

# AI magazine

Volume 45 Number 1

Spring 2024

## BENEFICIAL AI



INTRODUCING THE  
**NATIONAL AI INSTITUTES**

### AI Magazine

Volume 45 Number 1 Spring 2024

#### NSF'S NATIONAL AI INSTITUTES

Guest Editors: Goel, Ashok; Georgia Institute of Technology & AI-ALOE,  
Ou, Chaohua; Georgia Institute of Technology & AI-ALOE

#### EDITORIAL

Introduction to the Special Issue 4  
*Ashok Goel and Chaohua Ou*

#### SPECIAL TOPIC ARTICLES

The National Artificial Intelligence Research Institutes program and its significance to a prosperous future 6  
*James J. Donlon*

Athena – The NSF AI Institute for Edge Computing 15  
*Yiran Chen, Suman Banerjee, Shaundra Daily, Jeffery Krolik, Hai (Helen) Li, Daniel Limbrick, Miroslav Pajic, Rajashi Runtou and Lin Zhong*

Creating intelligent cyberinfrastructure for democratizing AI 22  
*Dhabaleswar K. Panda, Vipin Chaudhary, Eric Fosler-Lussier, Raghu Machiraju, Amit Majumdar, Beth Plale, Rajiv Ramnath, Ponnuswamy Sadayappan, Neelima Savardekar and Karen Tomko*

AI-EDGE: An NSF AI institute for future edge networks and distributed intelligence 29  
*Peizhong Ju, Chengzhang Li, Yingbin Liang and Ness Shroff*

Institute for Foundations of Machine Learning (IFML): Advancing AI systems that will transform our world 35  
*Adam Klivans, Alexandros G. Dimakis, Kristen Grauman, Jonathan I. Tamir, Daniel J. Diaz and Karen Davidson*

AI4OPT: AI Institute for Advances in Optimization 42  
*Pascal Van Hentenryck and Kevin Dalmeijer*

AI Institute in Dynamic Systems: Developing machine learning and AI tools for scientific discovery, engineering design, and data-driven control 48  
*J. Nathan Kutz, Steven L. Brunton, Krithika Manohar, Hod Lipson and Na Li*

The TILOS AI Institute: Integrating optimization and AI for chip design, networks, and robotics 54  
*Andrew B. Kahng, Arya Mazumdar, Jodi Reeves and Yusu Wang*

From learning optimization to learner flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (iSAT) 61  
*Sidney K. D'Mello, Quentin Biddy, Thomas Breideband, Jeffrey Bush, Michael Chang, Arturo Cortez, Jeffrey Flanigan, Peter W. Foltz, Jamie C. Gorman, Leanne Hirshfield, Mon-Lin Monica Ko, Nikhil Krishnaswamy, Rachel Lieber, James Martin, Martha Palmer, William R. Penuel, Thomas Philip, Sadhana Puntambekar, James Pustejovsky, Jason G. Reitman, Tamara Sumner, Michael Tissenbaum, Lyn Walker and Jacob Whitehill*

The AI Institute for Engaged Learning 69  
*James Lester, Mohit Bansal, Gautam Biswas, Cindy Hmelo-Silver, Jeremy Roschelle and Jonathan Rowe*

AI-ALOE: AI for reskilling, upskilling, and workforce development 77  
*Ashok Goel, Chris Dede, Myk Garn and Chaohua Ou*

AIFARMS: Artificial intelligence for future agricultural resilience, management, and sustainability 83  
*Vikram S. Adve, Jessica M. Wedow, Elizabeth A. Ainsworth, Girish Chowdhary, Angela Green-Miller and Christina Tucker*

The AIFS Institute: Building a better food system through AI 89  
*Ilias Tagkopoulos, Mason J. Earles, Danielle G. Lemay, Xin Liu, Nitin Nitin, Aaron D. Smith, Tarek I. Zohdi and Stephen F. Brown*

AIIRA: AI Institute for Resilient Agriculture <i>Baskar Ganapathysubramanian, Jessica M. P. Bell, George Kantor, Nirav Merchant, Soumik Sarkar, Patrick S. Schnable, Michelle Segovia, Arti Singh and Asheesh K. Singh</i>	94
AgAID Institute—AI for agricultural labor and decision support <i>Alan Fern, Margaret Burnett, Joseph Davidson, Janardhan Rao Doppa, Paola Pesantez-Cabrera and Ananth Kalyanaraman</i>	99
AI2ES: The NSF AI Institute for Research on Trustworthy AI for Weather, Climate, and Coastal Oceanography <i>Amy McGovern, Imme Ebert-Uphoff, Elizabeth A. Barnes, Ann Bostrom, Mariana G. Cains, Phillip Davis, Julie L. Demuth, Dimitrios I. Diochnos, Andrew H. Fagg, Philippe Tissot, John K. Williams and Christopher D. Wirz</i>	105
Institute for Artificial Intelligence and Fundamental Interactions (IAIFI): Infusing physics intelligence into artificial intelligence <i>Jesse Thaler, Mike Williams and Marisa LaFleur</i>	111
Molecule Maker Lab Institute: Accelerating, advancing, and democratizing molecular innovation <i>Martin D. Burke, Scott E. Denmark, Ying Diao, Jiawei Han, Rachel Switzky and Huimin Zhao</i>	117
AI-CARING: National AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups <i>Sonia Chernova, Elizabeth Mynatt, Agata Rozga, Reid Simmons and Holly Yanco</i>	124
<b>HIGHLIGHTS</b>	
Prosocial dynamics in multiagent systems <i>Fernando P. Santos</i>	131
Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety <i>Manas Gaur and Amit Sheth</i>	139
Physical scene understanding <i>Jiajun Wu</i>	156
<b>COLUMN</b>	
Generative AI: An AI paradigm shift in the making? <i>Risto Miikkulainen</i>	165



Cover art credit: The cover image was generated by Midjourney, responding to a prompt that included notions of, “people from various professions (teachers, nurses, farmers, engineers, and artists), working together to create and guide AI to facilitate collaboration, innovation, and problem-solving for the common good.” While this is a challenging concept for man or machine to represent in a single image, this issue’s articles describing the U.S. National AI Research Institutes will paint richer portraits!

# AI magazine

AI MAGAZINE (Online ISSN 2371-9621) is published quarterly on behalf of the Association for the Advancement of Artificial Intelligence by Wiley Periodicals LLC.

## Open Access and Copyright

All articles published by AI Magazine [submitted after November 15 2021] are fully open access: immediately freely available to read, download, and share. All articles [accepted from November 12 2021] are published under the terms of a Creative Commons license.

Copyright on any research article published by AI Magazine is retained by the author(s).

Further information about open access licenses and copyright can be found at: <https://authorservices.wiley.com/author-resources/Journal-Authors/licensing/licensing-info-faqs.html#4>.

Information on the licenses used by AI Magazine can be found at: [wileyonlinelibrary.com/journal/aaai](http://wileyonlinelibrary.com/journal/aaai).

## Co-Editors in Chief

Odd Erik Gundersen, Norwegian University of Science and Technology, Norway

K. Brent Venable, Institute for Human and Machine Cognition and University of West Florida, USA

## Associate Editors

J. Benton, NASA Ames Research Center, USA

Joshua Eckroth, i2k Connect, USA

Andrea Loreggia, University of Brescia, Italy

Sandip Sen, The University of Tulsa, USA

Marija Slavkovic, The University of Bergen, Norway

Michael Wollowski, Rose-Hulman Institute of Technology, USA

## Editors in Chief Emeritus

Ashok Goel, Georgia Institute of Technology, USA

David Leake, Indiana University, USA

## Editorial Board Columnists

Ian Beaver, Verint Next IT, USA

Pushpak Bhattacharyya, Indian Institute of Technology, Bombay

Bistra Dilkina, University of Southern California, USA

Ashok Goel, Georgia Institute of Technology, USA

Odd Erik Gundersen, Norwegian University of Science and Technology, Norway

Praveen Paritosh, Google Inc., USA

Jim Spohrer, IBM, USA

Michael Wollowski, Rose-Hulman Institute of Technology, USA

## Advisory Board

David Aha, Naval Research Laboratory, USA

Henrik Christensen, University of California, San Diego, USA

Ashok Goel, Georgia Institute of Technology, USA

James Hender, Rensselaer Polytechnic Institute, USA

Sarit Kraus, Bar Ilan University, Israel

David Leake, Indiana University, USA

Melanie Mitchell, Santa Fe Institute, USA

Jordan Pollack, Brandeis University, USA

Brian Scassellati, Yale University, USA

Elizabeth Sonenberg, University of Melbourne, Australia

Zhi-Hua Zhou, Nanjing University, China

## Publications Manager

Ida Camacho. AAAI

## AAAI Officials

President  
Francesca Rossi, IBM Research, USA

Past President  
Bart Selman, Cornell University, USA

President-Elect  
Stephen Smith, Carnegie Mellon University, USA

Secretary-Treasurer  
David E. Smith, USA

## Submission Instructions

For submission instructions, subscription and all other information visit: [wileyonlinelibrary.com/journal/aaai](http://wileyonlinelibrary.com/journal/aaai).

## Print on Demand

This journal is available Print on Demand (POD) from our supplier Sheridan. To place your order please contact your usual agent or Sheridan directly at <http://www.sheridan.com/LPM/Wiley>. POD copies are available from Sheridan at 100% of the Online Only list price.

## Disclaimer

The Publisher, the Association for the Advancement of Artificial Intelligence and Editors cannot be held responsible for errors or any consequences arising from the use of information contained in this journal. The views and opinions expressed do not necessarily reflect those of the Publisher, the Association for the Advancement of Artificial Intelligence or Editors, neither does the publication of advertisements constitute any endorsement by the Publisher, the Association for the Advancement of Artificial Intelligence Editors, or Authors of the products advertised.

Wiley Open Access articles posted to repositories or websites are without warranty from Wiley of any kind, either express or implied, including, but not limited to, warranties of merchantability, fitness for a particular purpose, or non-infringement. To the fullest extent permitted by law Wiley disclaims all liability for any loss or damage arising out of, or in connection with, the use of or inability to use the content.

## Councilors (through 2023)

Edith Elkind, University of Oxford, UK

Susan Epstein, The City University of New York: Hunter College, USA  
Laura Hiatt, US Naval Research Laboratory, USA

Barry O'Sullivan, University College Cork, Ireland

## Councilors (through 2024)

Xuelong Li, Northwestern Polytechnical University, China  
Felipe Meneguzzi, Pontifical Catholic University of Rio Grande do Sul, Brazil  
Ana Paiva, University of Lisbon, Spain  
Julie Shah, Massachusetts Institute of Technology, USA

## Councilors (through 2025)

Maria Chang, IBM Research, USA  
Kristian Kersting, Technical University of Darmstadt, Germany  
Jihie Kim, Dongguk University, South Korea  
Anita Raja, The City University of New York, USA

## Standing Committees

AI-Hub Liaison  
Michael Littman, Brown University, USA

Awards, Fellows, and Nominating Chair  
Bart Selman, Cornell University, USA

CRA Liaison  
Yolanda Gil, USC Information Sciences Institute, USA

Communications Chair  
K. Brent Venable, Institute for Human and Machine Cognition and University of West Florida, USA

Conference Chair  
Kevin Leyton-Brown, University of British Columbia, Canada

Diversity and Inclusion Chair  
Maria Gini, University of Minnesota, USA

Education Chair  
Laura Hiatt, Naval Research Laboratory, USA

Ethics Chair  
Susan Epstein, The City University of New York: Hunter College, USA

Finance Chair  
David E. Smith, USA

Global AI Initiatives and Policy Chair  
Francesca Rossi, IBM Research, USA  
and Barry O'Sullivan, University College Cork, Ireland

Membership Chair  
Julie Shah, Massachusetts Institute of Technology, USA

Publications Chair  
Anthony Cohn, University of Leeds, UK

Symposium Chair and Cochair  
Christopher Geib, SIFT, USA  
Ron Petrick, Heriot-Watt University, UK

US Initiatives and Policy  
Bart Selman, Cornell University, USA

AAAI Staff  
Executive Director  
Meredith Ellison

Director of Operations  
Chesley Grove

Program Manager  
Ashley Short

Membership Manager  
Jeanne Glover



## EDITORIAL

# Introduction to the Special Issue

### Abstract

We briefly introduce this special issue and describe the scheme for the organization of the 20 articles in it.

In *Looking back, looking ahead: Strategic initiatives in AI and NSF's AI Institutes program*, Donlon and Goel (2023) briefly introduced the U.S. National Science Foundation's National AI Research Institutes program to the readers of this magazine. This program so far has funded 25 AI Institutes at about \$20 M each over an initial five years. The investment of about \$500 M makes it one of the largest single public investments to date into AI research and development. Together the 25 AI Institutes capture the excitement around the potential social and scientific benefits of AI. Details for each of the AI Institutes can also be found on the website of the AI Institutes Virtual Organization: <https://aiinstitutes.org/institutes>.

This special issue presents brief reports on the first 18 AI Institutes, 7 launched in late 2020 and 11 in late 2021. At the time of writing of this overview in September 2023, the first 7 AI Institutes have completed about 3 years of research and the next 11 Institutes have completed about 2 years of work. Figure 1 illustrates the two cohorts by the mnemonics for the AI Institutes; the articles too are indexed by the same mnemonics. Each article describes the goals, themes, early results, and broader impacts of an AI Institute.

The 20 articles in the special issue are organized into six groups. The first two articles, including this introduc-

tion, present overviews. In the next article, James Donlon, the Director of National AI Research Institutes Program at NSF, discusses the significance of the program in guiding the US national research strategy on AI as well as the future of the program.

The remaining articles describe the 18 AI Institutes. As Figure 2 illustrates, the second group consists of articles on three Institutes (Athena, ICICLE, and AI-EDGE) that develop novel cyberinfrastructures for AI, the third group contains articles describing four Institutes (IFML, AI4OPT, Dynamics AI, and TILOS) that investigate new techniques in machine learning and optimization, and the fourth group includes articles reporting on three Institutes (iSAT, Engage AI, and AI-ALOE) that explore AI for human learning and education. The next group contains articles presenting four AI Institutes (AIFARMS, AIFS, AIIRA, and AgAID) that investigate AI for food and agriculture, and the final group consists of articles describing four Institutes (AI2ES, IAIFI, MMLI, and AI-CARING) that explore AI for various aspects of science and society. In each group, the early articles describe more mature AI Institutes in the first cohort and the later articles report on the Institutes in the second cohort (see Figure 1). For example, in the Learning and Education group, while the first article describes the iSAT Institute that was launched in 2020 and belongs to the first cohort, the next two articles report on EngageAI and AI-ALOE that were launched in 2021 with the second cohort of AI Institutes.

In 2023, NSF launched another 7 AI Institutes briefly summarized in Donlon and Goel (2023). The 25 AI

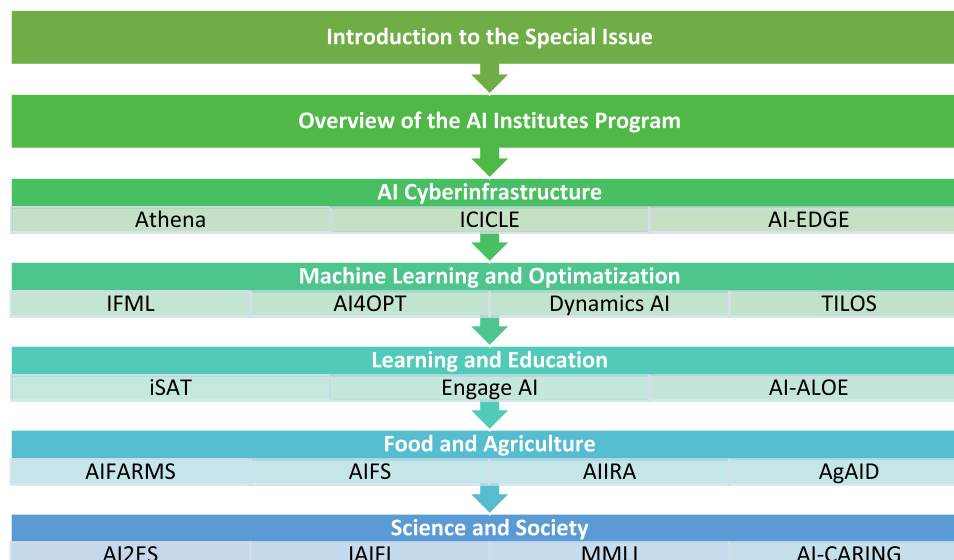
Launched in late 2020: AI2ES, AIFARMS, AIFI, AIFS, iSAT, MMLI, TILOS

Launched in late 2021: AI4OPT, AI-ALOE, AI-CARING, AI-EDGE, AgAID, AIIRA, Athena, Dynamics AI, EngageAI, ICICLE, IFML

**FIGURE 1** The first two cohorts of the AI Institutes.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AI Magazine* published by Wiley Periodicals LLC on behalf of the Association for the Advancement of Artificial Intelligence.




**FIGURE 2** Organization of the articles in this special issue.

Institutes together demonstrate that when its power is harnessed carefully, AI can be, should be, and in fact, *is* a force for social good.

#### CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

Ashok Goel   
 Chaohua Ou 

*Georgia Institute of Technology, Atlanta, Georgia, USA*

#### Correspondence

Ashok Goel, Georgia Institute of Technology, Atlanta,  
 GA, USA.  
 Email: [ashok.goel@cc.gatech.edu](mailto:ashok.goel@cc.gatech.edu)

#### ORCID

Ashok Goel  <https://orcid.org/0000-0003-4043-0614>  
 Chaohua Ou  <https://orcid.org/0000-0002-3065-2021>

#### REFERENCE

Donlon J., and A. Goel. 2023. "Looking back, Looking Ahead: Strategic Initiatives in AI and NSF's AI Institutes Program." *AI Magazine* 44(3): 345–48.

#### AUTHOR BIOGRAPHIES

**Ashok Goel** is a Professor of Computer Science at Georgia Tech and the PI and Executive Director of AI-ALOE.

**Dr. Chaohua Ou** is Assistant Director of Special Projects and Educational Initiatives for the Center for Teaching and Learning and the Managing Director of AI ALOE.



## SPECIAL TOPIC ARTICLE

# The National Artificial Intelligence Research Institutes program and its significance to a prosperous future

James J. Donlon 

National Science Foundation, Alexandria, Virginia, USA

## Correspondence

James J. Donlon, National Science Foundation, Alexandria, VA 22314, USA.  
Email: [jdonlon@nsf.gov](mailto:jdonlon@nsf.gov)

## Abstract

The U.S. National Artificial Intelligence (AI) Research Institutes program is introduced, and its significance is discussed relative to the guiding national AI research and development strategy. The future of the program is also discussed, including, the strategic priorities guiding the potential for new AI Institutes of the future, initiatives for building a broader ecosystem to connect Institutes into a strongly interconnected network, and the building of new AI capacity and fostering partnerships in minority-serving institutions.

