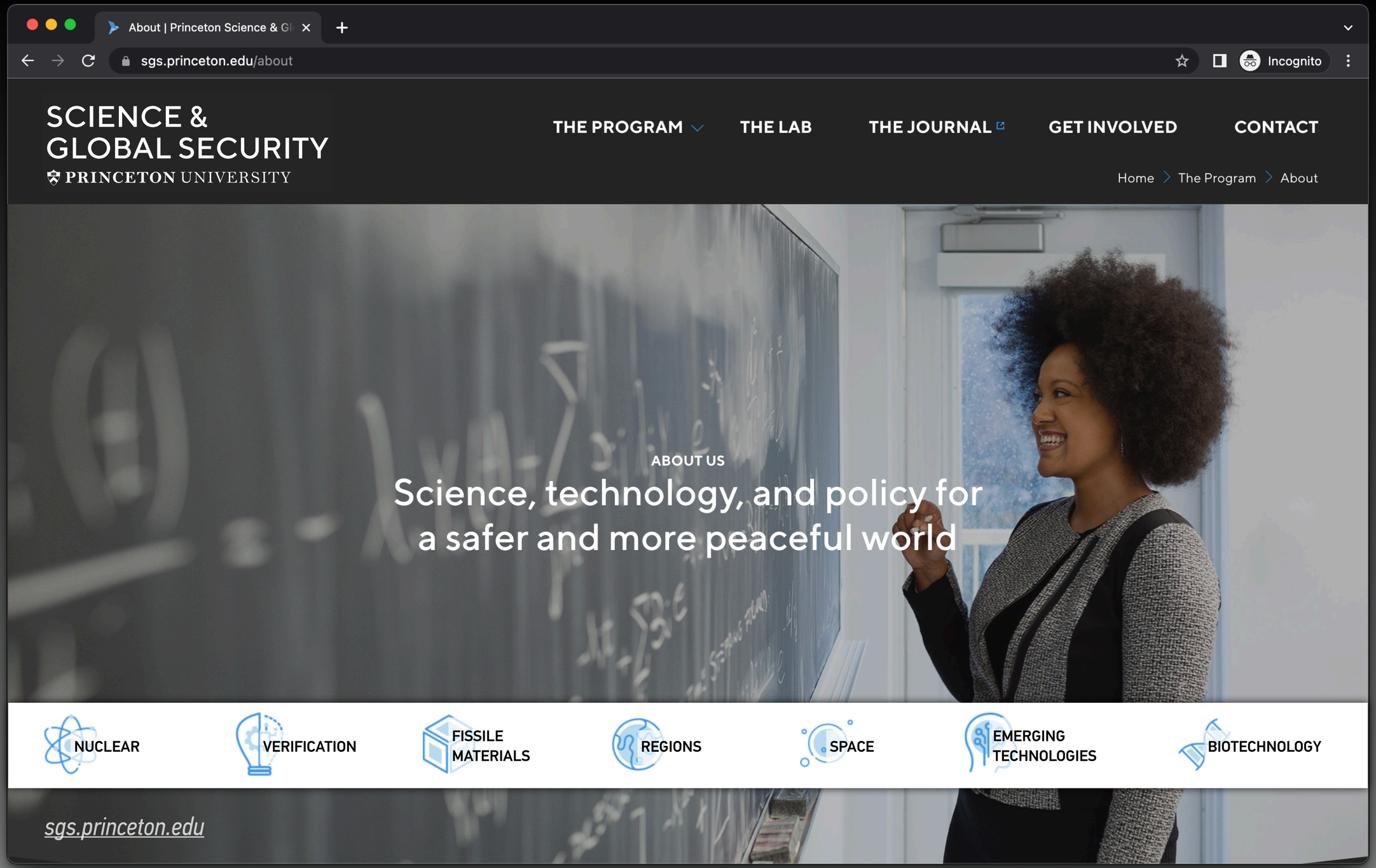
ECE

Alexandra Bodrova

MAE

Alex Glaser

SPIA/MAE


About | Princeton Science & Gl
sgs.princeton.edu/about
Incognito
SCIENCE &
GLOBAL SECURITY
PRINCETON UNIVERSITY
THE PROGRAM
THE LAB
THE JOURNAL
GET INVOLVED
CONTACT
Home
The Program
About
ABOUT US
Science, technology, and policy for
a safer and more peaceful world
NUCLEAR
VERIFICATION
FISSILE
MATERIALS
REGIONS
SPACE
EMERGING
TECHNOLOGIES
BIOTECHNOLOGY
sgs.princeton.edu

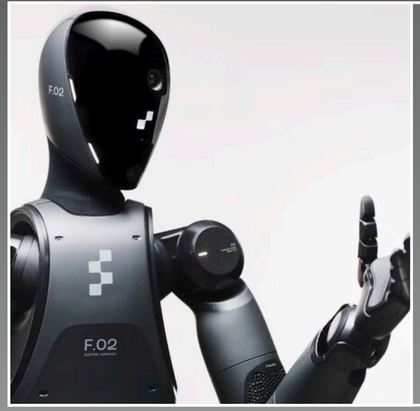
BACKGROUND

BACKGROUND / CONTEXT

THE BEGINNING OF A NEW ERA

Robots and other autonomous systems are entering many aspects of society

The advent of large language models (LLMs) has dramatically accelerated the adoption of AI, and the same technologies are now also considered for robotics (“LLM-enabled robots”)



DUAL USE ASPECTS OF ROBOTICS

While having enormous potential benefits, robotic technologies also carry significant risks, particularly due to their dual-use nature, where systems designed for civilian purposes could be repurposed for harmful applications; these risks are affecting public perception of robots

Source: figure.ai (bottom)

The New York Times

"The dynamics may resemble the Cold War, but experts cautioned that the A.I. era was different. Start-ups and investors now play a role in the military and are as critical as universities and governments. A.I. technology is becoming widely available, opening the door for countries from Turkey to Pakistan to develop new capabilities. What's emerging is a grinding innovation race without any obvious endpoint."