## INTRODUCTION TO THE AI INSTITUTES PROGRAM

The National Artificial Intelligence (AI) Research Institutes is the keystone program implementing the U.S. federal government research and development (R&D) strategy to advance a cohesive approach to studying AI-related opportunities and risks. Launched in 2019, the program has established 25 AI Institutes, each a founding investment of up to 20 million dollars over 5 years. Together these Institutes represent the collaborative efforts of over 100 funded organizations and over 680 professionals. Each Institute is an interdisciplinary collaboration that advances AI knowledge and methods, and builds new platforms for AI research infrastructure and innovation. The Institutes pursue this central goal in a use-inspired research framework—situating their research in critical domains whose needs are not addressed sufficiently by applications of existing approaches. The Institutes pursue new advances in AI motivated by those challenges, while at the same time, use-inspired research situates the work in broader, societally relevant research opportunities.

Considering the high expectations put on AI Institutes and their unprecedented opportunity to form nexus points

for AI leadership and collaboration, the world is closely following their progress. A recent illustration of the program's potential for inspiring needed strategic dialogue about AI and our future is a Congressional Showcase hosted by the co-chairs of the U.S. Senate AI Caucus on Capitol Hill in September of 2023. At this event, the leadership from all 25 AI Institutes engaged in discussions and demonstrations that highlight the power of these larger-scale investments focused on *convergence research*, research driven by a specific and compelling problem carried out through deep integration across disciplines. It was clear from that engagement that the AI Institutes are proving to be critical to both advancing new state-of-the-art AI and tackling the biggest challenges we face, including in climate, agriculture, energy, health, and information security, while promoting responsible innovation that protects people's safety and rights. Institutes also demonstrated the effectiveness and reach of their many ambitious programs to actively build the next generation of AI talent that the world will need for an increasingly AI-powered future. The event illustrated the unique strategic role of federally funded investments in AI R&D and the essential value these activities can provide in a holistic consideration of the innovation and policy environment needed for a prosperous future.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.



In September 2023, co-chairs of the Senate AI Caucus, Senators Martin Heinrich (D-NM, not pictured), and Mike Rounds (R-SD) (left), hosted all 25 AI Institutes in a Congressional Showcase on Capitol Hill. Here (continuing from left) NSF Director Sethuraman Panchanathan joins a discussion between Senator Rounds and the AIFS Director Ilias Tagkopoulou, and Managing Director Steve Brown. (Photo credit: Charlotte Geary/NSF)<sup>2</sup>

The U.S. National Science Foundation and its partners are delighted that *AI Magazine* has dedicated this special issue to highlighting the activities and outcomes of the first 18 AI Institutes, which were funded in response to focus areas identified in several issuances of the program solicitation, in groups sometimes referred to as “cohorts”. Now that the cohorts funded under the first two (2020 and 2021) cohorts are well established, we expect that the community will be eager to read about the plans, achievements, and lessons of these efforts. The Institute articles in this issue will provide that update to the community in the words of the Institutes themselves. In addition to these longer-established Institutes, a new cohort of seven additional Institutes was announced in Spring of 2023. Those new awards were introduced to the readers in a column appearing in the Fall 2023 issue (Donlon & Goel, 2023). Later we will describe the anticipated growth of the program for even more Institutes to be established in 2024 and 2025.

## AI INSTITUTES AND U.S. AI RESEARCH AND DEVELOPMENT STRATEGY

This issue is an effective backdrop against which to reflect on the importance of this program to the future of AI and its impact on society. The U.S. strategy in artificial intelligence is grounded in the National Artificial Intelligence Research and Development Strategic Plan, which was first published in 2016 and updated in 2019 and 2023 (<https://www.nitrd.gov/national-artificial-intelligence-research-and-development-strategic-plan-2023-update/>). The most recent plan outlines nine elements of strategic objectives for advancing AI R&D in the U.S., as well as guiding principles and implementation actions. The AI Institutes program contributes to each of these nine strategies. Indeed, each Institute contributes in some way to many of the objectives reflected in that document. In this section is a recap of the nine strategies in that plan,





along with highlights of how the Institutes individually and collectively lead in the pursuit of those strategies.

### Strategy 1: Make long-term investments in fundamental and responsible AI research

A central rationale for the AI Institutes program is a recognition of the need for larger scale investments over longer periods of time than has been typical in either traditional federally funded research or in corporate R&D projects. At 20 million dollars apiece, each is a large initial investment, yet Institutes are encouraged to establish themselves as leaders in their fields and conduct sustainability planning for viability beyond their establishing 5 years. The program also establishes longer-term support to AI research by soliciting and awarding Institutes in successive rounds of funding opportunities guided by contemporary priorities in AI. As of this printing, we have published the fourth solicitation for AI Institutes. By renewing funding opportunities guided by shared priorities, NSF and partners commit to a continually evaluated and renewed long-term mechanism for transformative AI.

Current Institutes exemplify the unique opportunity this moment presents to AI R&D. The national strategy points to a number of goals and benefits behind this strategy. Some Institutes take advantage of this timescale to focus longer term on the understanding of the theoretical capabilities and limitations of AI. *The AI Institute for Foundations of Machine Learning (IFML, page \_\_\_)* investigates the foundations of machine learning to impact the design of practical AI systems, while the recently funded *AI Institute for Artificial and Natural Intelligence (ARNI)* connects the revolution in our understanding of the brain to foundational progress in AI.

These longer-term investments also allow for more fundamental discoveries in other areas of science and engineering. The *Institute for Artificial Intelligence and Fundamental Interactions (IAIFI, page \_\_\_)* seeks physics advances in the understanding of fundamental interactions from the smallest to the largest scales, both advancing and utilizing innovative methods in AI built upon physics principles. The *Molecule Maker Lab Institute (MMLI, page \_\_\_)* combines AI and chemists in organic synthesis to create frontier AI tools, dynamic open access databases, and fast and broadly accessible small-molecule manufacturing and discovery platforms for AI-enabled synthesis planning, catalyst development, molecule manufacturing, and molecule discovery.

This first element of the national strategy specifically calls out the potential for longer research timeframes to lead to AI systems for simulations across real and virtual environments. In virtual environments, the *AI Institute*

*for Engaged Learning (Engage AI, page \_\_\_)* is developing AI for narrative-based learning. Several Institutes investigate the role of AI in digital twins. A prominent example is the *AI Institute for Resilient Agriculture (AIIRA, page \_\_\_)*, constructing new AI-driven, predictive digital twins for modeling agriculture systems at plant, field, and farm scale to increase the resiliency of the nation's agricultural systems.

### Strategy 2: Develop effective methods for human-AI collaboration

Many AI Institutes focus on effective methods for human-AI collaboration, focusing on aspects such as the science of human-AI teaming, cultivating trust in human-AI interactions, and pursuing greater understanding of the dynamics of such hybrid systems. Many Institutes have human-AI interaction as a significant feature of research and system design. An Institute that responded specifically to a call for proposals in human-AI interaction and collaboration is the *AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups (AI CARING, page \_\_\_)*. This Institute combines personalized longitudinal interaction, robust multiagent coordination, and principles of socially conscious and trustworthy AI into a new paradigm for AI interactions and collaborations in care networks for aging adults, their caregivers, and healthcare providers, especially in situations involving mild cognitive impairment. This Institute seeks to transform paradigms of AI interactions and collaborations through adaptation to both individual users and groups of users, modeling behavior, cooperation strategies, communication, and social norms.

Several Institutes with an emphasis on student learning have significant emphasis in this strategic area as well. One is the *AI Institute for Adult Learning and Online Education (AI-ALOE, page \_\_\_)*, in which researchers address human-AI collaboration in AI assistants capable of machine teaching, self-explanation, and guiding interactions via an implementation of a joint “theory of mind” between users and the system. Another, the *AI Institute for Inclusive and Intelligent Technologies for Education (INVITE)* investigates intelligent K-12 STEM learning environments designed to assess and promote noncognitive skills known to underlie effective learning, including persistence, academic resilience, and collaboration.

### Strategy 3: Understand and address the ethical, legal, and societal implications of AI

With significant methodological progress and the rapid deployment of impressive capabilities, AI has entered the

popular discussion, with significant emphasis on not only the significant opportunities this promises, but also with attention to the risks that can negatively impact individuals, groups, organizations, communities, society, the environment, and the planet. The AI Institutes address this risk directly in their strategies. Each develops and updates a detailed Ethics Plan to both govern their organizational operations as well as guide research in understanding and mitigating social and ethical risks of AI.

Two Institutes are chartered with a specific focus on the AI principles, methods, and strategies that lead to systems that are worthy of trust and foster appropriately calibrated public trust through effective engagement. The *AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES, page \_\_)* focuses on the creation of new trustworthy AI methods, novel approaches to understanding trust between humans and AI, and theories underlying communication about user trust and risk perceptions, applying these insights to critical decision making about environmental science and extreme weather prediction. Another Institute focused on trustworthy AI is the recently established *Institute for Trustworthy AI in Law & Society (TRAILS)*. This Institute integrates consideration of participation, technology and governance during the design, development, deployment, and oversight of AI systems to investigate what trust in AI looks like, how to create technical AI solutions that build trust, and which policy models are effective in sustaining trust.

Several Institutes have as a primary focus the use of AI to address ethical, legal, and societal issues in critical application areas. Among them, the *Institute for Agricultural AI for Transforming Workforce and Decision Support (AgAID, page \_\_)* pursues long-term human-AI collaboration via multistakeholder partnership between AI designers, agriculture stakeholders, and others to produce sustained agricultural productivity to meet future food demands. AgAID's partnerships seek to transform the way AI systems are built for complex societal problems in the real world. Another, the *National AI Institute for Exceptional Education (AI4ExceptionalEd)*, conducts sociotechnical system design in the intersection of AI and learning science to improve educational outcomes for children with speech and language related challenges. In the *AI Institute for Societal Decision Making (AI-SDM)*, ethical, legal, and societal issues are central to the development of human-centric AI that enables effective, agile, resource-efficient, and trustworthy decision-making in uncertain and dynamic situations arising in disaster management and public health.

## Strategy 4: Ensure the safety and security of AI systems

All Institutes must be attentive to safety (mitigating against a system producing new harm) and security (monitoring a system's integrity) as appropriate to research. Several Institutes are chartered with a primary focus on the security and safety of critical systems. The *AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE, page \_\_)* exploits the synergies between networking and AI to design the next generation of edge networks (6G and beyond) to be highly efficient, reliable, robust, and secure by way of a distributed intelligence plane to make networks self-healing, adaptive, and self-optimized. Another Institute, the *AI Institute for Edge Computing Leveraging Next Generation Networks (Athena, page \_\_)* focuses on AI for the functionality, efficiency, and trustworthiness of next generation networks to meet the demands placed on them by the extreme heterogeneity and volume of data generated by the internet-of-things and wireless devices, as well as the design of next-generation edge datacenters with high efficiency, availability, and security for cloud-based mobile networks. Finally, an Institute from the most recent cohort, the *AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION)*, takes an agent-based approach to protecting mission-critical systems against sophisticated, ever-changing security threats, incorporating knowledge representation, logic reasoning, and learning to model and reason about complex real-world security operations.

## Strategy 5: Develop shared public datasets and environments for AI training and testing

AI Institutes will increasingly become a rich source of datasets, large-scale and specialized AI computing and hardware resources; and open-source software libraries and toolkits. Two Institutes are responding to specific mandates in this arena. The *AI Institute for Dynamic Systems (Dynamics AI, page \_\_)* develops advanced machine learning tools for controlling complex physical systems by discovering physically interpretable and physics-constrained data-driven models through optimal sensor selection and placement. This Institute develops and releases a wealth of free and open materials for broad use by students, researchers, and the public. This public release of software and data is part of an explicit commitment from this Institute to promote open-science and reproducible research, and to encourage researchers to use benchmark problems.



Another Institute promotes the use of broadly available and scalable general-purpose AI systems and infrastructure. The *AI Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment (ICICLE, page \_\_\_)* seeks to build an integrated plug-and-play national AI cyberinfrastructure to make AI both accessible and secure for everyone and drive the democratization of AI in the larger society.

### Strategy 6: Measure and evaluate AI systems through standards and benchmarks

The recently established *Institute for Trustworthy AI in Law & Society (TRAILS)* promotes the development of AI systems that can earn the public's trust through broader engagement in the AI ecosystem through participatory design, methods and metrics for evaluating AI systems, study of how people make sense of AI systems, and the role of governance and policy in fostering trust in AI systems and incentivizing broader participation, accountability, and inclusiveness in the design and deployment of AI. Through their unique partnership with the National Institute of Standards and Technology (NIST), TRAILS will contribute to the national discussion of standards and benchmarks, the needed underpinnings in foundational research, and the participation of the broader AI research community in the establishment of such practices.

### Strategy 7: Better understand the national AI R&D workforce needs

Every AI Institute has a robust portfolio of education and workforce development activities for building the next generation of AI talent. Each Institute leverages the visionary nature of their research to drive new and innovative education and development at multiple levels including K-12 education, undergraduates, graduate students, and postdoctoral researchers. Institutes also contribute to the development of AI educators and practitioners in the professional setting. The educational program at the *AI Research Institute for Advances in Optimization (AI4OPT, page \_\_\_)* has as its goal to “democratize access to AI education, AI research, and the AI workforce, bridging the gap in opportunities that exist for multiple population segments,” starting with high school education and continuing with community colleges and universities. AI4OPT reaches high school students through popular and engaging summer camps in computational and data science hosted at multiple Institute campuses. The Institute's university-level training programs focus on a faculty training program

that reaches professors at Historically Black Colleges and Universities (HBCUs) and other minority-serving institutions. in at least five states with courses in AI and course design to support the establishment of new local AI education programs at their institutions. These selected programs are representative of the commitment of this and the other Institutes in building the next generation of AI talent.

The agricultural sector is an example of a significant opportunity to address workforce needs including automation as well as training and retraining. *The AI Institute for Next Generation Food Systems (AIFS, page \_\_\_)* aims to meet growing demands in our food supply by increasing efficiencies using AI and bioinformatics spanning the entire system from growing crops through consumption. They are well situated to meet the diverse and multidisciplinary needs of a sustainable food system. In the *Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability (AIFARMS, page \_\_\_)* Institute, focus areas include autonomous farming to alleviate the agricultural labor crisis, labor optimization for managing livestock with the use of computer vision, and machine learning to monitor livestock health and behavior and human-animal interactions, enabling small teams of skilled managers to achieve better outcomes in animal health and welfare with reduced labor requirements.

Another way the national strategy urges attention to workforce needs is through a focus on regional expertise and how such understanding intersects with research pertaining to those regions. The *AI Institute for Climate–Land Interactions, Mitigation, Adaptation, Tradeoffs and Economy (AI-CLIMATE)* has as a key goal the lifting of rural economies as well as the expansion and diversification of the AI workforce. These goals are met in partnership with regional stakeholders and organizations, as the Institute pursues new AI methods to lower the cost of and improve accounting for carbon in farms and forests to empower carbon markets and inform decision-making.

### Strategy 8: Expand public-private partnerships to accelerate advances in AI

This element of strategy calls for strengthening public-private partnerships to promote a vibrant and collaborative AI ecosystem. By working together, the federal government and the private sector can advance the scientific frontiers of AI and create global social and economic benefits. AI Institute partnerships foster collaboration and coordination among diverse stakeholders from government agencies, industry, academia, and nongovernmental organizations. Partnerships between the federal government and the private sector can play a vital role in



advancing AI by aligning their complementary goals, interests, and capabilities. The federal government provides strategic guidance, policy support, and long-term funding for AI research, especially for high-risk, high-reward projects that may not be pursued by the private sector. The private sector can offer cutting-edge technology, market insights, and entrepreneurial impetus for AI innovation, as well as access to data and computing resources that are essential for developing and deploying AI systems.

NSF has been joined by industry partners to make possible several of the Institutes featured in this special issue: Accenture in support of AI-ALOE; Amazon and Google in support of AI CARING; and Intel Corporation in support of AI4OPT and the *AI Institute for Learning-Enabled Optimization at Scale (TILOS, page \_\_\_)*. In addition, IBM Corporation has recently co-funded the newly established ACTION Institute.

Beyond the program's funding partners, Institutes themselves engage in their own robust portfolio of public-private partnerships, numbering over 480 alliances and counting. Such funding partners provide a unique opportunity for rapid development of mutually beneficial collaboration. For example, TILOS engages in close research collaboration with Intel to develop learning-enabled optimizations that transform chip design. TILOS and Intel have also collaborated on the integration of workforce development modules with university courses.

Public-private partnerships extend beyond the commercial sector, for example, to include partnerships bringing together federally funded research, local school districts, and nonprofit organizations. The *AI Institute for Student-AI Teaming (iSAT, page \_\_\_)* leads such a multistakeholder consortium to form trusting and reciprocal partnerships with students and teachers for transforming classrooms into more effective, engaging, and equitable learning environments.

### **Strategy 9: Establish a principled and coordinated approach to international collaboration in AI research**

This element of strategy recognizes the imperative for international partners to collaborate in pursuit of AI research and innovation, while respecting human rights and democratic values. This will lead to more innovative, diverse, and inclusive AI solutions that address global challenges and benefit all stakeholders. It will also ensure that the AI systems we develop and deploy are aligned with the values and interests of humanity as a whole. To achieve this goal, the AI Institutes program encourages international collaboration among AI researchers from different countries, cultures, and disciplines. International collaboration

can facilitate the exchange of ideas and expertise, as well as promote the sharing of data, resources, and best practices.

Several Institutes are active in pursuing their research in an international context. To accelerate this trend, NSF and USDA-NIFA have funded new international partnerships in more than a dozen AI Institutes to explore or expand international partnerships in Australia, Belgium, Chile, EU, France, Germany, India, Israel, Italy, Japan, Netherlands, New Zealand, South Korea, Spain, Switzerland, and the United Kingdom. As will be seen below in the discussion of the “network of networks”, NSF and the community are continuing to work toward the growth of even more international collaboration with Institutes through the nurturing of nascent partnerships.

### **THE FUTURE OF THE PROGRAM**

NSF and partners continue to leverage the multidisciplinary and strategic impact of the AI Institutes concept as we look ahead to adding more of these cornerstone national AI R&D investments. The current program solicitation (<https://new.nsf.gov/funding/opportunities/national-artificial-intelligence-research>) continues the program's prior structure in which proposals will respond to scientific and application foci described in funding tracks or “themes”.

Two of the themes in this round are focused on bringing together AI researchers with those in other scientific disciplines to advance AI while pursuing AI-driven scientific discovery. A theme on “AI for Astronomical Sciences” brings the AI Institutes approach to analyzing the immense volume of high-quality astronomical data coming from major astronomical facilities. This holds the promise to develop and use new AI tools for exciting new discoveries about our universe. In “AI for Discovery in Materials Research”, we invite the community to further our quest for the advancement of AI knowledge and methods that help us discover new materials with special properties that will be the building blocks of beneficial new products and technological capabilities.

The third theme calls the AI community to respond to some of the field's most central and important goals of creating robust AI that can adapt gracefully and quickly to new domains, is robust to surprise, and resists malicious manipulation. This theme, “Strengthening AI”, has three component goals: “grounding” of our AI methods (i.e., understanding the concepts those methods reason over and operate with); “instructibility” of AI (i.e., the ability of AI to change its behavior appropriately in response to feedback provided by users); and “alignment” of these technologies with our goals and values (i.e., correspondence of an AI's operations with



objective truths, societal expectations, and human intentions). These principles are more relevant than ever as society reckons with the rapid development and deployment of impressive advances such as those in generative models. The emergence of powerful commercial AI based on generative models has challenged the AI research community into new consideration of what it means to develop general AI. In addition, the broader discussion about applications of AI and their potential societal impact will require consideration of a more comprehensive range of AI methods and approaches. A beneficial use of AI will rely on the characteristics described in this new theme. These issues are highly ambitious, long-term research objectives that are suitable to be addressed at Institute scale.

## GROWING A “NETWORK OF NETWORKS”

Just as each AI Institute is itself a hub for academia, industry, and government to accelerate discovery and innovation in AI, a broader aspiration of the National AI Research Institutes Program is to connect the program community into a whole that is greater than the sum of all Institutes. Toward this end, NSF is building a broader ecosystem to connect Institutes into a “network of networks”. This approach includes the continued addition and growth of AI Institutes, expansion of AI capacity to institutions not yet significantly engaged in AI, and a community infrastructure for collaboration and communication.

Key to connecting these parts into the “network of networks” vision is a strong community-driven hub activity. The AI Institutes Virtual Organization (AIVO, <https://aiinstitutes.org/>), established in 2022, is an NSF-funded, community-led activity for coordination, collaboration, and community building around the AI Institutes program. AIVO is a program-wide focal point for enhancing engagement between the program and the public through the use of events such as public exposition of Institute activities, social media presence, and the amplification of the Institute announcements and achievements.

Notably, AIVO convenes the annual Summit of AI Institutes Leadership (SAIL) conference. SAIL is a meeting of Institutes leadership that facilitates program-wide knowledge sharing, exchange of best practices, and community input to the strategic direction of the program. In the first two SAIL conferences, Institute personnel have engaged with one another and with guests in the broader research community to discuss topics of common interest in research (e.g., generative models, robotics, AI ethics, and security), AI impact (scientific discovery, societal applications, education, and broadening participation), Institute effectiveness (e.g., team science, project

management, sustainability planning, and partnerships), and more. Co-located with SAIL-23, AIVO also hosted a public-facing outreach event to promote AI Institute activities and connect Institutes to a wider range of stakeholders in both the public and private sectors. This AI Institutes Exposition & Engagement Showcase (“AI Institutes Expo”) included exhibitions from all 25 Institutes, presentations and panel discussions in a plenary “showcase stage” setting, and ample opportunities for networking and ad hoc gatherings.

While keystone events like SAIL and the AI Institutes Expo are high value large gatherings, AIVO connects leadership and expertise across Institute boundaries in a sustained and topic-driven way through support to multi-institute “special interest groups” and funded workshops oriented on those shared topics and priorities. Past workshops have been supported in topics including Adult Learning and Workforce Training, Data Management, Ethics and Trustworthiness, Project Management, Outreach and Communications, and others.

AIVO also administers several robust programs to fund the development of new collaborations beyond the AI Institutes ecosystem. For example, AIVO allocates a portion of its NSF funding to support travel grants to increase researcher engagement at AI conferences. Another AIVO program, the International Engagement Support Program, recognizes the increasingly global relevance of AI and the impact of combining contributions from researchers worldwide. AIVO administers this partnership development program to facilitate early exchange of ideas and expertise and to encourage the growth of future international partnerships in Institutes.

## EXPANDING AI RESEARCH IN MINORITY-SERVING INSTITUTIONS

Another objective for broadening the AI Institutes network is the continued growth of a broad and diverse interdisciplinary research community for the advancement of AI and AI-powered innovation. The Expanding AI Innovation through Capacity Building and Partnerships (ExpandAI) program (<https://new.nsf.gov/funding/opportunities/expanding-ai-innovation-through-capacity-building>) aims to significantly broaden participation in AI research, education, and workforce development through capacity development projects and through partnerships within the National AI Research Institutes ecosystem. This program is a funding opportunity for qualifying minority-serving institutions (MSIs) to engage in one of two tracks. In Capacity Building Pilots, MSIs with little to no existing AI programs pursue planning and growth of new AI capabilities and early



The second annual Summit of AI Institutes Leadership (SAIL) was held in October 2023 and included a day of workshops, a two-day main conference for knowledge exchange among Institutes, and a co-located AI Institutes Showcase and Exhibition for public engagement.

exploration of potential partnerships with AI Institutes. In ExpandAI Partnerships, participating MSIs engage in larger scale collaborative projects to build their AI research and/or education programs and engage in new, mutually beneficial collaborations with AI Institutes. ExpandAI awards made by both NSF and the U.S. Department of Agriculture's National Institute of Food and Agriculture (USDA-NIFA) have already begun to expand our national network of AI capability in new and exciting ways.

NSF and partners are grateful to *AI Magazine* for this special issue showcasing the progress and achievements so far in the 18 Institutes funded under the first two cohorts. The author would also like to congratulate and acknowledge all of our AI Institutes. We are excited for the continuing growth of the AI Institutes and look forward to a prosperous future of transformational innovation and leadership from this community.

#### ACKNOWLEDGMENTS

The National Artificial Intelligence Research Institutes Program is a joint government effort and multisector initiative in the U.S. led by the National Science Foundation (NSF). NSF currently funds 20 Institutes, some of them with the support of other U.S. government agencies

and U.S. industrial partners, as will be seen throughout this issue. NSF gratefully acknowledges the financial and intellectual contributions of its funding partners in these Institutes: U.S. Department of Education (ED) Institute of Education Sciences (IES), U.S. Department of Homeland Security (DHS) Science & Technology Directorate (S&T), National Institute of Standards and Technology (NIST), Department of Defense (DOD) Office of the Under Secretary of Defense for Research and Engineering (OUSD (R&E)), Accenture, Amazon, Google, IBM Corporation, and Intel Corporation. In addition, under this program, the U.S. Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) fully funds an additional five AI Institutes. NSF also thanks the following new partners for joining in the current solicitation in the AI Institutes program: Capital One Financial Corporation and the Simons Foundation. The author thanks Dr. Michael Littman, Division Director of the NSF Division of Information and Intelligent Systems, for his leadership of NSF AI strategy and for his thoughts on the alignment of current Institute activities to the National AI R&D Strategic Plan.

#### CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.



## ORCID

James J. Donlon  <https://orcid.org/0009-0001-9846-1037>

## REFERENCES

- “AI Institutes Virtual Organization.” AIVO. Accessed July 1, 2023. <https://aiinstitutes.org/>
- Donlon, J., and A. Goel. 2023. “Looking Back, Looking Ahead: Strategic Initiatives in AI and NSF’s AI Institutes Program.” *AI Magazine* 44(3): 345–48. <https://doi.org/10.1002/aaai.12107>
- “Expanding AI Innovation through Capacity Building and Partnerships (ExpandAI).” NSF. October 17, 2022. Accessed July 1, 2023. <https://new.nsf.gov/funding/opportunities/expanding-ai-innovation-through-capacity-building>
- “National Artificial Intelligence Research and Development Strategic Plan 2023 Update.” NITRD. Posted on May 23, 2023. Accessed July 1, 2023. <https://www.nitrd.gov/national-artificial-intelligence-research-and-development-strategic-plan-2023-update/>
- “National Artificial Intelligence Research Institutes.” NSF. August 1, 2023. <https://new.nsf.gov/funding/opportunities/national-artificial-intelligence-research>
- “National Artificial Intelligence (AI) Research Institutes Accelerating Research, *Transforming Society, and Growing the American*

*Workforce.*” NSF. Accessed July 1, 2023. <https://new.nsf.gov/funding/opportunities/national-artificial-intelligence-research>

“NSF announces 7 new National Artificial Intelligence Research Institutes.” NSF. Posted on May 23, 2023. Accessed July 1, 2023. <https://new.nsf.gov/news/nsf-announces-7-new-national-artificial>

**How to cite this article:** Donlon, JJ. “The National Artificial Intelligence Research Institutes program and its significance to a prosperous future.” *AI Magazine* 45: 6–14. <https://doi.org/10.1002/aaai.12153>

## AUTHOR BIOGRAPHY

**James J. Donlon** is a Program Director at the U.S. National Science Foundation where he leads the National AI Research Institutes Program.



**SPECIAL TOPIC ARTICLE**

# Athena – The NSF AI Institute for Edge Computing

**Yiran Chen<sup>1</sup>** | **Suman Banerjee<sup>2</sup>** | **Shaundra Daily<sup>1</sup>** | **Jeffery Krolik<sup>1</sup>** |  
**Hai (Helen) Li<sup>1</sup>** | **Daniel Limbrick<sup>3</sup>** | **Miroslav Pajic<sup>1</sup>** | **Rajashi Runton<sup>1</sup>** |  
**Lin Zhong<sup>4</sup>**

<sup>1</sup>Department of Electrical and Computer Engineering, Duke University, North Carolina, USA

<sup>2</sup>Department of Computer Sciences, University of Wisconsin, Wisconsin, USA

<sup>3</sup>Electrical and Computer Engineering Department, North Carolina A&T University, North Carolina, USA

<sup>4</sup>Department of Computer Science, Yale University, Connecticut, USA

## Correspondence

Yiran Chen, Department of Electrical and Computer Engineering, Duke University, 130 Hudson Hall, Durham, NC 27708, USA.

Email: [yiran.chen@duke.edu](mailto:yiran.chen@duke.edu)

## Funding information

National AI Research Institutes program supported by NSF and DHS, Grant/Award Number: NSF-2112562; Division of Computer and Network Systems, Grant/Award Number: 2112562

## Abstract

The National Science Foundation (NSF) Artificial Intelligence (AI) Institute for Edge Computing Leveraging Next Generation Networks (Athena) seeks to foment a transformation in modern edge computing by advancing AI foundations, computing paradigms, networked computing systems, and edge services and applications from a completely new computing perspective. Led by Duke University, Athena leverages revolutionary developments in computer systems, machine learning, networked computing systems, cyber-physical systems, and sensing. Members of Athena form a multidisciplinary team from eight universities. Athena organizes its research activities under four interrelated thrusts supporting edge computing: Foundational AI, Computer Systems, Networked Computing Systems, and Services and Applications, which constitute an ambitious and comprehensive research agenda. The research tasks of Athena will focus on developing AI-driven next-generation technologies for edge computing and new algorithmic and practical foundations of AI and evaluating the research outcomes through a combination of analytical, experimental, and empirical instruments, especially with target use-inspired research. The researchers of Athena demonstrate a cohesive effort by synergistically integrating the research outcomes from the four thrusts into three pillars: Edge Computing AI Systems, Collaborative Extended Reality (XR), and Situational Awareness and Autonomy. Athena is committed to a robust and comprehensive suite of educational and workforce development endeavors alongside its domestic and international collaboration and knowledge transfer efforts with external stakeholders that include both industry and community partnerships.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.





## INTRODUCTION

Edge computing has emerged as a formidable paradigm that brings data processing and analysis closer to where data originates, enabling real-time insights, reduced latency, and improved privacy. However, the success of contemporary artificial intelligence (AI) technologies, such as deep neural networks (DNNs), is largely dependent on the upscaling of the model size and the availability of data. This trend is at odds with the highly restricted computational resources of edge devices, which are usually constrained by size, cost, and energy consumption limitations. In addition, the management complexity of all resources of the edge computing system, such as computation, communication, and data sensing, keeps increasing. These trends necessitate the need for wide deployment of new AI solutions to infuse more efficient and powerful intelligence into edge devices and to facilitate closer connections and better coordination between the edge devices. In essence, this concept can be referred to as “*Big AI for Small Devices.*”

Athena (Athena, 2021) aims at making scientific contributions in both edge computing and AI with an emphasis on computer systems research. These include (i) novel practical and algorithmic foundations of AI to ensure the new *functionalities*, efficiency, scalability, security, privacy, and fairness of the AI solutions adopted in future edge computing systems; (ii) new edge computing systems for efficient and reliable execution of AI-operations; (iii) edge networked computing system designs and operations across the stack by leveraging a data-driven, AI-based approach; and (iv) novel use-inspired services and applications, focused on diverse cyber-physical systems that leverage the innovations of the other thrusts. The researchers of Athena work closely with external collaborators to translate research outcomes to industrial practice and policy making. The educational and outreach activities of Athena empower students and postdocs to develop their interests, build skills, and acquire knowledge about AI and computer and network systems through research experiences, industry internships, and community engagements. The Inclusive AI Initiative—one of Athena’s innovations in education and workforce development—strengthens the ethical AI competencies of all Athena members to better promote and be aware of equity and fairness in their research and the communities impacted by the institute’s research.

## ORGANIZATION AND MANAGEMENT

Athena is a multi-university and transdisciplinary AI Institute including eight academic institutions (Duke

University, Arizona State University, Massachusetts Institute of Technology, North Carolina Agricultural and Technical State University, Princeton University, University of Michigan-Ann Arbor, University of Wisconsin-Madison, and Yale University) and five industry collaborators (AT&T, Microsoft, Motorola Solutions, EdgeMicro, and 5NINES), as shown in Figure 1. The Athena team includes 28 senior researchers with diverse backgrounds from eight participating universities, three international collaborators respectively from Switzerland, Germany, and Israel, and 1 program manager. Athena also sponsors more than 60 students and PostDocs to perform research with our faculty members. All the researchers are associated with one or more research pillars (more details can be found in the section [Technical Thrusts and Pillars](#)). An external advisory board (EAB) has been created to provide advice and suggestions on the operation of Athena, including six members from both industry and academia. Athena is also working with several community and educational partners on its broader impact activities, including Cary Township, NC School of Science and Math (NCSSM), STEM Early College@NC A&T, STEM from Dance, etc. In the recently announced NSF Expand AI awards, the Athena team is working with UT San Antonio to promote research and education for under-represented groups. Athena also works closely with the Duke-led NSF Industry-University Cooperative Research Center (IUCRC) for Alternative Sustainable Intelligent Computing (ASIC, 2018) to create a consortium to foster research collaborations with industrial and government partners.

Every year Athena hosts an annual review meeting and Tech Showcase, which is open to all Athena members, students, and invited government and industry stakeholders. The meeting is composed of keynote talks, technical sessions, panels, and poster and demo sessions. All the research, educational, and other projects are also reviewed in the meeting and the feedback from internal and external reviewers is provided to the researchers and PIs.

## TECHNICAL THRUSTS AND PILLARS

The Athena’s research builds upon four interrelated thrusts: *Foundational AI*, *Computer Systems*, *Networked Computing Systems*, and *Services and Applications*, which cover multiple layers of modern edge computing systems and cross both foundational and use-inspired research. On top of the advances and breakthrough of AI technologies, the research outcomes are integrated into and demonstrated by three primary research pillars, namely, *edge computing AI systems*, *collaborative extended reality (XR)*, and *situational awareness and autonomy*.

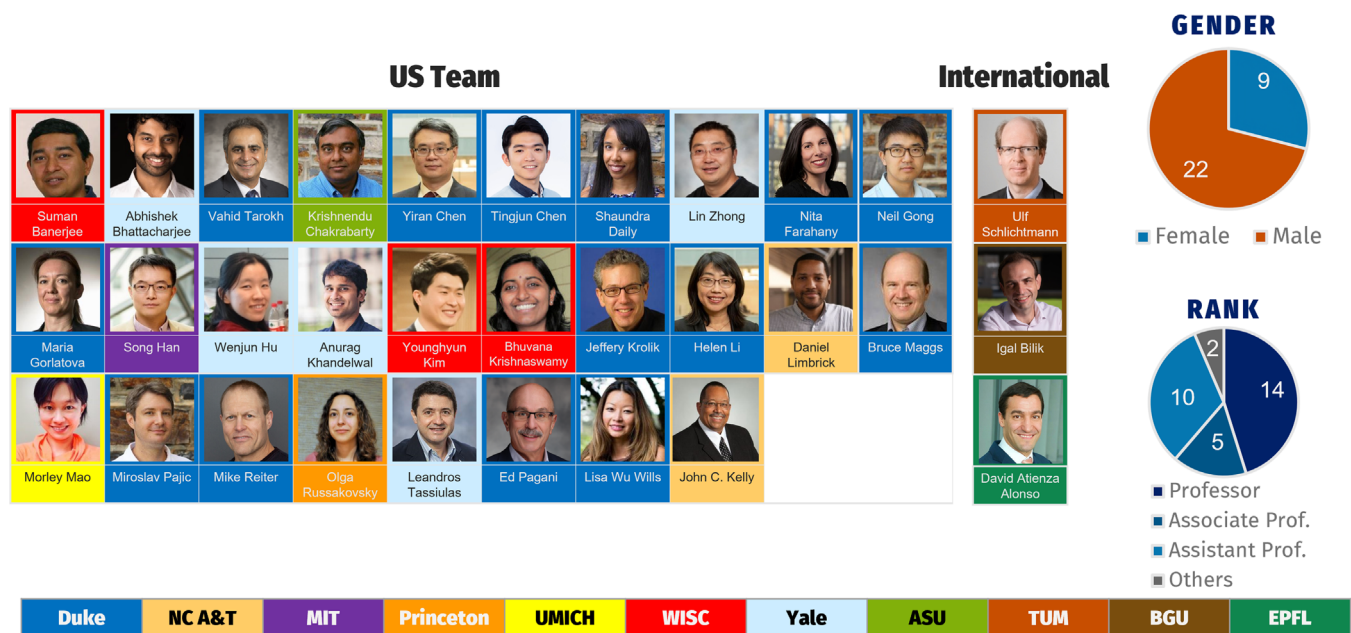


FIGURE 1 Athena team members.