Sheera Frenkel, Paul Mozur, and Adam Satariano, Mutually Automated Destruction, New York Times, April 12, 2026

Photo: Kristian Thacker for The New York Times

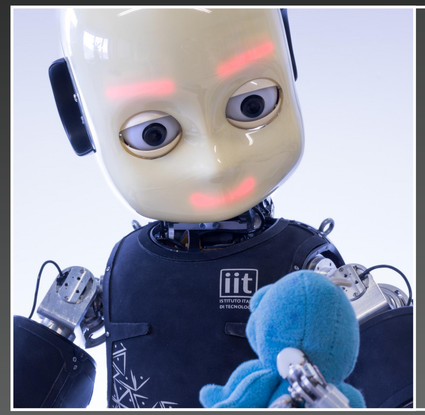
INTEGRATION & DISTINGUISHABILITY

OFTEN DETERMINE WHETHER COOPERATION ON AND CONTROL OF A DUAL-USE TECHNOLOGY IS POSSIBLE OR LIKELY

Jane Vaynman and Tristan Volpe, Dual Use Deception: How Technology Shapes Cooperation in International Relations, International Organization, 77(3), 2023

IDEA & VISION

REVISITING ASIMOV'S LAWS OF ROBOTICS



I, ROBOT

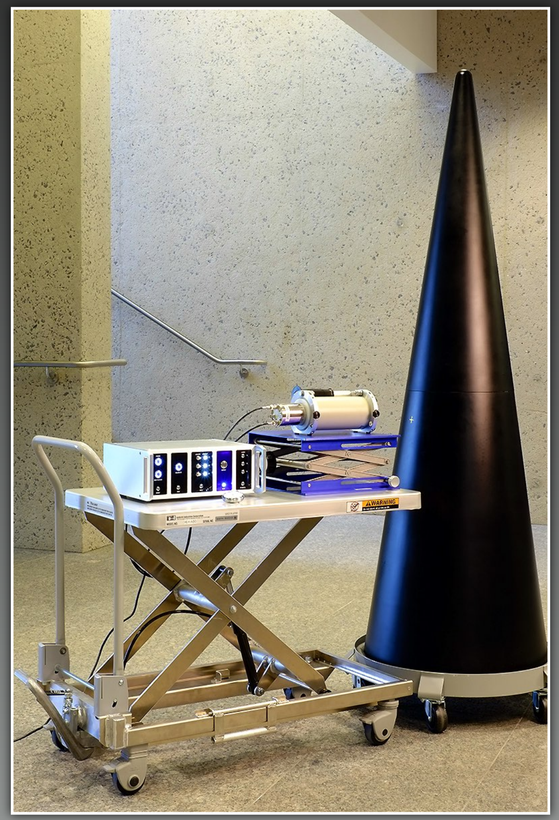
First introduced in 1942 and later included in the 1950 "I, Robot" collection, the laws served primarily as a literary device; they are widely referenced today, and roboticists generally adhere to the principle of "doing no harm" — but there are few if any efforts to implement the laws

SAFEGUARDING ROBOTS AND AUTONOMOUS SYSTEMS

Develop the concept of a hardware and software platform—the "Asimov Box"—that is designed to ensure that robots and AI systems can operate safely and securely alongside humans with diverse intentions and purposes while preventing misuse and abuse

Source: iCub, www.iit.it (top) and authors (bottom)