## AI foundation

We envision that *functionality*, *efficiency*, *heterogeneity*, and *trustworthiness* are four major challenges in the AI research of Athena to deliver the promised next-generation edge computing systems. The unique operational requirements of AI-powered networks and computing systems demand for new functions of AI, such as the abilities to handle the unseen circumstances and to debug root reason of concerned system behaviors. Efficiency and scalability of AI models is also crucial in the target large-scale and heterogeneous edge computing system. The heterogeneity of the edge devices and the data to be processed enables great opportunities to advance the computing paradigm of AI, while also imposing many challenges on maintaining reliability and robustness of the computation. The distributed and heterogeneous architecture introduces many security, privacy, and fairness concerns of the adopted AI models.

The members of Athena's AI thrust are developing novel causality analysis methods to grant the AI models an ability to handle the unforeseen circumstances based on out-of-distribution (OOD) data analysis (Linderman et al., 2023). The team is also investigating the techniques to enable efficient deployment and execution of the AI models and automate the model design by considering the system heterogeneity (Lin et al., 2022). The new federated learning (FL) framework and computing hierarchy are being developed to combat the system heterogeneity and preserve the data integrity. Finally, the team members

are also developing defenses against security and privacy attacks with provable guarantees and exploring the societal fairness implications of the AI models and the corresponding algorithmic bias mitigation methods. All these techniques are being seamlessly integrated into the three research pillars together with the techniques developed by other thrusts.

## Edge computing AI systems

Due to the limited scale hindering economies of scale benefits, naively architected edge data centers face issues of not only high operating costs but also constrained performance and flexibility in edge computing applications such as AI/ML workloads. Hence, the team of the Edge Computing AI Systems pillar of Athena proposes to leverage resource disaggregation in edge data centers where compute and memory/storage resources are physically separated into network-attached resource blades, particularly through network-centric operating systems (OSes). Based on our team members' expertise in the OS's memory management subsystem, the team focuses on key OS components and services for disaggregated edge data centers. Specifically, we looked at (a) efficient synchronization (Yu et al., 2023), and (b) near-memory accelerators (Tang et al., 2023). In addition, the ongoing research aims to achieve (c) performance isolation and fairness among edge applications, such as real-time AI inference and database workloads.

Our research activities include two major directions: (i) *Systems for AI*. We are using programmable network hardware to design and implement in-network memory and process management. Our goal is to boost resource elasticity and hardware flexibility, ultimately improving the efficiency of serving AI workloads. (ii) *AI for systems*. We are exploring system optimizations using ML algorithms, utilizing an overarching view of all inter-resource communications to establish AI-optimized system management. Our efforts toward the key directions have been demonstrated across research focuses such as (a) Efficient and Scalable Synchronization Primitive, (b) Distributed Near-memory Accelerator, and (c) Performance Isolation and Fairness. The team is putting efforts to expedite AI/ML and traditional tasks in disaggregated edge data centers through a network-centric OS design and extend the scope to other OS components and devices. A key goal is to simplify data synchronization across diverse computing resources like CPUs, GPUs, and accelerators. Our ambition is to enable universal data sharing using a unified virtual address across accelerators, enhancing development convenience and data center modularity.

## Collaborative Extended Reality (XR)

XR is an application domain of Athena where the goal is to create an end-to-end system where multiple augmented reality users enter an environment and share data with each other to collaboratively solve specific problems. This includes sharing 2D/3D point-cloud style data to create 3D maps. It can also focus on allowing cloud/edge supported object detection and tracking in various challenging environments. Collaborating with the Situational Awareness and Autonomy Pillar (see section [Situational Awareness and Autonomy](#)) to provide an end-to-end sensing service, which involves gathering data from wide-area sensors, collecting them at edge sites, suitably compressing them, and then leveraging necessary decoding capabilities in the cloud. Figure 2 shows a demo that uses a VR headset to simulate brain surgery, which showcased in the NSF AI Hill day on September 19, 2023.

The pillar focuses on challenging systems problems to deliver a seamless end-user experience to users of XR systems. The activities of the project have different systems components that primarily run on end devices, and on edge systems and edge servers, with coordination and support of a data center's computational power. The activities are split into the following categories: (i) Scalable object detection and learning, (ii) Continuous localization and mapping, (iii) Energy efficiency, (iv) FL, (v) Scalable data compression, (vi) Edge dimensioning, and (vii) Multi-modal interfaces. The ultimate goal of these projects



**FIGURE 2** National Science Foundation (NSF) Director Sethuraman Panchanathan tried a demo of using a VR headset to simulate brain surgery, which was developed by Athena PI Maria Gorlatova in NSF Artificial Intelligence (AI) Hill Day.

is to create an end-to-end system demonstrating showcasing different aspects of the collaboration in XR. It will showcase an indoor environment where a group of users can collaboratively create 3-dimensional maps and floor plans using data gathered by their headsets, be able to efficiently detect different object types in the environment, leave markers for each, and collaboratively solve various tasks assigned. An example use of this capability is for public safety agency personnel, for example, firefighters, attempting to navigate an unknown building to conduct search and rescue operations.

## Situational awareness and autonomy

Over the past decade, the scientific foundations of model-based design have established rigorous design and analysis approaches for safety-critical systems. However, in the context of data-driven, AI-enabled systems, such methods still have limited use in the real-world scenarios. Consequently, our goal is to provide assured, robust, and resilient services for autonomous systems at the edge. To achieve

this, we focus on two critical, interdependent components: edge-based trusted autonomy and situational-awareness, that provide strong safety and performance (i.e., functionality) guarantees both at design- and run-time, as the system, its functional requirements and environment evolve. In particular, we focus on two applications that will be used to illustrate the promise and effectiveness of the next generation of edge-based autonomy: (1) multi-drone search and target tracking in contested and uncertain environments, and (2) collaborative sensing in Connected and Automated/Autonomous Vehicles (CAVs).

Our approaches to achieve the above goals include the following. We have been developing trustworthy and adversarially robust models for the fusion of asynchronous heterogeneous sensing used for perception and situational awareness in autonomous systems (e.g., cameras, LiDAR, radar). To deal with contested environments, realistic threat models and security analyses for such scenarios have been introduced (e.g., Hallyburton et al., 2022) demonstrating the vulnerability of existing perception models). Further, to exploit on-demand computation availability, we have been developing autonomous services capable of employing a combination of centralized (e.g., on an edge-server) and decentralized data aggregation, as well as the development of secured FL models (e.g., Sun et al., 2022). Similarly, we have focused on the design of adversarially robust decision-making policies. For example, we have recently extended the standard adversarial training approach for robust reinforcement learning agents, based on two-player max-min games; specifically, two-player games were extended by introducing an adversarial herd (i.e., a group of adversaries), in order to reduce the difficulty of the inner optimization problem, and the potential over-pessimism due to selection of an adversary set that may include unlikely scenarios (Dong et al., 2023).

## EDUCATION AND WORKFORCE DEVELOPMENT

In an era driven by technological advancements, Athena is dedicated to cultivating a diverse workforce for the future. Athena actively tailors educational innovations to cater to students from kindergarten to postdoctoral researchers, emphasizing the broadening participation of underrepresented groups in science and engineering.

Our Embedded STEM Labs is a novel mechanism for supporting community-based k-12 education outreach. Through collaborative partnerships with nonprofit organizations, Athena develops programs that enhance the STEM educational capacity of nonprofit staff, creates culturally relevant and design-based STEM curricula, and provides university faculty and students with community

engagement opportunities. By summer of 2023, we will have reached over 450 students with 96% of our k-12 participants from minoritized groups, ensuring our institute contributes to broadening participation in AI. Recognizing the significance of undergraduate involvement, Athena offers academic year and summer research opportunities. Moreover, Athena organizes an annual graduate bootcamp with virtual and online components to equip students with essential skills for the application process, statement writing, and selecting the right educational path. This holistic approach ensures that students are empowered to pursue graduate studies.

Athena's team has also made remarkable strides in the realm of AI ethics education. By developing multiple modules that can be utilized broadly, they have bolstered the understanding of ethical considerations in the field. Furthermore, a newly approved undergraduate course titled "Let's Talk About Digital You" will enable a broader reach and serve as a blueprint for ethical technology education for higher educational institutions.

To foster advancements in autonomous vehicles (AV), teams have created AVstack—a groundbreaking, open-source software platform. AVstack facilitates the development, evaluation, and analysis of AI-based modules for AV autonomy and situational awareness. The teams' ongoing efforts focus on producing course materials and laboratory modules that utilize AVStack and open-source datasets/simulators to educate students on AI-based AV design. In the graduate studies realm, faculty have actively engaged master's and PhD students in Athena-related research.

Athena exemplifies a visionary approach to shaping the next generation of ethical AI talent. Through our comprehensive educational programs, research opportunities, and impactful partnerships, Athena is at the forefront of shaping a future where AI is leveraged responsibly and for the benefit of all.

## COMMUNITY BUILDING AND BROADER IMPACT

Athena is collaborating with three foreign universities: École polytechnique fédérale de Lausanne (EPFL), Technische Universität München (TUM), and Ben-Gurion University of the Negev (BGU). The PI/co-PIs from both Athena and the international institutions have collaborated on various activities including joint research projects, for example, FL framework for biomedical applications (EPFL + Duke), machine learning accelerator design based on emerging nanoscale devices (TUM + Duke), and new machine learning methods for integrating situational awareness radar and communications for autonomous

vehicles (BGU + Duke). Besides collaborative research, Athena continues to work with international collaborative institutions on curriculum developments, graduate student and young faculty training and mentoring, personnel exchange and visit, workshops, and summer camps. Athena is also participating in the Next Generation Internet (NGI) Enrichers Program, an initiative funded by the European Union, where Athena will be hosting the selected “Fellow” for a 3–6-month fellowship at Duke University.

The Athena Monthly Seminar Series has successfully entered its third year. Speakers range from internationally acclaimed scientists to Athena graduate students. Of the audience members, typically, 5% are government representatives, 10% are industry members, and 85% are students or faculty from both domestic and international universities. These seminars are open to the public and are advertised widely on our social media pages.

Athena is also committed to technology transfer and has been actively working with industrial partners to apply the research outcomes to practical problems. So far, the team members of Athena have filled 5 patent disclosures in the applications of system safety, biometric data analysis, autonomous driving, and healthcare. A company of the team members has recently been acquired by NVIDIA.

## CONCLUSION

Sponsored by the National Science Foundation and Department of Homeland Security, Athena will deliver the key technologies for next-generation edge computing systems powered by AI with unprecedented efficiency, reliability, and performance. Athena is now serving as the nexus point of community, facilitating the ecosystem of the emerging technologies, and cultivating diverse next-generation technical leaders having the values of ethics and fairness. The success of Athena will disrupt the future edge computing industry, create new business model and entrepreneurial opportunities, and transform the competition model of future edge computing industry and research.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National AI Research Institutes program supported by NSF and DHS under the NSF Award No. 2112562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

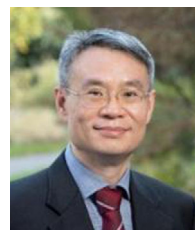
Yiran Chen  <https://orcid.org/0000-0002-1486-8412>

## REFERENCES

- Athena. <https://athena.duke.edu/>, 2021.
- ASIC. <http://asic.pratt.duke.edu/>, 2018.
- Linderman, R., J. Zhang, N. Inkawhich, H. Li, and Y. Chen. “Fine-Grain Inference on Out-of-Distribution Data with Hierarchical Classification.” Paper presented at Conference on Lifelong Learning Agents (CoLLAs), August, 2023.
- Lin, J., L. Zhu, W.-M. Chen, W.-C. Wang, C. Gan, and S. Han. 2022. “On-Device Training Under 256 KB Memory.” Paper presented at Annual Conference on Neural Information Processing Systems (NeurIPS), December, 2022.
- Yu, Y., S.-S. Lee, A. Khandelwal, and L. Zhong. 2023. “GCS: Generalized Cache-Coherence For Efficient Synchronization.” arXiv preprint arXiv:2301.02576.
- Tang, Y., S.-S. Lee, and A. Khandelwal. 2023. “CHASE: Accelerating Distributed Pointer-Traversals on Disaggregated Memory.” arXiv preprint arXiv:2305.02388.
- Hallyburton, S., Y. Liu, Y. Cao, Z. M. Mao, and M. Pajic. “Security Analysis of Camera-LiDAR Fusion Against Black-Box Attacks on Autonomous Vehicles.” Paper presented at 31st USENIX Security Symposium (USENIX SECURITY), August, 2022, 1903–20.
- Sun, J., A. Li, L. Duan, S. Alam, X. Deng, X. Guo, H. Wang, M. Gorlatova, M. Zhang, H. Li, and Y. Chen. “FedSEA: A Semi-Asynchronous Federated Learning Framework for Extremely Heterogeneous Devices.” Paper presented at The ACM Conference on Embedded Networked Sensor Systems (SenSys), November, 2022, 106–19.
- Dong, J., H. L. Hsu, Q. Gao, V. Torokh, and M. Pajic. 2023. “Robust Reinforcement Learning through Efficient Adversarial Herding.” arXiv preprint arXiv:2306.07408.

**How to cite this article:** Chen, Y., S. Banerjee, S. Daily, J. Krolik, H.(H.) Li, D. Limbrick, M. Pajic, R. Runton, and L. Zhong. 2024. “Athena – The NSF AI Institute for Edge Computing.” *AI Magazine* 45: 15–21. <https://doi.org/10.1002/aaai.12147>

## AUTHOR BIOGRAPHIES



**Yiran Chen** is the John Cocke Distinguished Professor of Electrical and Computer Engineering at Duke University. He is the Principal Investigator (PI) and the Director of Athena. His research interests include emerging memory technologies, machine learning acceleration, neuromorphic computing, federated learning, and mobile computing. He is the Fellow of AAAS, ACM, IEEE and NAI.



**Suman Banerjee** is the David J. DeWitt Professor of the Department of Computer Sciences at the University of Wisconsin. He is the co-PI and the Collaborative XR Pillar Lead of Athena. His research interests focus on networking and distributed systems with a primary focus on wireless and mobile networking. He is the Fellow of ACM and IEEE.



**Shaundra Daily** is the Cue Family Professor of the Practice of Electrical and Computer Engineering at Duke University. She is the Education and Workforce Development Lead of Athena. Her research interests include design, implementation, and evaluation of technologies, programs, and curricula to support diversity, equity, and inclusion in STEM fields.



**Jeffery Krolík** is the Professor of Electrical and Computer Engineering at Duke University. He is the Managing Director of Athena. His research interests include statistical signal processing for surveillance radars and microwave remote sensing, active and passive sonar, and medical imaging. He is the Fellow of IEEE.



**Hai (Helen) Li** is the Clare Boothe Luce Professor of Electrical and Computer Engineering at Duke University. She is co-PI and the AI Foundation Lead of Athena. Her research interests include neuromorphic computing systems, machine learning acceleration and trustworthy AI, emerging memory technologies, and lower power circuits and systems. She is the Fellow of ACM and IEEE.



**Daniel Limbrick** is the Associate Professor of the Electrical and Computer Engineering Department at the North Carolina Agricultural and Technical State University. He is the Broader Participation Lead of Athena. His research interests include reliability and scalability of integrated circuits through logic and physical synthesis.



**Miroslav Pajic** is the Dickinson Family Associate Professor of Electrical and Computer Engineering at Duke University. He is co-PI and the Situational Awareness Pillar Lead of Athena. His research interests focus on design and analysis of cyber-physical systems with varying levels of autonomy and human interaction at the intersection of (more traditional) areas of embedded systems, AI, etc.



**Rajashi Runton** is the Senior Program Coordinator of the Department of Electrical and Computer Engineering at Duke University. She is the Program Manager of Athena. She received BSEE from Massachusetts Institute of Technology and MBA from Arizona State University. She has been a product engineer at Motorola and Program Manager and Business Manager at Freescale Semiconductor.



**Lin Zhong** is the Professor of the Department of Computer Science at Yale University. He is co-PI and the Edge Computing AI Systems Pillar Lead of Athena. His research interests include software systems by embracing analog hardware, formal methods, and new systems programming languages. He is the Fellow of ACM and IEEE.



## SPECIAL TOPIC ARTICLE

# Creating intelligent cyberinfrastructure for democratizing AI

Dhableswar K. Panda<sup>1</sup> | Vipin Chaudhary<sup>2</sup> | Eric Fosler-Lussier<sup>1</sup> |  
 Raghu Machiraju<sup>1</sup> | Amit Majumdar<sup>3</sup> | Beth Plale<sup>4</sup> | Rajiv Ramnath<sup>1</sup> |  
 Ponnuswamy Sadayappan<sup>5</sup> | Neelima Savardekar<sup>1</sup> | Karen Tomko<sup>6</sup>

<sup>1</sup>The Ohio State University, Columbus, Ohio, USA

<sup>2</sup>Case Western Reserve University, Cleveland, Ohio, USA

<sup>3</sup>San Diego Supercomputer Center, Diego, California, USA

<sup>4</sup>Indiana University, Bloomington, Indiana, USA

<sup>5</sup>University of Utah, Salt Lake City, Utah, USA

<sup>6</sup>Ohio Supercomputer Center, Columbus, Ohio, USA

## Correspondence

Dhableswar K. Panda, The Ohio State University, Columbus, OH, USA.  
 Email: [panda@cse.ohio-state.edu](mailto:panda@cse.ohio-state.edu)

## Funding information

National Science Foundation; Division of Advanced Cyberinfrastructure, Grant/Award Number: OAC-2112606

## Abstract

Artificial intelligence (AI) has the potential for vast societal and economic gain; yet applications are developed in a largely ad hoc manner, lacking coherent, standardized, modular, and reusable infrastructures. The NSF-funded Intelligent CyberInfrastructure with Computational Learning in the Environment AI Institute (“ICICLE”) aims to fundamentally advance *edge-to-center*, AI-as-a-Service, achieved through intelligent cyberinfrastructure (CI) that spans the edge-cloud-HPC *computing continuum*, *plug-and-play* next-generation AI and intelligent CI services, and a commitment to design for broad accessibility and widespread benefit. This design is foundational to the institute’s commitment to democratizing AI. The institute’s CI activities are informed by three high-impact domains: *animal ecology*, *digital agriculture*, and *smart foodsheds*. The institute’s workforce development and broadening participation in computing efforts reinforce the institute’s commitment to *democratizing AI*. ICICLE seeks to serve as *the national nexus for AI and intelligent CI*, and welcomes engagement across its wide set of programs.