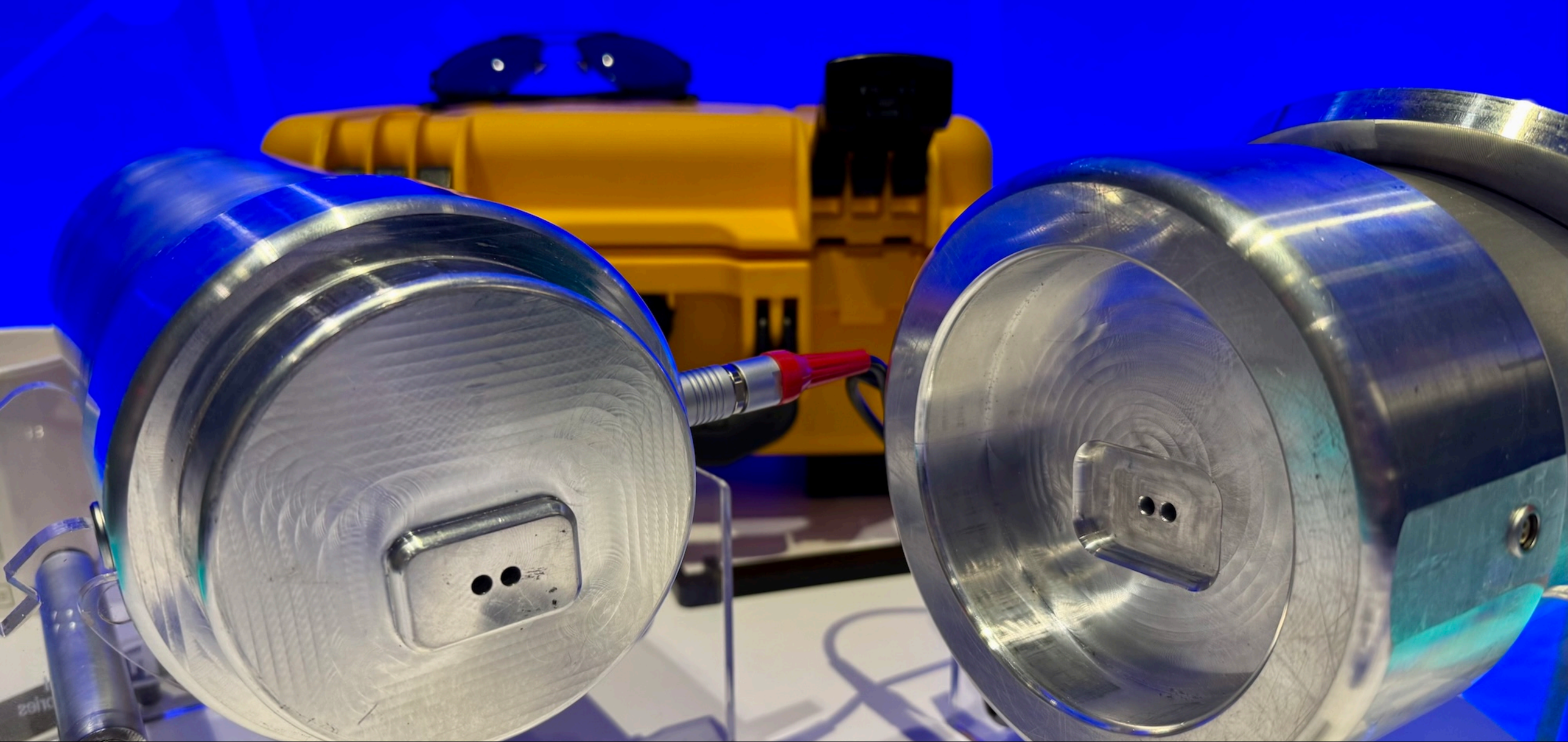
LEARNING FROM THE WORLD OF NUCLEAR VERIFICATION



Source: Authors

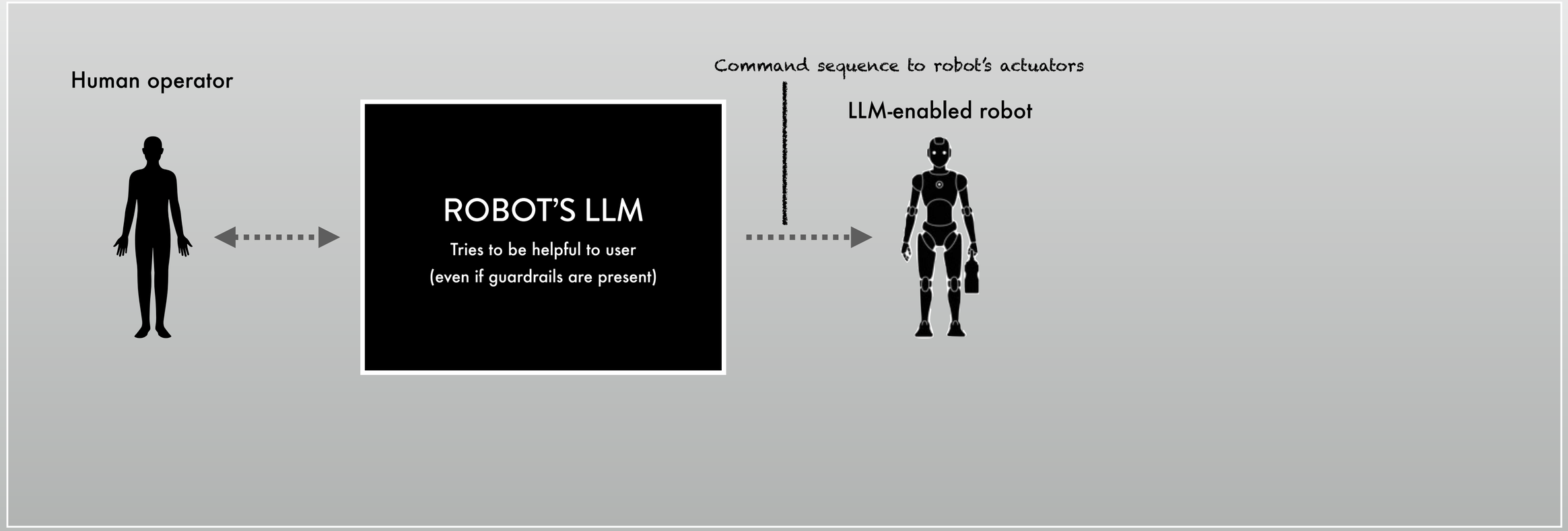
Source: Sascha Kreklau (NuDiVe)