## INTRODUCING THE ICICLE AI INSTITUTE

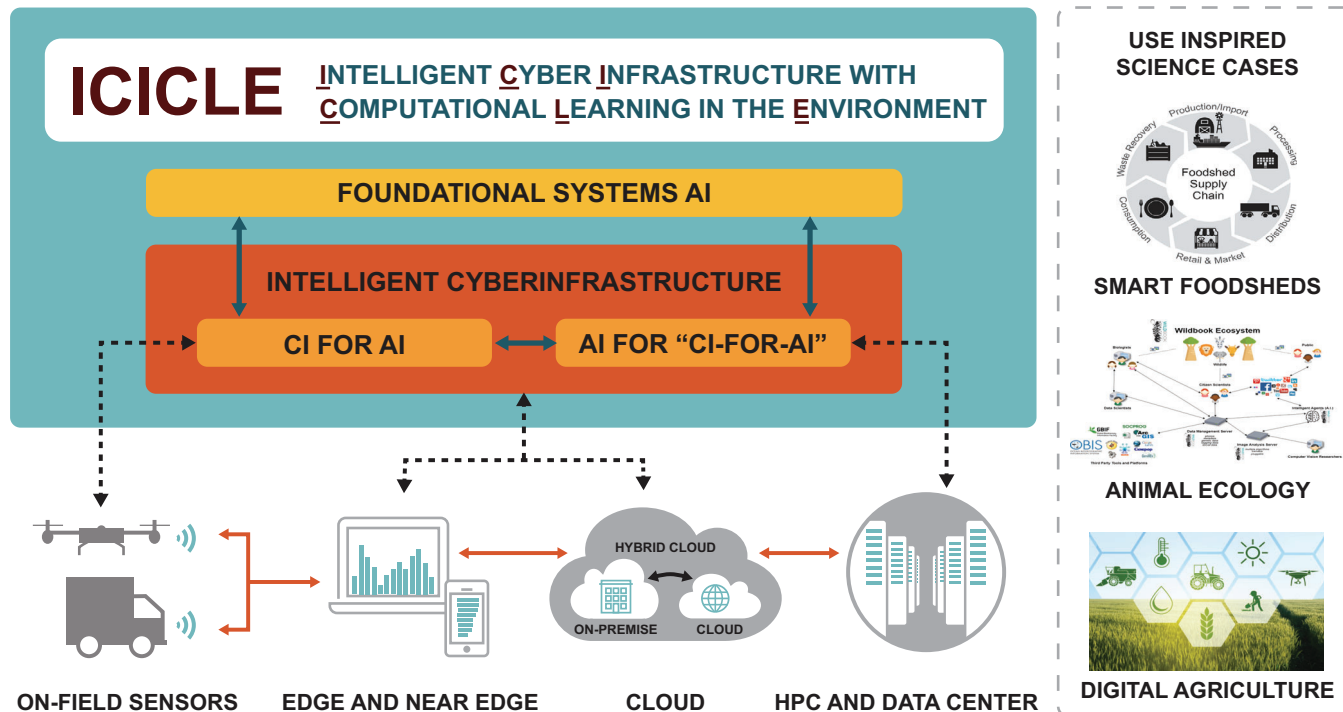
**Institute rationale:** Artificial intelligence (AI) is transforming every sector of society, from transportation and autonomous vehicles to biodiversity and wildlife conservation to food production and smart foodsheds. It will be the driving force for the next information revolution. However, there is a massive gap between available AI techniques and end-user availability to various application domains. The ad hoc development of many AI applications lack coherent, standardized, modular, and reusable infrastructures.

Successful AI solutions for one use case rarely directly generalize to other use cases, or even the same use case with a slightly different context.

Infrastructure implementation necessitates proper AI abstractions. Concomitant with generalizability challenges, as the cyberinfrastructure (CI) grows increasingly complex, end users face bewildering choices when purposing AI toward insightful analytics, modeling complex systems, or developing automation. In environment-focused settings, scientists studying natural/managed ecosystems and resource flows necessarily deal with complex dynamics represented by a wide array of public and private data

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Overview and scope of the ICICLE project.

sources. Efficiently using AI infrastructures requires navigating model choice, cloud-or-local compute decisions, privacy and data protection, and heterogeneous architectures.

**Addressing the challenge:** The **ICICLE** (*Intelligent CyberInfrastructure with Computational Learning in the Environment*) AI Institute (ICICLE 2023) addresses these challenges by building the first and foremost *edge-to-center* AI-as-a-service enterprise, advancing foundational AI research and next-gen CI for AI (*CI4AI*), while using AI techniques in building this CI (*AI4CI*). This CI supports AI democratization, extending the accessibility and benefit of AI to a wider population (Figure 1) and across the computing continuum—from traditional large-scale HPC systems and data centers to newer activity in edge- and near-edge devices as well as cloud resources.

Development and deployment of this next-gen CI is driven by *use-inspired research* in three high-impact domains: *animal ecology*, *digital agriculture*, and *smart foodsheds*. ICICLE's design extracts both common and differentiated workflows across use cases, inspiring new CI for AI and enabling new discoveries in these domains. End users and AI application developers must be able to keep data access appropriately authenticated or private, and choose where computation occurs (e.g., data centers, the cloud, or edge devices). Thus, the “E” in ICICLE—environment—reflects two interacting themes within the institute: end use cases from real-world environments utilizing AI and in expanding computational environments.

**The team:** Realizing ICICLE's vision requires a highly cohesive and interdisciplinary team of scientists, engineers, and practitioners from thirteen universities and research institutes. Team expertise spans CI (high-performance computing, networking), AI (statistical machine learning, computer vision, knowledge graphs, and conversational AI), data privacy and trust, visualization, and use-inspired science in animal ecology, digital agriculture, and smart foodsheds. Additionally, ICICLE team members have expertise in the areas of broadening participation, workforce development (WFD), knowledge transfer, and collaboration. Under the umbrella of the ICICLE Institute, these diverse, individually talented, and individually focused researchers have formed a cohesive team capable of taking up the grand challenge of solving our proposed *plug-and-play* AI vision by tackling cross-cutting problems in CI, AI, and use-inspired science.

**Building connections:** The diverse expertise in ICICLE is being harnessed to grow the next generation of scientists, scholars, and workers—from kindergarten through professionals—using an inclusive plan that is synergistic with NSF's overall WFD portfolio. ICICLE's broadening participation in computing (BPC) plan leverages both the foundations of the NSF CISE BPC program and our own institutional resources to impact WFD, particularly reaching out to women, historically disadvantaged persons, and persons with disabilities.

As a national scale AI Institute, ICICLE also leverages NSF-funded infrastructure including *Tapis* from the Texas





Advanced Computing Center, *Jetstream* from Indiana University, and *Voyager* from the San Diego Supercomputer Center, thus bringing together multiple organizations to synergistically work together as a *nexus for this collaborative effort*.

## VISION AND RESEARCH DIRECTIONS

ICICLE addresses a crucial reality: today, applying AI techniques to any specific use case is a nontrivial task for AI experts, let alone for users whose expertise lies in the domains outside AI. The vision of the ICICLE Institute is to empower users to seamlessly *plug-and-play* AI models in any given local computing environment across multiple use cases.

ICICLE aims to provide conversational and visualization-powered user interfaces where users are able to ask questions about the essential underlying context, available computational resources, and other information in a natural language much like speaking to agents such as Siri and Alexa. Next, ICICLE includes a novel realization of an AI model commons, that is, an ecosystem for producing, profiling, sharing, and distributing AI models. The commons will use existing models, exploit available unified semantic data representation provided by expansive knowledge graphs to create new models, and direct users to models most relevant to their needs. Finally, ICICLE is environment-aware: it manages AI holistically across the computing continuum from the edge and on-premise devices used for data collection to model training, curation, and storage at large HPC/cloud centers. Recognizing AI's inherent contextual sensitivity, adaptivity in ICICLE is an essential property for AI deployment with innovations for automated or human-in-the-loop adaptation to different application contexts (including tasks, environments, and user preferences) at the edge.

**Research directions:** All these foundational systems AI innovations will be powered by an intelligent CI that will provide high-performance model training, data management, edge intelligence, and wireless control and coordination for the computing continuum, which is in turn continuously improved by advanced AI techniques. Throughout, ICICLE includes cross-cutting design considerations to manage data privacy, accountability, and integrity, as well as visual analytics for explainability.

The institute itself provides an opportunity to study *AI ethics* because of its unique multimodal orientation on CI for AI, AI for CI, and on the democratization of AI. Through dialog, thought experiments, and a mapping exercise we asked “Where could issues of AI ethics arise in AI infused CI infrastructure for AI?” Through this exercise,

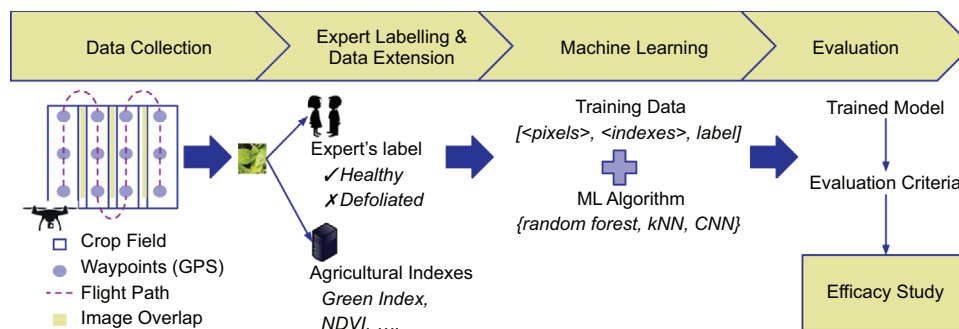
we categorized concerns as falling into six areas: democratization, fairness, accountability, trustworthiness, privacy, and the unknowns that arise from the use of the CI, (manuscript under review). In a step then towards operationalizing the project's statement of ethics, we developed a AI ethics framework. The framework, a living document, proposes a collection of guiding principles that are tailored to render ICICLE as a responsibly designed and managed CI. The guiding principles are a work in progress; they, however, must be complemented by outreach and WFD effort focused on the unique challenges of AI ethics in research infrastructure.

To empower plug-and-play AI, ICICLE must provide abstraction layers to shield users and AI and CI developers from the pervasive heterogeneities in the full AI development and application life-cycle. Such successful abstractions require the following advances in foundational systems AI: *knowledge graphs* as the knowledge backbone to provide a semantically rich abstraction for addressing data heterogeneity; a *model commons* to provide the abstraction for addressing model heterogeneity and improve discoverability, matchability, and interoperability of AI models; *adaptive AI* to enable in situ adaptation of AI models at the edge; *federated learning* to support edge-to-center across the computing continuum, decentralized, and privacy-preserving learning; and *conversational AI* to provide unified and user-friendly interfaces for human-machine interaction and improve the accessibility and usability of the entire ICICLE system. All these systems' AI research areas embrace and aim to address the full complexity and heterogeneity in the environment-bound, edge-to-center, and user-supporting scenarios targeted by ICICLE.

The ICICLE Institute brings under one roof a broad range of experts, institutions, and partners to develop next-generation AI and CI allowing for **democratized and plug-and-play AI modeling across the computing continuum** to rapidly impact salient use-inspired projects while training the next generation of AI-aware talent. Following the mantra “today's AI is tomorrow's CI,” the ICICLE Institute is integrating tenets of foundational AI into CI in a virtuous cycle, whereby AI innovations in areas such as conversational AI and federated learning become integrated into the CI, while foundational CI research supports AI innovation by improving the ability to rapidly train, deploy, and secure models and data.

## HIGHLIGHTED ACCOMPLISHMENTS

Even early in the life of the ICICLE Institute, we are starting to see synergistic accomplishments that advance our ambitious goals of combining foundational AI research



**FIGURE 2** AI inspired agriculture.

with CI ecosystem advancements. ICICLE must be a sustainable, transformational nexus of collaborations among the AI, CI, and domain sciences while at the same time BPC and developing an AI-savvy and ethically aware workforce.

Institute research outputs, outcomes, and impacts take place across several dimensions but are oriented towards contributions to application domains (animal ecology, digital agriculture, smart foodsheds).

**Animal ecology:** Two guiding applications for this area are smart camera traps that can intelligently monitor animal behavior from a fixed (typically remote) location, and drone-based observations that can monitor behavior more dynamically. Early results include the release of benchmark datasets and models, as well as the release of a reference simulation application (camera traps). Research foci include knowledge graphs, resource profiling, model training, and model commons.

Early outreach and interactions include the Ohio Department of Natural Resources, and Division of Wildlife. Collaboration with the NSF-funded HDR DIRSE Imageomics Institute has resulted in contributions to the Experiential Field Course in Kenya, focusing on animal behavior inference from drone videos.

**Digital agriculture:** This application area enables democratized access of AI technologies to digital agriculture services. Its guiding applications are aerial crop scouting, crop health modeling, and autonomous in-field machinery; this requires HPC model training and inference, edge wireless, model commons (Figure 2). Early results include released data sets and AI models on soybean crop health (Ockerman et al. 2023). The digital agriculture group has initiated collaborative outreach to the Ohio Soybean Council, AgAID, AIFARMS, Tata Consulting Services and other industry partners. We have also established the US-Indo research collaboration with Technology Innovation Hub on IoT (with IIT-Bombay).

**Smart foodsheds:** This application area envisions democratizing AI for food systems workers of the future

through access to computational workflows and tools. It strives to achieve this goal by means of a standardized ICICLE Integrated Knowledge and Learning Environment (IKLE) that serves as an entry point for food systems planners to explore their own ontologically encoded data sets and eventually mixed with public data (Tu, Wang, et al. 2023b). The group's early research results include the Persons-Projects-Organizations-Datasets ("PPOD") ontology, now publicly available on GitHub. It is developing a conversation agent and federated data capabilities. They have worked with the visual analytics research thrust on a prototype visualization platform for exploring KG and food flow data.

**AI foundations:** Within this thrust, the team is developing plug-and-play technology to allow construction of systems utilizing key AI infrastructure components, including models from model commons, knowledge graphs, federated learning and adaptive AI techniques. One example is a new holistic transfer learning methodology (Tu, Chen, et al. 2023a) that allows adaptation of pretrained models to new distributions without disrupting the original performance; when used in the smart camera traps setting, this empowers ecologists to adapt species classification models to work in new locations with few data. Similarly, information access in food system knowledge graphs will soon be driven by new adaptive conversational AI technology that provides effective few-shot in-context learning for knowledge-base question answering with large language models (Gu, Deng, and Su 2022).

**CI4AI:** Within the CI4AI thrust, early results include new top-k gradient sparsification methods and all-to-all sparse communication algorithms for high-performance model training that is 20% faster than SOTA method on BERT training with convergence (Hussain et al. 2022). The novel solution includes a new distributed scheduling algorithm for ensuring AI-adaptive, probabilistic per-packet real-time guarantees has the potential for improving network capacity by a factor of 5–20x. Finally, an intelligent resource provisioner with reinforcement learning has



reduced average wait time by 54.2% on three production GPU clusters (Ding et al. 2023).

**AI4CI:** Within the *AI4CI* thrust, early results include a hybrid analytical/ML model for searching functionally equivalent CUDA code (Xu et al. 2022), a family of CNN architectures for the detection of soybean defoliation (Ockerman et al. 2023), and improved predictive capability of the ML-based resource prediction models (Vallabhajosyula and Ramnath 2023). Within visual analytics, early results include reusable software components for exploring knowledge graph data that allow bi-directional interaction between multiple-coordinated views to gain a deeper understanding of the graph data. These components have been used to create an interactive knowledge and learning environment for smart foodsheds (as mentioned above), and also for visualizing loss landscapes of multiple training clients in a federated learning system (Tu, Wang, et al. 2023b).

**PADI:** Within the “privacy, accountability and data integrity” thrust (PADI), early results include novel privacy-preserving techniques applied to sequential data sharing (Jiang, Yilmaz, and Ayday 2023), the sharing of summary statistics from sensitive databases, and collaborative quality control for research databases (Dervishi et al. 2023). In ongoing research, we are leveraging model cards and model ontologies for accountability AI models and their use.

## CI/SOFTWARE DEVELOPMENT AND RELEASE

The ICICLE team is also focused on translating research to publicly available CI/software components. It has embarked on a robust CI/software development, testing, and release plan thanks to strong cooperation between the software thrust members and the developers of various CI/software components. Many of these components are being integrated with the TAPIS framework (TAPIS 2023) to provide solutions in the computing continuum environment. Multiple rounds of releases with more than 20 components have been released so far (ICICLE-Software 2023). The team plans to continue periodic releases in realizing the edge-to-center AI-as-a-service enterprise.

## OVERVIEW OF THE BROADER IMPACT ACTIVITIES

The ICICLE Institute is committed to an inclusive environment for all. Its code of conduct guides all members of the ICICLE community regardless of position or seniority and includes a mechanism for reporting and addressing issues

that might arise. The institute further hosted a 3-h facilitated workshop on inclusion at its 2022 All Hands Meeting. Attendees were asked to make a personal pledge to take action to ensure inclusive academic spaces with follow-up surveys to evaluate effectiveness.

The strategy adopted for realizing the broader impacts goals of the institute is anchored in the Collective Impacts model where a single backbone group works collectively across WFD, BPC, and knowledge transfer on shared goals and shared measures.

The institute’s overriding goal for WFD is an ethically aware AI workforce. We begun by defining an ICICLE Ethical AI Framework, before applying the framework to professional development of ICICLE team members themselves, to professional training opportunities, and to research experiences for undergrads, and K-12 student immersive experiences. The BPC goals are twofold: (i) making BPC part of the fabric of the institute, and (ii) building awareness and fostering actions leading to broadened participation.

Broader impacts are realized by ICICLE through deep interaction with stakeholder communities. The institute has diverse engagement vehicles including a student affiliates program, an academic affiliates program, an industry partners consortium for industry partners, and a stakeholder roundtable. The institute is flexible to new engagements and research interactions that we envision as it accelerates in its software availability and CI platform offerings.

Broader impacts programmatic activities are having early success in extending the impact of the institute. In particular, the ICICLE NexGens, an affiliate group composed of ICICLE students, is increasing student voices on institute decisions and increasing a students sense of belonging. The ICICLE Educational Fellows program (Indiana University 2023), which awarded its first cohort of five fellows in 2023, brings to the project an external perspective on stakeholder engagement and democratizing AI.

## NEXUS OF COLLABORATION ACTIVITIES

The institute is actively becoming the nexus of multidisciplinary and multi-organizational collaborative activities for creating intelligent CI to democratize AI. Representative activities include (i) leading the CI resource concierge service for all AI institutes; (ii) leading a Special Interest Group (SIG) on AI Ethics across all AI institutes; (iii) collaboration between ICICLE and AI4Opt AI institute, optimizing food logistics for the USDA-LFPA program; (iv) Joint ICICLE-TIH-Bombay (India) Digital Agriculture activities; (v) diverse set of users using the released

CI/Software components; (vi) joint cross-cutting research and publications across groups/thrusts/organizations; and (vii) new research grants/proposals leveraging ICICLE activities.