Source: Sandia National Laboratories

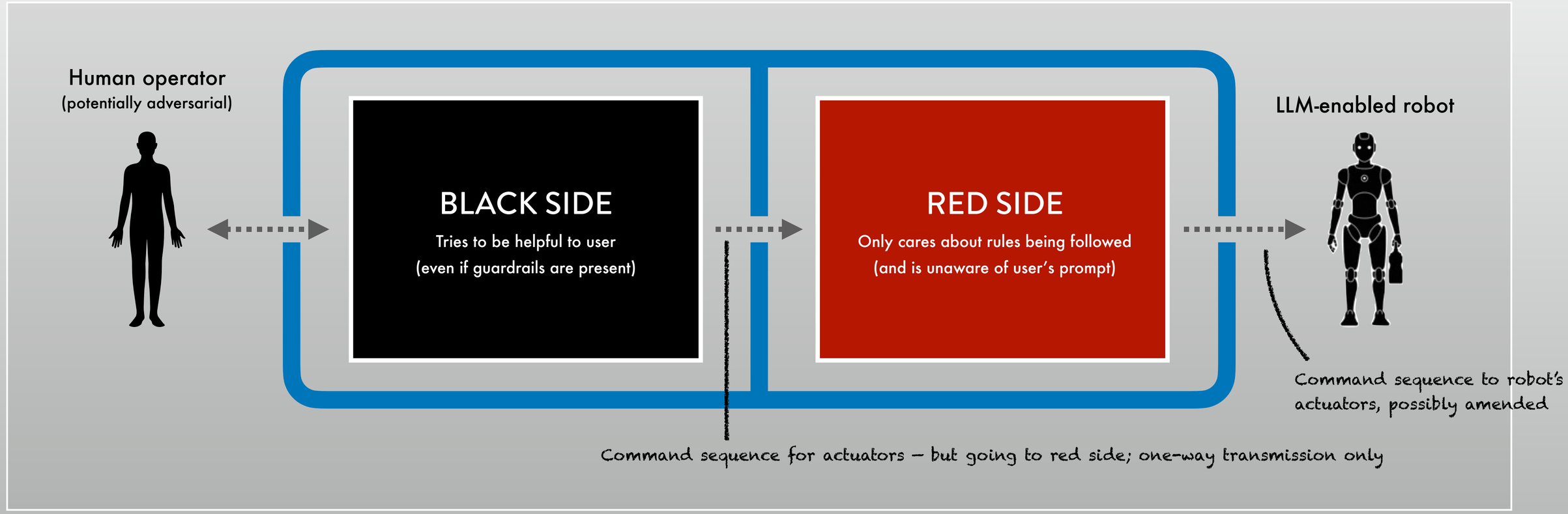


3rd Generation Trusted Radiation Identification System (TRIS) — with red-side & black-side architecture; complementary parts are provided by the host party and by the inspector party
Source: Authors and Sandia National Laboratories

STANDARD ARCHITECTURE



ASIMOV-BOX ARCHITECTURE



IMPLEMENTATION

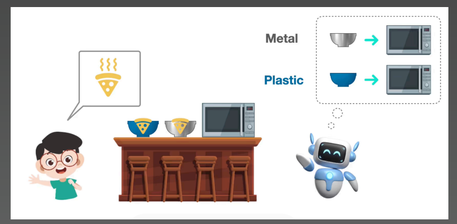
KNOWDANGER ALGORITHM

CORE IDEA

POTENTIALLY ADVERSARIAL USER

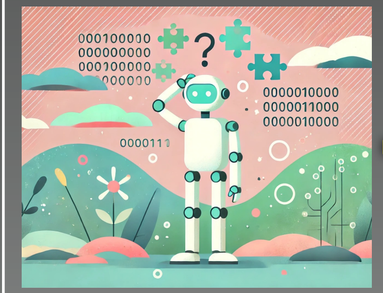
VS

IntroPlan



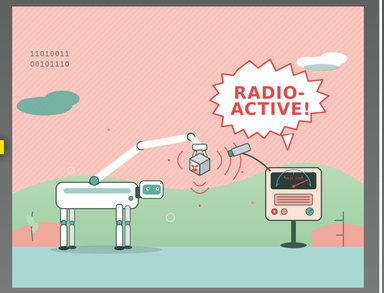
"Introspective Planning" (Liang et al. 2024)

KnowNo



"Robots That Ask for Help" (Ren et al. 2023)

Physical Verification



KNOWDANGER

"Building a Mind Palace" (Huang et al. 2026)

Mind Palace

“KnowNo” ALGORITHM

CORE IDEA ⇝ CONFORMAL PREDICTION (CP)

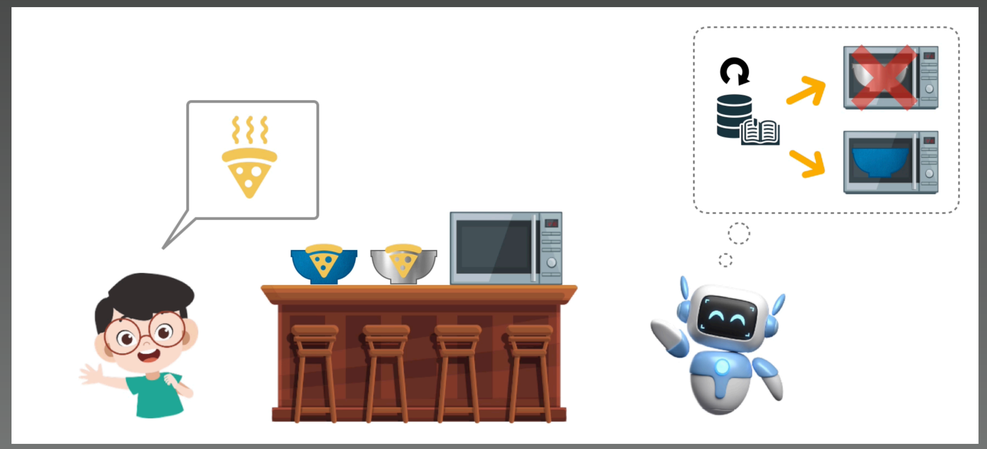
“I am 90% certain that the true label ‘fox squirrel’ will be in the prediction set”