## FUTURE ACTIVITIES

Over the course of its first 5 years, ICICLE will establish itself as a leading institute for CI in the field of AI. Its impact will resonate within the domains of high-performance computing, AI and machine learning (ML), as well as in the application-inspired disciplines of animal ecology, digital agriculture, and smart foodsheds. There will be a notable influence on the foundational systems of AI and the interdisciplinary approaches that inherently incorporate AI for enhanced performance and robustness. Given ICICLE's emphasis on carefully selected projects, rubrics and assessment metrics are employed at the conclusion of every virtuous cycle to measure the scholarly impact on both foundational and translational domains. In addition, ICICLE is focused on training and engaging an AI informed and ethically aware workforce by means of efforts that are broadly inclusive. The broader impacts efforts complement technical developments to ensure that both research products and people associated with the institute enter the world accounting for and reflecting different perspectives.

## CONCLUSIONS

ICICLE is positioned as the nexus for foundational and use-inspired AI for CI and CI for AI research among its collaborating institutions, domain scientists, and other NSF AI Institutes, along with external partners. Our overarching objectives are to (i) facilitate collaboration and synergy between CI researchers, AI researchers, and domain scientists; (ii) involve stakeholders throughout the research process; (iii) build and support a community of AI4CI and CI4AI activities; (iv) support the adoption of ICICLE technology, and (v) engage and partner with external organizations for technology transfer and commercialization. ICICLE aims to create a robust and effective ecosystem that accelerates both AI and CI research while maximizing their utility across various scientific disciplines and industries.

## ACKNOWLEDGMENTS

The authors would like to thank all members (students, staff, and faculty) of the ICICLE team for their hard work toward realizing the vision of the institute. The ICICLE project is funded in part by the National

Science Foundation (NSF) under Grant Number OAC-2112606.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Dhabaleswar K. Panda  <https://orcid.org/0000-0002-0356-1781>

## REFERENCES

- ICICLE-Software. 2023. <https://icicle.osu.edu/cyberinfrastructure/software/> (accessed July 4, 2023).
- ICICLE. 2023. <https://icicle.osu.edu/> (accessed July 4, 2023).
- Dervishi, L., W. Li, A. Halimi, X. Jiang, J. Vaidya, and E. Ayday. 2023. "Privacy Preserving Population Stratification for Collaborative Genomic Research." In ISOC Network and Distributed System Security Symposium.
- Ding, Q., P. Zheng, S. Kudari, S. Venkataraman, and Z. Zhang. 2023. "Mirage: Towards Low-interruption Services on Batch GPU Clusters with Reinforcement Learning. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23)." Association for Computing Machinery, New York, NY, USA, Article 25, 1-13. <https://doi.org/10.1145/3581784.3607042>
- Gu, Y., X. Deng, and Y. Su. 2022. "Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments." In Annual Meeting of the Association for Computational Linguistics. <https://api.semanticscholar.org/CorpusID:254853929>.
- Hussain, M. T., G. S. Abhishek, A. Buluç, and A. Azad. 2022. "Parallel Algorithms for Adding A Collection of Sparse Matrices." In 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 285-94. IEEE.
- Indiana University. 2023. "ICICLE Educational Fellows." June 20, 2023. <https://icicle.osu.edu/education-and-outreach/icicle-educational-fellows-program>.
- Jiang, Y., E. Yilmaz, and E. Ayday. 2023. "Robust Fingerprint of Location Trajectories Under Differential Privacy." In Proceedings on Privacy Enhancing Technologies Symposium, volume 2023, 5-20. NIH Public Access.
- Ockerman, S., J. Wu, C. Stewart, and Z. Zhang. 2023. "A Reflection On AI Model Selection for Digital Agriculture Image Datasets." In Workshop on AI for Agriculture and Food Systems.
- TAPIS. 2023. <https://tapis-project.org/> (accessed July 4, 2023).
- Tu, C.-H., H.-Y. Chen, J. Zhong, Z. Mai, V. Pahuja, T. Berger-Wolf, S. Gao, C. Stewart, Y. Su, and W.-L. Chao. 2023. "Holistic Transfer: Towards Non-Disruptive Fine-Tuning with Partial Target Data." In Conference on Neural Information Processing Systems.
- Tu, Y., X. Wang, R. Qiu, H.-W. Shen, M. Miller, J. Rao, S. Gao, et al. 2023. "An Interactive Knowledge and Learning Environment in Smart Foodsheds." *IEEE Computer Graphics and Applications* 43(3): 36-47. <https://doi.org/10.1109/MCG.2023.3263960>.
- Vallabhajosyula, S., and R. Ramnath. 2023. "Towards Characterizing DNNs to Estimate Training Time using HARP (HPC Application Resource (runtime) Predictor)." In Practice and Experience in Advanced Research Computing (PEARC'23).
- Xu, Y., Q. Yuan, E. C. Barton, R. Li, P. Sadayappan, and A. Sukumaran-Rajam. 2022. "Effective Performance Modeling and



Domain-Specific Compiler Optimization of CNNs for GPUs.” In Proceedings of the International Conference on Parallel Architecture and Compilation Techniques, 252–64. <https://doi.org/10.1145/3559009.3569674>.

**How to cite this article:** Panda, D. K., V. Chaudhary, E. Fosler-Lussier, R. Machiraju, A. Majumdar, B. Plale, R. Ramnath, P. Sadayappan, N. Savardekar, and K. Tomko. 2024. “Creating intelligent cyberinfrastructure for democratizing AI.” *AI Magazine* 45: 22–28. <https://doi.org/10.1002/aaai.12166>

## AUTHOR BIOGRAPHIES

**Dhabaleswar K. Panda**, ICICLE Director, is a Professor and University Distinguished Scholar of Computer Science and Engineering at The Ohio State University. His research interests are high-performance computing, exascale computing, deep/machine learning, big data, and cloud computing. He is an IEEE Fellow and the recipient of the 2022 IEEE Charles Babbage Award.

**Vipin Chaudhary**, ICICLE co-PI, is the Kevin J. Kranzusch Professor and Chair of Computer and Data Science at the Case Western Reserve University. His research interests include High Performance CI, AI, Data Science, Medical Computing, and Quantum Computing.

**Eric Fosler-Lussier**, ICICLE co-PI, is the John Makhoul Professor and Associate Chair in Computer Science and Engineering at the Ohio State University. He has research interests in Conversational AI, and is a Fellow of the IEEE and ICSA.

**Raghu Machiraju**, ICICLE co-PI, is a Professor of Biomedical Informatics, Pathology, and Computer Science and Engineering at OSU; he also serves as an Associate Chair for Growth in CSE and previously

served as the Founding Faculty Lead of the Translational Data Analytics Institute. Research interests include domain-specific AI methods, particularly for clinical medicine, and systems AI.

**Amit Majumdar** is the Director of the Data Enabled Scientific Computing Division at the San Diego Supercomputer Center, UCSD and an Associate Professor in the Department of Radiation Medicine and Applied Sciences. Research interests include parallel computing, science gateways and CI resources, and software for HPC and AI.

**Beth Plale**, ICICLE co-PI, is the Michael A and Laurie Burns McRobbie Bicentennial Professor of Computer Engineering and Chair of the Department of Intelligent Systems Engineering at Indiana University Bloomington. Her research interests are in computational and data infrastructure, open science, AI ethics, and data accountability.

**Rajiv Ramnath** is a Professor of Practice in Computer Science and Engineering at OSU with extensive industry and interdisciplinary collaborations. His work covers data science, cyberinfrastructure, software engineering, enterprise strategy, and computing in education.

**Ponnuswamy Sadayappan** is a Professor in the Kahlert School of Computing at the University of Utah. His primary research interests are in compiler optimization and high-performance computing. He is a Fellow of the IEEE.

**Neelima Savardekar** is the Managing Director of ICICLE at The Ohio State University.

**Karen Tomko** is the Director of Research Software Applications. Her research interests include performance improvement for research software and cyberinfrastructure for AI.



## SPECIAL TOPIC ARTICLE

# AI-EDGE: An NSF AI institute for future edge networks and distributed intelligence

Peizhong Ju | Chengzhang Li | Yingbin Liang | Ness Shroff

Ohio State University, Columbus, Ohio, USA

## Correspondence

Peizhong Ju, Ohio State University, Columbus, OH 43016, USA.  
Email: ju.171@osu.edu

The last two authors represent the entire AI-EDGE faculty team.

## Funding information

NSF, Grant/Award Numbers: 2112471, 2225561

## Abstract

This paper highlights the overall endeavors of the NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE) to create a research, education, knowledge transfer, and workforce development environment for developing technological leadership in next-generation edge networks (6G and beyond) and artificial intelligence (AI). The research objectives of AI-EDGE are twofold: “AI for Networks” and “Networks for AI.” The former develops new foundational AI techniques to revolutionize technologies for next-generation edge networks, while the latter develops advanced networking techniques to enhance distributed and interconnected AI capabilities at edge devices. These research investigations are conducted across eight symbiotic thrust areas that work together to address the main challenges towards those goals. Such a synergistic approach ensures a virtuous research cycle so that advances in one area will accelerate advances in the other, thereby paving the way for a new generation of networks that are not only intelligent but also efficient, secure, self-healing, and capable of solving large-scale distributed AI challenges. This paper also outlines the institute’s endeavors in education and workforce development, as well as broadening participation and enforcing collaboration.

## INTRODUCTION

The recent breakthroughs of AI and Machine Learning (ML) as well as their successful applications in a broad range of domains provide a unique opportunity for designing AI-driven next-generation networks that are “intelligent” in many different ways. The NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE) was established to fully explore the power of AI to revolutionize next-generation edge networks. AI-EDGE will develop new foundational AI tools for designing next generation intelligent edge networks (6G and beyond), and enabling true distributed intelligence. Our objective is to

cultivate an environment that will not only foster innovation in cutting-edge technologies but also prepare a new wave of professionals equipped to navigate this rapidly evolving landscape.

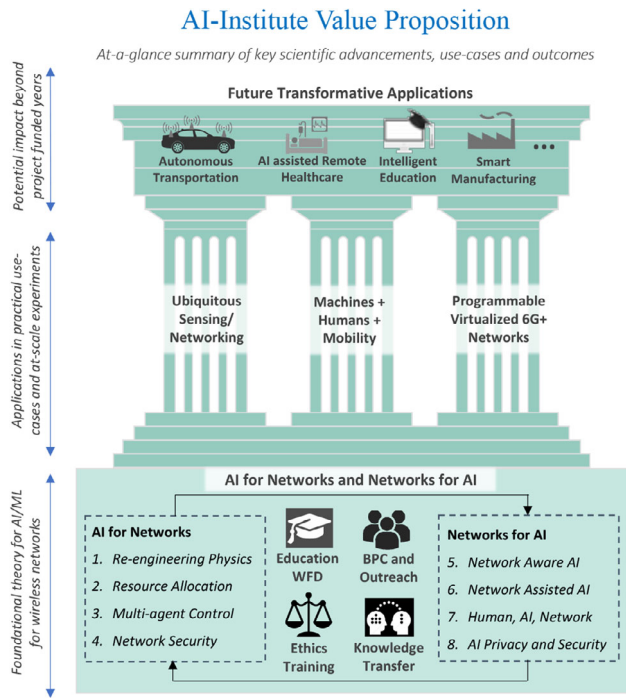
To address the above mentioned challenges, the institute is made of a strong, diverse, and growing consortia of research leaders from university, industry, and government labs that will work collaboratively to realize the overall vision of designing next generation edge networks and distributed intelligence (AI-EDGE 2023).

The overarching mission of the institute’s research is to design next-generation hyper-scalable, heterogeneous and dynamic networks that are highly efficient, reliable, robust, and secure. New AI tools and techniques will be developed to ensure that these networks are self-healing

Peizhong Ju and Chengzhang Li contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Overview of AI-EDGE.

and self-optimized. These networks will in turn be designed to unleash the power of collaboration to help solve long-standing distributed AI challenges, making AI more efficient, interactive, and privacy-preserving. Applications and practical use cases at experimental scale are the pillars (Figure 1) that will be built on and will interact with this foundation in a virtuous cycle to deepen our understanding and magnify the impact of our work. *Use Case 1* is built around sensing and re-engineering the wireless network environment. *Use Case 2* is built around extreme mobility applications where machines, humans, and in some cases network elements could be in motion at the same time. *Use Case 3* focuses on the development of programmable and virtualized edge networks. Looking to the future, our institute will develop the key underlying technologies for distributed and networked intelligence that will have an impact on future transformative applications well-beyond the institute's funded years.

The success of the institute will be measured on several fronts. On the research front, success will be measured by the development of (i) a new intelligence plane for designing next generation edge networks, (ii) how that plane will be used to enable distributed AI, and (iii) the integration of the research tasks to successfully implement the three key use cases identified by the institute. Success in education and workforce development will be measured by the impact on AI-EDGE in helping to build a large and diverse workforce in AI and networks.

The research is currently conducted across eight inter-related thrust areas under two main themes: *AI for Networks* and *Networks for AI*, as shown in Figure 1.

*AI for Networks*: While preliminary successes of AI for Networks have been promising, developing ML tools to networking with minimal or no human oversight poses many important research challenges that need to be explored. In Thrust 1, we re-engineer the physical fabric for 6G and beyond (6G+) wireless communications through AI, thus treating the fabric itself as a controllable entity. In Thrust 2, we develop new AI techniques for the design and control of these next-generation networks taking into account practical resource constraints. In Thrust 3, we further generalize these techniques to allow for multi-agent, possibly noncooperative, network entities. In Thrust 4, we develop AI methods to secure the network edge.

*Networks for AI*: The vision of distributed intelligence requires capabilities beyond today's AI algorithms. The research foundation here starts with Thrust 5, where distributed AI adapts its operation seamlessly by taking into account computation, communication, and data constraints. In Thrust 6, we go further by aiming to re-engineer the networks themselves to better serve the needs of distributed AI operations. In Thrust 7, we investigate collaborative (human + machine) analysis of data produced by our application drivers, which will be far more effective than with either humans or machines in isolation. In Thrust 8, we design and control the network such that it is privacy-aware and can be optimized to facilitate protection from information leakage and attacks.

In the rest of this paper, we will provide a sampling of the institute's recent representative work along the above two main themes. Additionally, we will outline the institute's endeavors in education and workforce development, as well as broadening participation (BP) and enforcing collaboration. More details can be found on the institute webpage (AI-EDGE 2023).

## RESEARCH THEMES OF AI FOR NETWORKS

The growth of edge networks in scale and complexity implies that intelligent and autonomous design becomes paramount. We describe below in detail two key learning areas that the institute has been exploring that will have a demonstrable impact on edge network design and control: (i) reinforcement learning and (ii) meta learning and active learning.

### Reinforcement learning (RL) empowered edge network control

RL is a major AI paradigm in which an online learner interacts with an unknown environment sequentially in order to achieve a large expected cumulative reward. Such a

mathematical framework precisely captures how edge networks acquire, process, and communicate data, and hence can play a vital role in designing data-driven network resource allocation and control mechanisms. AI-EDGE has made a number of fundamental contributions to reinforcement learning with a primary focus on addressing challenging issues that practical edge networks encounter, including various constraints such as resource and safety constraints and switching cost, partial observability, model misspecification, and so forth.

Edge devices are generally subject to resource constraints such as power, computation, and memory constraints. Their data processing can be modeled as constrained Markov decision processes (CMDPs). AI-EDGE researchers have developed several innovative algorithms to handle such constraints while maintaining the best possible performance. For example, Wei, Liu, and Ying (2022) developed model-free RL algorithms with optimal regret and constraint violation guarantees. Ghosh, Zhou, and Shroff (2022) further expanded our exploration to larger, potentially infinite, state spaces via a new model-free constrained RL algorithm that also enjoys near-optimal performance bounds. These novel algorithms exhibit significant performance improvements in comparison to existing solutions, especially in edge networks with latency and energy constraints.

Unlike energy constraints which are typically *cumulative* over time, *safety and power* constraints are instantaneous at each step/time. For example, in autonomous driving, improper operations at any time can cause catastrophic consequences such as crashing of a car. Thus, such safety-critical systems need to avoid violation of safety constraints at each time instance. It is highly challenging because the safety at each step should both depend on history steps due to the system evolution and predict the impact on future steps. Recently, we made a fundamental breakthrough in Shi, Liang, and Shroff (2023) to address such a challenge, where we constructed novel *safe subgraphs* to design safe exploration and planning, and showed that our approach achieves a regret that nearly matches the state-of-the-art regret in the *unsafe-action* setting and that in the *unconstrained* setting.

## Meta-learning for autonomous transportation

Meta-learning is a powerful learning technique to reduce sample requirement for learning tasks while still maintaining the desired performance. AI-EDGE researchers have made several significant contributions to these learning methodologies and have further demonstrated their successful applications in multimodal fusion design for beamforming over autonomous vehicles.

Autonomous vehicles typically move in dynamic conditions, and thus a car may encounter new environments that were not explicitly included in an apriori training dataset. A major challenge here is that a single pre-trained model is unable to generalize well. We approached such a problem by leveraging a popular meta-learning framework of model-agnostic meta-learning (MAML), for which we first made several innovative contributions, including characterizing the optimization performance (Ji, Yang, and Liang 2021), understanding the statistical convergence guarantee (Collins et al. 2022), and characterizing the generalization performance in the overparameterized regime (Ju, Liang, and Shroff 2023). Equipped with these powerful tools, we further demonstrated the advantages of MAML over traditional deep-learning techniques wherein a model retrained in the unseen test environment (i) uses a fraction of the data compared to classical retraining, which, in turn, simplifies data collection and storage, and (ii) results in equal or higher accuracy in optimal beam selection compared to the case when the new environment dataset is fully available during initial training. Our results in Gu, Collins et al. (2023) reveal that our MAML implementation on camera images for the objective of beam selection improves test accuracy by up to 86% with fine-tuning when encountering an unseen non-line-of-sight (NLOS) environment compared to conventional supervised learning. We are currently comparing MAML-based approaches with traditional transfer learning, using initial results in our earlier work (Gu, Salehi, et al. 2023).

## RESEARCH THEMES OF NETWORKS FOR AI

Future AI problems will be at large scale, for which federated learning (FL) implemented over edge networks is a powerful paradigm to decentralize the computation, protect data privacy, and offer maximum design flexibility. To this end, we have made substantial efforts in designing innovative *network-aware* FL algorithms that address various challenges of edge networks such as node heterogeneity, communication constraints, unequal client participation, uncertainty of wireless communication medium, and so forth. We also investigated more complex AI problems, decentralizing which requires further novel designs. Below, we discuss in more detail about our work along these directions.

### Network-aware federated learning (FL)

Nodes in edge networks typically have different computational capabilities, quality of datasets, and levels of





participation. Hence, we have extensively addressed these types of network heterogeneity and proposed new FL algorithms that are communication-efficient, personalized, straggler-resilient, and privacy-preserving.

As one such work, we studied model heterogeneous FL for classification where different clients have different model architectures (Regatti et al. 2022). Unlike existing works on model heterogeneity, we do not require access to a public dataset, nor do we impose constraints on the model architecture of clients. We also ensure that the clients' model architectures and data are private. We developed a communication-efficient algorithm with provable convergence and generalization properties that aligns conditional distributions of each client in the feature space and achieves a consensus on the final layer classification weights.

We also proposed a novel straggler-resilient FL method that incorporates statistical characteristics of clients' data to adaptively select the clients in order to speed up the learning procedure (Reisizadeh et al. 2022). The key idea is to start the training with faster nodes and gradually involve slower nodes once the statistical accuracy of the data corresponding to the current participating nodes is reached. The proposed approach reduces the overall runtime (theoretically and via numerical evaluation) required to achieve the statistical accuracy of data of all nodes, as the solution for each stage is close to the solution of the subsequent stage with more samples and can be used as a warm-start.

## Federated learning with diverse structures

AI problems can have diverse structures in order to capture various practical domains, encompassing the basic risk *minimization* problem such as supervised learning, the *minimax* problem such as robust adversarial training and generative adversarial networks, and the *bilevel* problem that captures meta-learning, hyperparameter optimization, neural architecture search, and so forth. Building on our work on decentralizing the basic minimization as discussed in the preceding subsection, we have recently pioneered the research on FL into the much wider scope of minimax and bilevel problems while still addressing network constraints and heterogeneity.

More specifically, we recently proposed a general federated minimax optimization framework (Sharma, Panda, and Joshi 2023), that accounts for heterogeneity across clients and subsumes several existing methods like local stochastic gradient descent-ascent (SGDA). To fix this mismatch caused by heterogeneity, we proposed normalizing the client updates by the number of local steps undertaken between successive communication rounds. Our

algorithms showed significant improvements over existing computation and communication complexity results.

## EDUCATION AND BROADENING PARTICIPATION

There is a tremendous demand for AI-related skills not only in computer science and engineering, but also across many disciplines, from agriculture and manufacturing to business and entertainment. Further, the growth of future-generation wireless edge networks will require a large workforce with expertise in wireless networking and communications, security, and privacy. Thus, it is a priority to educate students, professionals, and practitioners in AI and networks and substantially grow and diversify the workforce. Since it was founded, AI-EDGE has made significant progress in educating undergraduate and graduate students, practitioners, and professionals, and expanding our reach to under-represented communities. Some highlights are provided below.

AI-EDGE team members received a major grant for establishing and growing the workforce needed to expand 5G and broadband access in Ohio 5G-Ohio website (2023). This has led to synergistic work between AI-EDGE and the 5G-OH Broadband and 5G Connectivity Center to establish BS specialization and Professional MS programs at OSU.

AI-EDGE has developed an educational repository for AI and networks, where the course contents have been partitioned into self-contained modules covering different topics, each with student notes, lecturer notes, and appropriate paper reading assignments. The course repository also includes bridge projects to integrate learning across AI and the networks, personalized curriculum for individual students and professionals, and professional MS programs via "stackable certificates" with modular contents from across universities.

Further, AI-EDGE consistently provides industrial experience and training opportunities for students and postdocs. We currently have multiple ongoing collaborations, where PhD students and postdocs work with scientists and engineers across several industries.

To facilitate synergistic research interactions and expand our students' scope of knowledge, the institute hosts a highly successful regular online seminar series that feature research talks and tutorials given by students, faculty, and postdocs of the institute, as well as world-leading researchers in networking and AI. Recordings of the talks are maintained on the institute's website for easy access. We also organize regular institute-wide meetings, during which the students and postdocs are given opportunities to present lightning talks and posters about their

research outcomes, which significantly encouraged their engagement in this multidisciplinary research program.

*Inclusivity and BP* are cornerstones of AI-EDGE's ethos. We continually keep this focus at the forefront of our institute's activities. We next summarize three key BP activities and efforts that we have made during the initial 2 years of AI-EDGE's establishment.

*Building a large diverse workforce:* AI-EDGE has partnered with another OSU-AFRL led large-MSI consortium to develop a community of under-represented engineers and scientists and provide them with opportunities to develop science, technology, and leadership skills. By leveraging this consortium, AI-EDGE has recruited undergraduate students from under-represented groups (URGs) and run two 8-week summer programs during which the students work with their mentors who are faculty and postdocs/graduate students at AI-EDGE. The research projects that students can select cover a broad range of topics in ML, networking, and security.

*Women in AI program:* The institute has launched an institute-wide virtual "Women in AI program" participated by all female faculty, students, and postdocs as well as other institute's members. The program events are organized by a student committee and monitored by the BP committee. The institute has organized several networking events with different themes for women students and postdocs, including research highlight presentations, panel discussions on topics of interest such as the process of academic job search, work-life balance, and strategies for a successful career, and more.

*Community outreach and engagement:* AI-EDGE faculty members have given several interviews to local and international news organizations, at libraries, and keynotes/invited talks on the importance of AI and networking technologies. Our faculty, students, and postdocs across the institute have actively participated in various outreach events for high-school and undergraduate students and teamed up to give tutorials at various high-school summer camps.

## CONCLUSION

In this paper, we have introduced the AI-EDGE institute, which seeks to build on the synergies between AI/ML and networking to develop intelligent next-generation edge networks and distributed AI. We have provided an overview of the institute's key focus areas and summarized some of our exciting research advancements on various learning methods (such as RL and FL) and their applications. We have also highlighted some main AI-EDGE endeavors in education, workforce development, and BP.

## ACKNOWLEDGMENTS

The AI Institute is primarily funded by the NSF and DHS through NSF grant 2112471. The Institute has also received additional supplements for NSF 2225561 for extending the work of the institute.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Peizhong Ju  <https://orcid.org/0000-0002-4569-3539>

Chengzhang Li  <https://orcid.org/0000-0003-3142-0131>

## REFERENCES

- 5G-Ohio website. 2023. "5G Ohio." <https://5g-oh.osu.edu/> (accessed June 27, 2023).
- AI-EDGE. 2023. "AI-EDGE Institute." <https://aiedge.osu.edu/> (accessed June 27, 2023).
- Collins, L., A. Mokhtari, S. Oh, and S. Shakkottai. 2022. "MAML and ANIL provably Learn Representations." In *ICML*.
- Ghosh, A., X. Zhou, and N. Shroff. 2022. "Provably Efficient Model-Free Constrained RL with Linear Function Approximation." In *NeurIPS*.
- Gu, J., L. Collins, D. Roy, A. Mokhtari, S. Shakkottai, and K. R. Chowdhury. 2023. "Meta-Learning for Image-Guided Millimeter-Wave Beam Selection in Unseen Environments." In *ICASSP*.
- Gu, J., B. Salehi, S. Pimple, D. Roy, and K. R. Chowdhury. 2023. "TUNE: Transfer Learning in Unseen Environments for V2X mmWave Beam Selection." In *ICC*.
- Ji, K., J. Yang, and Y. Liang. 2022. "Theoretical Convergence of Multi-Step Model-Agnostic Meta-Learning." *JMLR*. 23(29): 1–41.
- Ju, P., Y. Liang, and N. Shroff. 2023. "Theoretical Characterization of the Generalization Performance of Overfitted Meta-Learning." In *ICLR*.
- Regatti, J., S. Lu, A. Gupta, and N. Shroff. 2022. "Conditional Moment Alignment for Improved Generalization in Federated Learning." In *NeurIPS*.
- Reisizadeh, A., I. Tziotis, H. Hassani, A. Mokhtari, and R. Pedarsani. 2022. "Straggler-Resilient Federated Learning: Leveraging the Interplay Between Statistical Accuracy and System Heterogeneity." *JSAIT* 3: 197–205.
- Sharma, P., R. Panda, and G. Joshi. 2023. "Federated Minimax Optimization with Client Heterogeneity." *Published in Transactions on Machine Learning Research*, December 2023.
- Shi, M., Y. Liang, and N. Shroff. 2023. "A Near-Optimal Algorithm for Safe Reinforcement Learning Under Instantaneous Hard Constraints." In *ICML*.
- Wei, H., X. Liu, and L. Ying. 2022. "Triple-Q: A Model-Free Algorithm for Constrained Reinforcement Learning with Sublinear Regret and Zero Constraint Violation." In *AISTATS*.

**How to cite this article:** Ju, P., C. Li, Y. Liang, and N. Shroff. 2024. "AI-EDGE: An NSF AI institute for future edge networks and distributed intelligence." *AI Magazine* 45: 29–34. <https://doi.org/10.1002/aaai.12145>

## AUTHOR BIOGRAPHIES

**Peizhong Ju** received his B.S. degree in Electrical Engineering from Peking University, Beijing, China, in 2016, and Ph.D. degree in Electrical and Computer Engineering from Purdue University, West Lafayette, IN, USA, in 2021. He is currently a Postdoc Researcher at the ECE Department at The Ohio State University and is also affiliated to AI-EDGE Institute. His current research interests are about analyzing the performance of various machine-learning models and designing new machine-learning algorithms for multi-agent systems.

**Chengzhang Li** received the B.S. degree in Electronics Engineering from Tsinghua University, Beijing, China, in 2017, and the M.S. and Ph.D degrees in Computer Engineering from Virginia Tech, Blacksburg, VA, USA, in 2020 and 2022, respectively. He is currently a Postdoc Researcher at the AI-EDGE Institute, The Ohio State University. His current research interests are machine learning in edge networks, real-time scheduling in 5G, and Age of Information (AoI).

**Yingbin Liang** received the Ph.D. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign in 2005. She is currently a Professor with the Department of Electrical and Computer Engineering, The Ohio State University (OSU), and a core Faculty Member of the Translational Data Analytics Institute (TDAI), OSU. Currently, she is also serving as the Deputy Director for the AI-Edge Institute, OSU. Before she joined OSU, she served on the Faculty of the University of Hawaii and Syracuse University. Her research interests include machine learning, optimization, information theory, and statistical

signal processing. She received the National Science Foundation CAREER Award and the State of Hawaii Governor Innovation Award in 2009. She also received the EURASIP Best Paper Award in 2014.

**Ness B. Shroff** received the Ph.D. degree in Electrical Engineering from Columbia University, New York, NY, USA, in 1994. He joined Purdue University, West Lafayette, IN, USA, immediately thereafter as an Assistant Professor with the School of Electrical and Computer Engineering. At Purdue, he became a Full Professor of ECE and the Director of a University-Wide Center on Wireless Systems and Applications in 2004. In 2007, he joined The Ohio State University, Columbus, OH, USA, where he holds the Ohio Eminent Scholar Endowed Chair in networking and communications, with the Departments of ECE and CSE. He is currently the Institute Director of the NSF AI Institute for Future Edge Networks and Distributed Intelligence. He holds or has held Visiting (chaired) Professor positions with Tsinghua University, Beijing, China, Shanghai Jiaotong University, Shanghai, China, and the Indian Institute of Technology Bombay, Mumbai, India. He was the recipient of numerous best paper awards for his research and is listed in Thomson Reuters' on The World's Most Influential Scientific Minds, and has been noted as a Highly Cited Researcher by Thomson Reuters in 2014 and 2015. He is currently the Steering Committee Chair for ACM Mobihoc and was Editor in Chief of the IEEE/ACM TRANSACTIONS ON NETWORKING. He also was the recipient of the IEEE INFOCOM Achievement Award for seminal contributions to scheduling and resource allocation in wireless networks.



**SPECIAL TOPIC ARTICLE**

# Institute for Foundations of Machine Learning (IFML): Advancing AI systems that will transform our world

**Adam Klivans** | **Alexandros G. Dimakis** | **Kristen Grauman** | **Jonathan I. Tamir** | **Daniel J. Diaz** | **Karen Davidson**

Department of Computer Science, The University of Texas at Austin, Austin, Texas, USA

### Correspondence

Adam Klivans, Department of Computer Science, The University of Texas at Austin, Austin, TX, USA.  
Email: [karen\\_davidson@utexas.edu](mailto:karen_davidson@utexas.edu)

### Funding information

National AI Research Institutes program

### Abstract

The Institute for Foundations of Machine Learning (IFML) focuses on core foundational tools to power the next generation of machine learning models. Its research underpins the algorithms and data sets that make generative artificial intelligence (AI) more accurate and reliable. Headquartered at The University of Texas at Austin, IFML researchers collaborate across an ecosystem that spans University of Washington, Stanford, UCLA, Microsoft Research, the Santa Fe Institute, and Wichita State University. Over the past year, we have witnessed incredible breakthroughs in AI on topics that are at the heart of IFML's agenda, such as foundation models, LLMs, fine-tuning, and diffusion with game-changing applications influencing almost every area of science and technology. In this article, we seek to highlight the application of foundational machine learning research on key use-inspired topics:

- Fairness in Imaging with Deep Learning: designing the correct metrics and algorithms to make deep networks less biased.
- Deep proteins: using foundational machine learning techniques to advance protein engineering and launch a biomanufacturing revolution.
- Sounds and Space for Audio-Visual Learning: building agents capable of audio-visual navigation in complex 3D environments via new data augmentations.
- Improving Speed and Robustness of Magnetic Resonance Imaging: using deep learning algorithms to develop fast and robust MRI methods for clinical diagnostic imaging.

IFML is also responding to explosive industry demand for an AI-capable workforce. We have launched an accessible, affordable, and scalable new degree program—the MSAI—that looks to wholly reshape the AI/ML workforce pipeline.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 UT Austin. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



## INTRODUCTION

The Institute for Foundations of Machine Learning (IFML) was part of the first cohort of the National Science Foundation's artificial intelligence (AI) institutes. We are home to a dynamic and diverse team of researchers focusing on machine learning foundations—the apparatus that underpins AI. Headquartered at The University of Texas at Austin, IFML researchers collaborate across an ecosystem that spans University of Washington, Microsoft Research, and Wichita State University. Together, we conduct foundational research that has the potential to impact every person on the planet. Over the past year, we have witnessed incredible breakthroughs in AI on topics that are at the heart of IFML's agenda: foundation models, generative models, fine-tuning, diffusion, algorithms, data augmentation, robustness, and reinforcement learning, to name a few. It was a thrilling year for our team, as we saw game-changing applications influencing almost every area of science and technology.

IFML works to understand the key foundational questions that need to be solved so that machine learning can continue its upward trajectory. Part of IFML's mission, as mandated by the NSF, is to function as a nexus point for other institutes and centers that may have overlapping interests in machine learning. We believe in the power of foundational research and its importance to both short and long-term innovation in machine learning. Within the current empirical framework, reducing the amount of trial and error using principled heuristics is extremely impactful. New algorithms and analyses have the potential to dramatically reshape the field. For example, the invention of polynomial-time interior-point methods for solving linear programs has influenced nearly every aspect of optimization, an area at the core of modern machine learning systems. Additionally, foundational researchers are now routinely hired by the largest technology companies. Given this tight integration, theoretical research is positioned to have major influence.

To meet future demand for a highly skilled AI workforce, IFML members were instrumental in developing coursework for a new [Master of Science in Artificial Intelligence](#) (MSAI) degree program at The University of Texas at Austin. The MSAI was featured in [The New York Times](#) due to its potential to reshape the landscape of AI education.

The MSAI is explicitly designed to deliver affordability, accessibility, and scalability. These are the same traits that make this a uniquely valuable tool for workforce development:

- **Accessibility:** Because the program is online and part-time for most students, it can be completed without

the student having to leave the workforce, relocating, or even making night and/or weekend trips to campus as is required by many executive education programs.

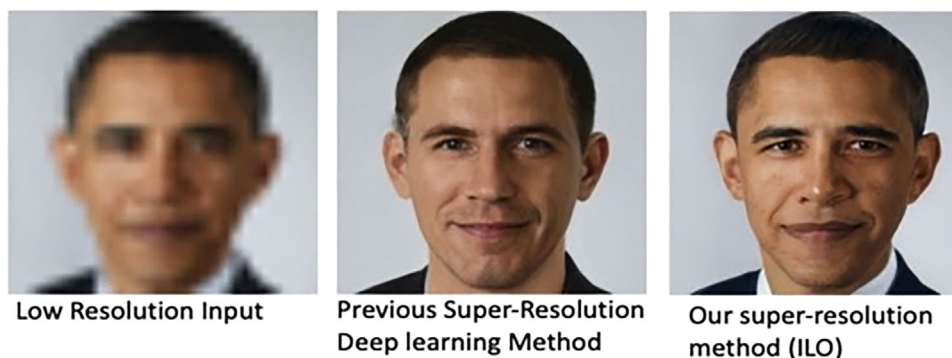
- **Affordability:** The program is intentionally priced to mitigate one of the most significant barriers to graduate study: prohibitively high tuition. Whereas traditional specialty MS programs are typically priced between \$500,000 and \$100,000, the MSAI program will be priced at approximately \$10,000.
- **Scalability:** In steady state, we expect to enroll more than 2500 MSAI students while graduating more than 700 students per year. We already have clear evidence of the viability of this delivery model through our online MS programs in Computer Science (CS) and Data Science (DS), which together currently enroll more than 3000 students. We began accepting applications for the MSAI on June 1 and received more than 2900 applications for the inaugural cohort.

## RESEARCH HIGHLIGHT 1 – FAIRNESS IN IMAGING WITH DEEP LEARNING

We have all seen the moment in the movies—detectives gather around a computer screen staring at a blurry image captured from a store security camera. With just a few clicks, magical “*zoom and enhance*” software unblurs the fuzzy image, revealing the culprit's license plate or the face of the villain. It is still science fiction, but the reality is that Deep Learning methods are getting closer to realizing this dream. Modern AI techniques can render high-quality images from blurry and low-resolution samples. These techniques can transform computational photography, medical imaging, increase resolution, and accelerate MRI and microscopy.

In 2020, a deep learning generative model with groundbreaking performance was posted on the web. The model could turn low-resolution images to high-quality photos. A user uploaded a low-resolution image of President Obama, to obtain an image that is now called “*White Obama*.” This was a startling display of bias in AI imaging algorithms. The deep learning model was reconstructing images with predominantly white features and a heated debate arose across Twitter and other social media. Alex Dimakis, IFML co-director and a Professor in the Chandra Family Department of Electrical and Computer Engineering, led a team of researchers to investigate. *Was the source of the problem the bias in the training data, or in the reconstruction algorithm, and how could we fix it?*

The algorithm that generated these images, PULSE, is using a deep generative model called StyleGAN that can produce artificial faces (like the ones shown on



**FIGURE 1** Low-resolution image input with comparisons of previous super-resolution deep learning method and the research team's super-resolution method.

thispersondoesnotexist.com) but is optimizing the generated image to match the low-resolution input image after downscaling. StyleGAN was trained on a dataset of images of predominantly white people. Many researchers argued that this was the source of the bias, and that simply creating a balanced training dataset would solve the problem. The team's recent work (Figure 1) shows that this is not the case: the reconstruction algorithm needs to be modified, in addition to the training set.