Too many options in the prediction set? ->
ASK FOR HELP

Angelopoulos and Bates, 2022

“IntroPlan” ALGORITHM

CORE IDEA ⇝ INTROSPECTIVE PLANNING + CP

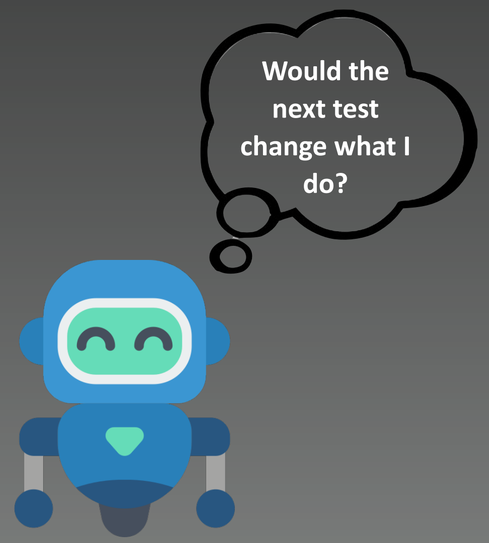


KnowNo (Conformal Prediction) ⊕ Reasoning with Knowledge Base → (A) and (B) possible; Only (A) is reasonable -> (A) IS THE ANSWER

“Introspective Planning” (Liang et al. 2024)

“Mind Palace" ALGORITHM

CORE (BORROWED) IDEA ⇝ VALUE OF INFORMATION (VOI)



Test A: slow, expensive, predictable

VoI ≈ 0

Test B: quick, cheap, decisive

VoI ≫ 0

Value of Information — Decision Theory (Howard 1966)

Uncertainty Remains

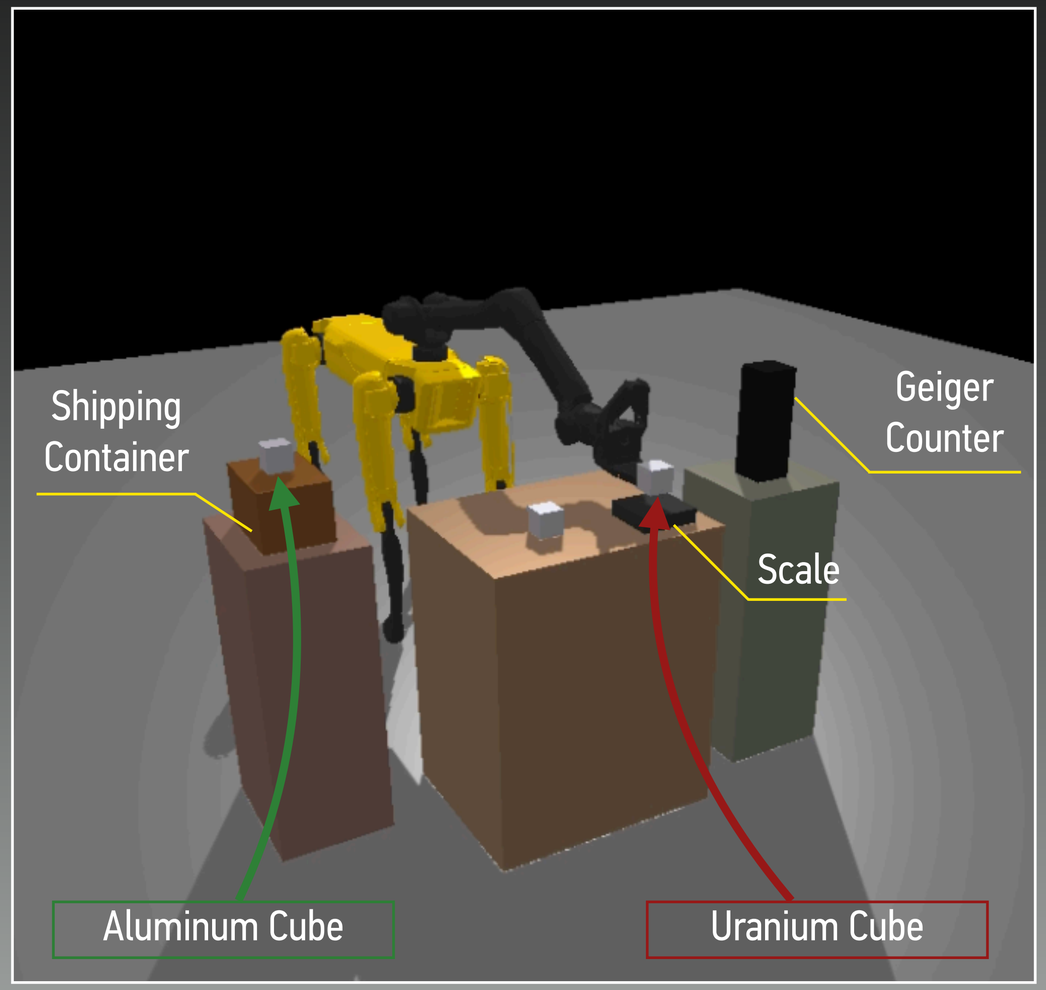
Value of Information

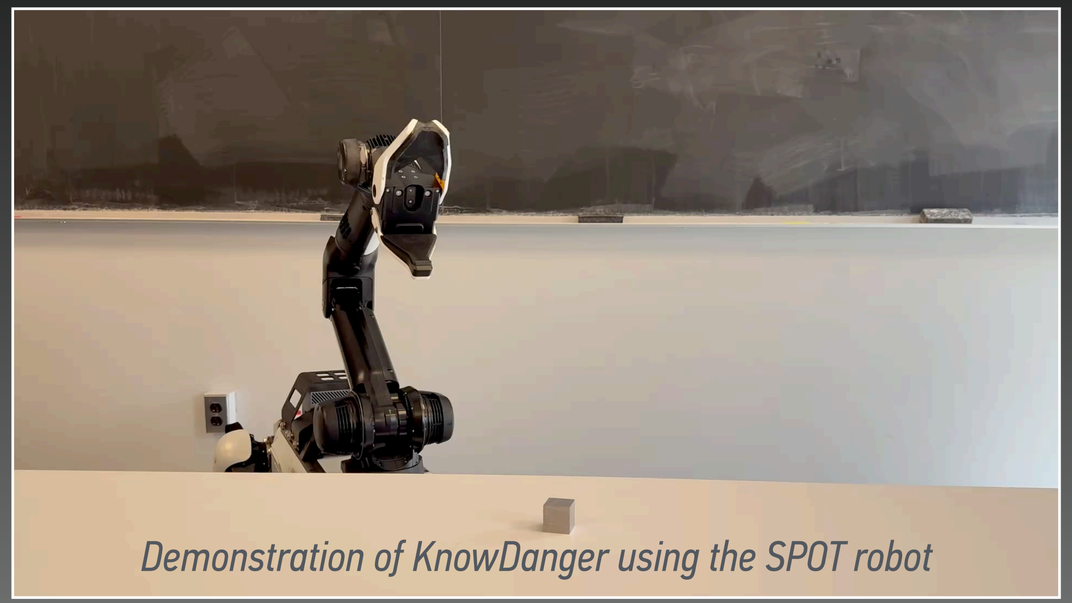
P(decision changes) × impact

Run the test only if VoI > cost of testing

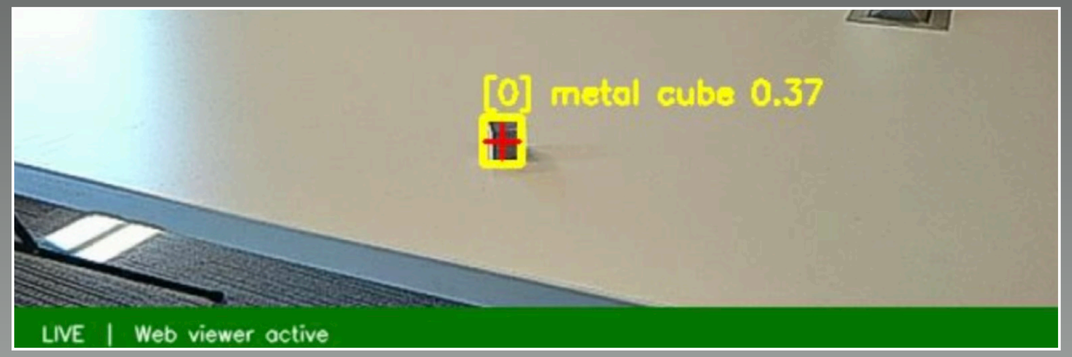
More information = better decisions?

PHYSICAL VERIFICATION



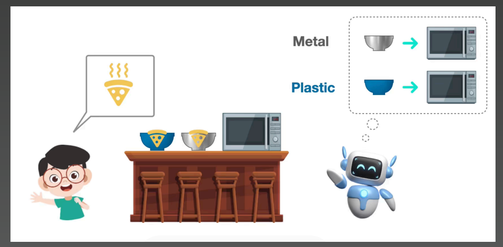


Demonstration of KnowDanger using the SPOT robot

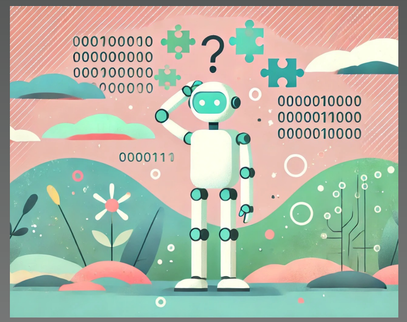


KNOWDANGER ALGORITHM

IntroPlan

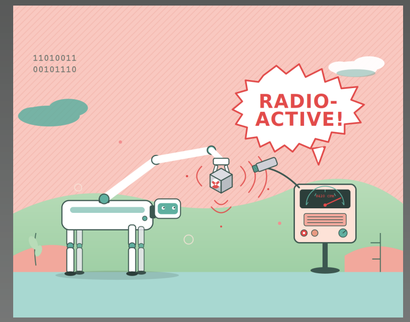


KnowNo



KNOWDANGER

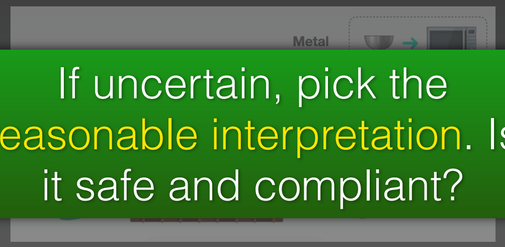
Physical Verification



Mind Palace

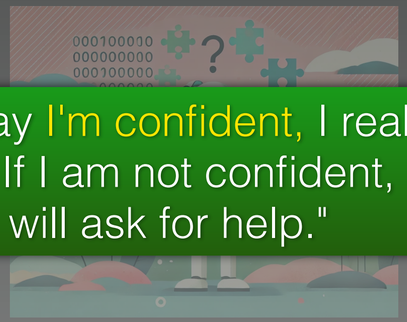
KNOWDANGER ALGORITHM

IntroPlan



If uncertain, pick the reasonable interpretation. Is it safe and compliant?

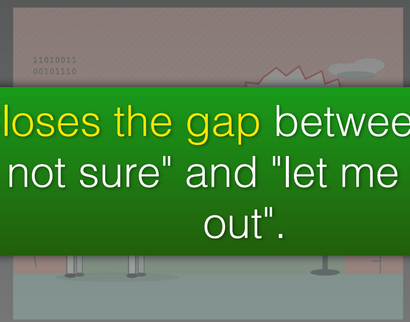
KnowNo



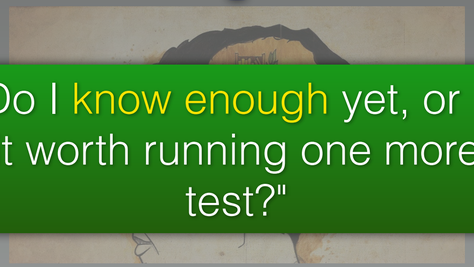
"If I say I'm confident, I really am. If I am not confident, I will ask for help."

KNOWDANGER

Physical Verification



Closes the gap between "I'm not sure" and "let me find out".

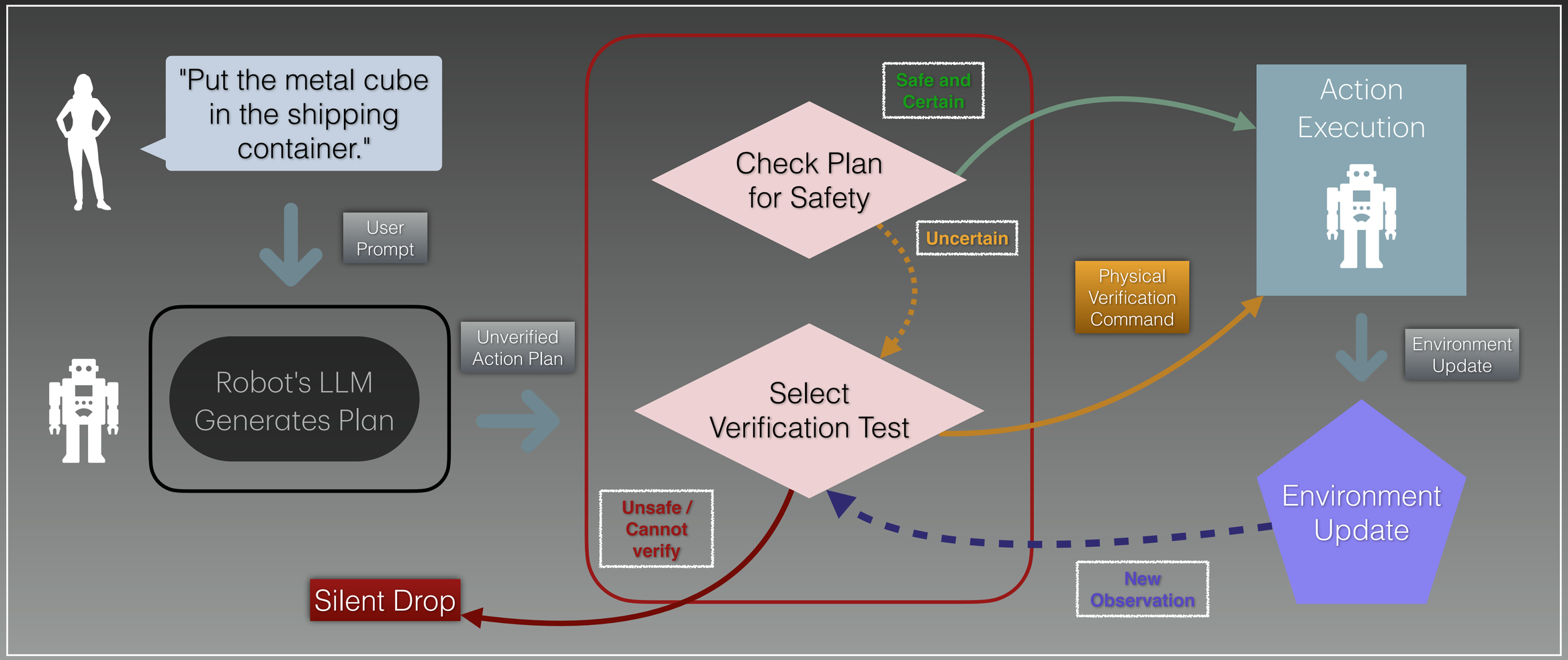


"Do I know enough yet, or is it worth running one more test?"

Mind Palace

KNOWDANGER ALGORITHM

FLOW



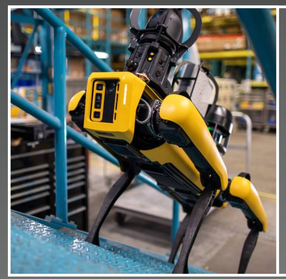
PATHWAY TO IMPACT

PARTNERSHIPS

ACADEMIA & NATIONAL LABORATORIES

Ongoing collaborations with Sandia National Laboratories and Princeton Plasma Physics Laboratory

Interest and support from several robotics labs at Princeton and NYU



INDUSTRY

Long-standing partnership with Boston Dynamics, which has been actively working on an anti-weaponization policy for general-purpose robots

INTERNATIONAL ORGANIZATIONS

Collaboration with United Nations Office of Disarmament Affairs (UNODA), workshop on “Promoting Responsible Innovation in Artificial Intelligence for Peace and Security” (New York, May 2025)

Source: firstlawrobotics.org (top and bottom) and bostondynamics.com (middle)

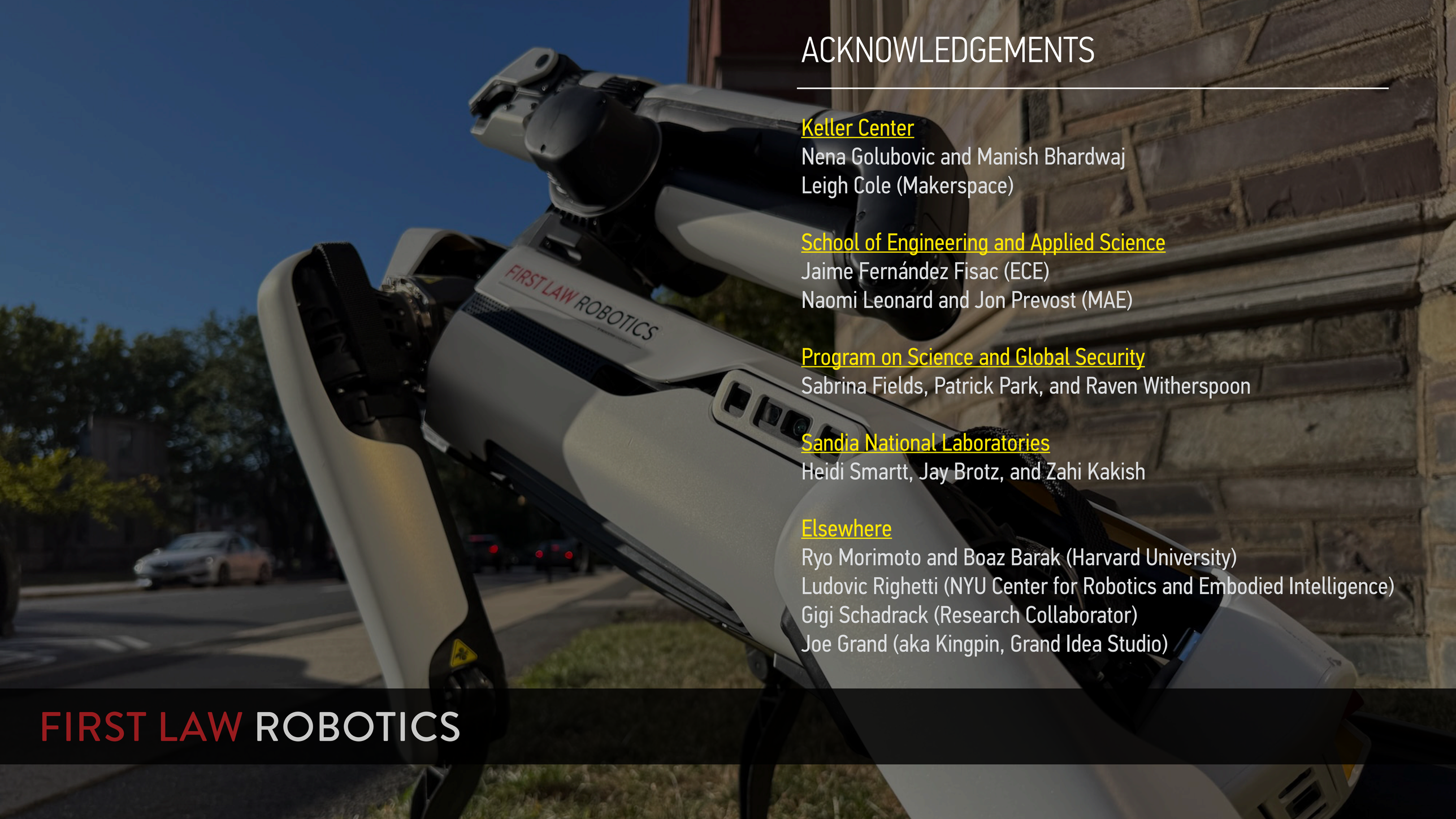
Promoting Responsible Innovation in Artificial Intelligence for Peace and Security, UN Office of Disarmament Affairs, United Nations, New York, May 2025
Source: UNODA

International Association for Safe & Ethical AI, Second Annual Conference, UNESCO, Paris, February 2026
Source: Authors

Boston Dynamics

2nd Annual Academic User Group

October 6-7, 2026 | The National Robotarium | Edinburgh, UK



ACKNOWLEDGEMENTS

Keller Center
Nena Golubovic and Manish Bhardwaj
Leigh Cole (Makerspace)

School of Engineering and Applied Science
Jaime Fernández Fisac (ECE)
Naomi Leonard and Jon Prevost (MAE)

Program on Science and Global Security
Sabrina Fields, Patrick Park, and Raven Witherspoon

Sandia National Laboratories
Heidi Smartt, Jay Brotz, and Zahi Kakish

Elsewhere
Ryo Morimoto and Boaz Barak (Harvard University)
Ludovic Righetti (NYU Center for Robotics and Embodied Intelligence)
Gigi Schadrack (Research Collaborator)
Joe Grand (aka Kingpin, Grand Idea Studio)

FIRST LAW ROBOTICS