The first problem is that we needed to think carefully about defining fairness in reconstructing images of people with various attributes. Traditional group fairness definitions are defined with respect to specified protected groups—camouflaging the fact that these groupings are artificial and carry historical and political motivations. For instance, should South and East Asians be viewed as a single group or separate groups? Should we consider one race as a whole or further split by gender? Choosing which groups are valid and who belongs in them is an impossible dilemma and being “fair” with respect to Asians may require being “unfair” with respect to South Asians. This motivates our introduction of oblivious fairness. The machine learning algorithm needs to work for all possible groupings of the population.

Our first result is that several intuitive notions of group fairness are incompatible and impossible to achieve simultaneously. We show that the natural extension of demographic parity is strongly dependent on the grouping, and impossible to achieve obliviously. We introduce a new definition of fairness called Conditional Proportional Representation which can be achieved obliviously (i.e., without defining specific protected groups) through a natural algorithm that we propose.

Our second result is that the reconstruction algorithm previously used (MAP inference) is amplifying any bias that can be present in the data. The essence of why is illustrated in the following toy example: Alice flips a biased coin that comes Heads with probability 0.6 and Tails with

0.4. Bob has to guess Heads or Tails, knowing this is a biased coin. If Bob wants to maximize his probability of winning, he will always guess Heads, even when the bias is only 60%. This increases the 60% dataset bias to 100%, and we observe this bias amplification phenomenon experimentally in our study. On the contrary, if Bob uses the algorithm we propose (posterior sampling), Bob will randomize his guess and propose Heads only 60% of the time, matching but not amplifying the bias in the training data (Figure 2).

As deep learning imaging algorithms become ubiquitous across smartphones, social networks, MRI scanners, and a broad array of other applications, designing the correct metrics and understanding the foundations of deep learning is going to be key to ensure future deployments reflect the diverse and inclusive reality of our world.

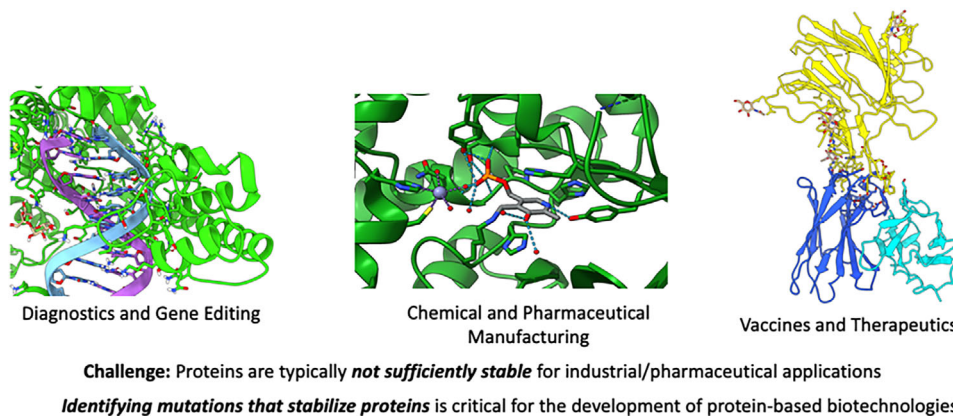
## RESEARCH HIGHLIGHT 2 – DEEP PROTEINS

One of the holy grails of biotechnology is the ability to engineer proteins by introducing mutations into their sequences in order to improve their function. To create protein-based biotechnologies, a thermostable scaffold is often needed for downstream industrial and therapeutic applications. For example, we may wish to thermostabilize the COVID-19 spike protein to improve the ability of our innate adaptive immune system to produce neutralizing antibodies. The computational stability prediction community, however, has struggled to find models that can generalize due to the large and highly complex nature of atomic-level data, as well as a lack of properly curated training sets. IFML members are focused on the fundamental problem of developing machine learning models that can predict stabilizing mutations on a given protein. Over the past year, IFML has developed new foundational methods for data augmentation, new representations and



**FIGURE 2** An example of reconstructions using posterior sampling compared to the previous method PULSE which uses MAP inference and amplifies the bias (Jalal et al., 2021a).

## Proteins are the backbone of biotechnology



**FIGURE 3** Identifying mutations that stabilize proteins can help advance innovations for industrial or pharmaceutical use.

algorithms for fast inference, and novel deep learning architectures (Figure 3).

The standard approach for building a thermostability predictor is to use a billion-parameter sequence model as a black box and fine-tune on a small-labeled training set. As our aim is to move beyond the limitations of these methods, we created new training sets using a simple and powerful data augmentation technique and leveraged both structure and sequence-based models to address both geometric and physical properties of proteins. The results

required us to dig into foundational deep learning problems involving new architectures and representations. Our data augmentation technique opens the door for an order-of-magnitude increase in available data across a wide variety of protein engineering domains, as noted in “Stability Oracle: A Structure-Based Graph-Transformer for Identifying Stabilizing Mutations.”

Due to experimental limitations, most datasets of mutations are biased toward particular amino acids. This makes it extremely difficult to make predictions on the entire

space of mutations. Thermodynamic permutations (TP) is the first thermodynamic stability training set that sampled all 380 mutation types. We are currently extending this to higher order mutations. To address data leakage, we first generated a more comprehensive test set that better sampled the 380 mutation types and then used protein similarity to ensure no overlap between training test splits.

With these data enhancements in hand, we developed Stability Oracle, a structure-based stability predictor that, given a protein microenvironment and “to” and “from” amino acids, can rapidly predict the associated thermodynamic change. Prior structure-based work in thermostability prediction requires a complete structure of both the wildtype and mutated protein. This is computationally prohibitive, as it requires running AlphaFold to compute a single prediction. Further it has been empirically shown that AlphaFold struggles to update a protein structure based on a single mutation. Instead, we use a self-supervised pretraining phase based on previous work by Daniel Diaz on the MutCompute model but with an updated graph transformer architecture. This allows us to use representations of the “from” and “to” amino acids along with a microenvironment instead of obtaining entirely new structures. Fine tuning this Stability Oracle model using our TP data augmentation results in state-of-the-art thermostability prediction across all standard benchmarks and outperforms on all metrics.

Next, we discovered a way to extract “from” and “to” amino acid representations using AlphaFold, which is typically used for protein structure prediction. In fact, several publications have highlighted the inability of AlphaFold to capture changes in free energy of point mutations in proteins. Our approach and resulting model, StabilityFold, contradicts this conventional wisdom and demonstrates that AlphaFold can be used to learn changes in free energy of point mutations. Further, we developed a new parallel algorithm called MutateEverything, which not only fine-tunes AlphaFold embeddings for single point mutations but can be generalized to high order mutations and can make millions of predictions using a single forward pass through AlphaFold’s evoformer. Our method is quite general, and we can “plug-in” sequence-based models such as Prostate to obtain state of the art sequenced-based thermostability predictors that outperform for example ESM.

IFML is collaborating with UT Austin’s Ellington lab to stabilize a lipase that can accelerate field diagnostics for Tuberculosis and stabilize a polymerase for synthetic biology applications. With the Georgiou Lab, we are stabilizing two human enzymes with applications in breast and colon cancer, respectively. We are working with the McClellan lab to accelerate the stabilization of several viral spike pro-

teins for vaccine design. Finally, we are working with the Adapt lab at Houston Methodist Research Institute where we are working on improving the expression of COVID-19 antibodies and exploring the ability of Stability Oracle to predict mutations that improve binding affinity to the Omicron spike protein.

Long term, our goal is to create a suite of deep learning tools that accelerate the engineering of different protein phenotypes and provide the computational foundation for the biomanufacturing revolution of the chemical, pharmaceutical, agrochemical industries.

### RESEARCH HIGHLIGHT 3 – SOUNDS AND SPACE FOR AUDIO-VISUAL LEARNING

Humans learn by interacting with the world. IFML researchers have adopted a similar strategy to teach virtual agents new skills. We use all our senses when we navigate the world, but today’s embodied AI agents—like robots or virtual assistants—are typically restricted to using only visual perception of their environment. IFML member Kristen Grauman, a Professor in the Department of Computer Science at UT Austin and a Research Scientist in Facebook AI Research (FAIR), is working to fill this void by building agents capable of audio-visual navigation in complex, acoustically and visually realistic 3D environments.

SoundSpaces is a first-of-its-kind audio-visual platform for embodied AI. We are working to produce realistic audio rendering based on room geometry, materials, camera position, and sound location— so we can create smart agents that can better respond to real-world situations. In this [demo](#), we see the platform responding to a real-life scenario, such as a fire alarm going off during a piano lesson.

Building on their SoundSpaces platform, the team has developed multimodal deep reinforcement learning approaches to train navigation policies end-to-end from a stream of egocentric audio-visual observations. These policies allow the agent to (1) discover elements of the geometry of the physical space indicated by the reverberating audio and (2) detect and navigate to sound-emitting targets. Additionally, the project introduces a dataset of audio renderings based on geometrical acoustic simulations for two sets of publicly available 3D environments (Matterport3D and Replica), and extends Habitat to support the new sensor, making it possible to insert arbitrary sound sources in an array of real-world scanned environments. The research shows that audio greatly benefits from embodied visual navigation in 3D spaces and lays the groundwork for new research in embodied AI with audio-visual perception.



The latest iteration of SoundSpaces delivers on-the-fly geometry-based audio rendering for 3D environments. Given a 3D mesh of any real-world environment, SoundSpaces can generate highly realistic acoustics for arbitrary sounds captured from arbitrary microphone locations. Together with existing 3D visual assets, it supports an array of audio-visual research tasks, such as audio-visual navigation, mapping, audio source localization and separation, and visual-acoustic matching.

Compared to existing resources, SoundSpaces 2.0 has the advantages of allowing continuous spatial sampling, generalization to novel environments, and configurable microphone and material properties. This is the first geometry-based acoustic simulation that offers high fidelity and realism while also being fast enough to use for embodied learning. SoundSpaces and SoundSpaces 2.0 are publicly available to facilitate wider research for perceptual systems that can both see and hear.

## RESEARCH HIGHLIGHT 4 – IMPROVING SPEED AND ROBUSTNESS OF MAGNETIC RESONANCE IMAGING

“Stress inducing” is often a phrase associated with magnetic resonance imaging (MRI) and more than 12 million people undergo the procedure each year. Jon Tamir, assistant professor in the Chandra Family Department of Electrical and Computer Engineering at UT Austin, works with a team of researchers to develop fast and robust MRI methods for clinical diagnostic imaging. Working with colleagues at UT Austin’s Dell Medical School, Tamir and his team are developing machine learning methods to shorten the times MRI exams can take while likewise extracting more data from the process.

MRI is an exceptional imaging modality because it does not use harmful radiation, but scan time is a major barrier to wider clinical adoption. Making the scanner faster and more comfortable will go a long way to serving a larger population, especially in pediatric and neonatal settings where a sick patient cannot be expected to stay still for very long.

Tamir and his team are using foundational AI algorithms developed at IFML to combine deep learning with biomedical imaging in a principled manner. A significant body of prior work focuses on developing end-to-end deep learning methods that reconstruct images from MRI measurements. The use of end-to-end deep learning methods is hindered by their black-box nature, due to lack of interpretability, trust, and robustness guarantees. Such methods perform extremely well when evaluated in carefully controlled environments but are fragile when used in clinical environments with disparate scanner hardware,

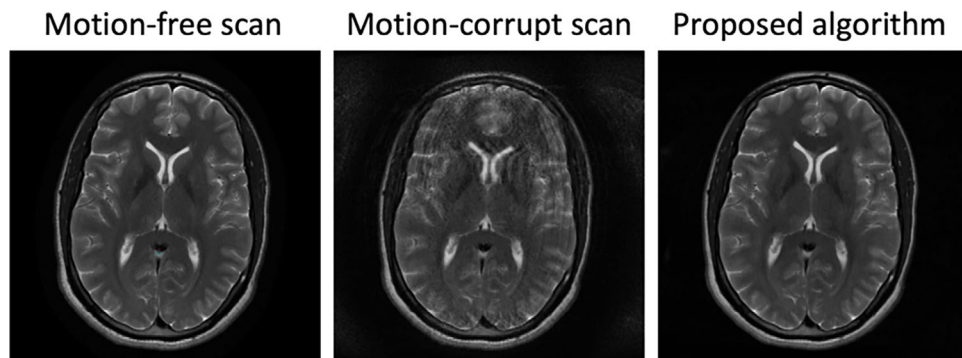
imaging protocols, and patient populations. The reason for this fragility is the explicit coupling of the measurement model and the statistical image prior during training (Jalal et al., 2021b). In essence, a prescribed measurement operator and a dataset must both be specified at train time. When the test-time conditions differ, then the reconstructions will suffer from artifacts due to generalization error. In practice, clinical MRI demands the flexibility to change measurement settings on a per-patient basis to accommodate the natural heterogeneity in patient populations.

To overcome these limitations, the team designed an algorithm that explicitly decouples the measurement model and the statistical image prior. They used established and principled statistical physical models to separately describe the likelihood function of the measurement system. Then, they trained a foundation generative model on MRI images, similar to how Dall-E and Stable Diffusion were trained on images from the web. They combined the measurement model and the generative model and posed the image reconstruction task as sampling from the posterior distribution. As a result, they were able to accelerate the MRI scan by factors of 4x–12x while maintaining high image quality (Levac, Jalal, & Tamir, 2023). A benefit of this approach is the ability estimate uncertainty in the reconstruction, potentially providing more nuanced information about the quality of the result to the clinician. In addition to estimating uncertainty in the reconstruction, Tamir and coworkers extended their framework to incorporate uncertainty in the measurement model itself. For example, patient motion during the scan leads to ambiguity in the acquired measurements. By modeling the motion as an unknown random variable, they were able to extend the reconstruction to sample from the joint posterior of both the image and the motion parameters. Motion is a serious issue in MRI and this approach lets us get diagnostic images even when the patient moves during the scan.

Experimental scan performed by Tamir’s team at Dell Medical School on a healthy volunteer with institutional review board approval. The left image shows a brain MRI scan performed when the subject was still. The middle image is the result when the subject moved during the scan and shows artifacts near the ventricles. The right image is the proposed algorithm that removes the motion-induced artifacts (Figure 4).

## SUMMARY

IFML research in use-inspired areas—imaging, video, navigation, and protein design/engineering—are the applications that have been at the heart of some exciting recent breakthroughs. IFML senior personnel are closely



**FIGURE 4** The research team extended their framework to incorporate uncertainty, such as patient movement, in the measurement model itself (Levac, Jalal, & Tamir, 2023).

connected to industrial partners in these areas, resulting in research that is aligned with the latest developments and has the potential for near-term deployment.

Our use-inspired work has become considerably more collaborative, as common foundational techniques (e.g., transformers) find applications across multiple modalities. As such, foundational research is now vital for achieving impact in diverse application areas. We use new algorithms, data sets, and architectures to ensure fairness in generative imaging, to enable robots to navigate in ever-changing environments, to deploy more robust MRI in clinical health settings, and to develop new biologics and therapeutics.

Leveraging our strong partnerships in education and broadening outreach efforts, IFML continues to address the demand for an increasingly AI-centric workforce. We strive to be a leader in AI education by providing a globally available, low-cost online Master of Science in AI.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National AI Research Institutes program supported by NSF under Award No. 2019844. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Karen Davidson  <https://orcid.org/0009-0000-4528-7998>

## REFERENCES

Jalal A., S. Karmalkar, J. Hoffmann, A. G. Dimakis, and E. Price. 2021a. "Fairness for Image Generation with Uncertain Sensitive Attributes." In International Conference on Machine Learning (ICML).

Jalal A., M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. I. Tamir. 2021b. "Robust Compressed Sensing MRI with Deep Generative Priors." In Advances in Neural Information Processing Systems (NeurIPS).

Levac B., A. Jalal, and J. I. Tamir. 2023. "Accelerated Motion Correction for MRI Using Score-Based Generative Models." In IEEE International Symposium on Biomedical Imaging (ISBI).

**How to cite this article:** Klivans, A., AG. Dimakis, K. Grauman, JI. Tamir, DJ. Diaz, and K. Davidson. 2024. "Institute for Foundations of Machine Learning (IFML): Advancing AI systems that will transform our world." *AI Magazine* 45: 35–41. <https://doi.org/10.1002/aaai.12163>

## AUTHOR BIOGRAPHIES

**Adam Klivans** is IFML director and a professor in the Department of Computer Science at UT Austin.

**Alexandros G. Dimakis** is IFML co-director and a professor in the Chandra Family Department of Electrical and Computer Engineering at UT Austin.

**Kristen Grauman** is a professor in the Department of Computer Science at UT Austin.

**Jonathan I. Tamir** is an assistant professor in the Chandra Family Department of Electrical and Computer Engineering at UT Austin.

**Daniel J. Diaz** is an IFML postdoctoral researcher at UT Austin.

**Karen Davidson** is communications coordinator for IFML and the Machine Learning Lab at UT Austin.



## SPECIAL TOPIC ARTICLE

# AI4OPT: AI Institute for Advances in Optimization

Pascal Van Hentenryck | Kevin Dalmeijer

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

### Correspondence

Pascal Van Hentenryck, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.  
Email: [pvh@gatech.edu](mailto:pvh@gatech.edu)

### Funding information

US National Science Foundation, Grant/Award Number: 2112533

### Abstract

This article is a short introduction to AI4OPT, the NSF AI Institute for Advances in Optimization. AI4OPT fuses AI and optimization, inspired by societal challenges in supply chains, energy systems, chip design and manufacturing, and sustainable food systems. By combining machine learning and mathematical optimization, AI4OPT strives to develop AI-assisted optimization systems that bring orders of magnitude improvements in efficiency, perform accurate uncertainty quantification, and address challenges in resiliency and sustainability. AI4OPT also applies its “teaching the teachers” philosophy to provide longitudinal educational pathways in AI for engineering.

## INTRODUCTION

The mission of the NSF Artificial Intelligence (AI) Institute for Advances in Optimization (AI4OPT) is

*to revolutionize decision making at massive scales by fusing AI and mathematical optimization and delivering scientific breakthroughs that the two fields cannot achieve independently.*

This Institute pursues this objective by integrating the model-driven paradigm typically followed in operations research with data-driven methodologies coming from AI. The research in AI4OPT is use-inspired, addressing fundamental societal and technological challenges of our times. They include the following: How to design agile, sustainable, resilient, and equitable supply chains? How to operate energy systems powered by distributed renewable energy resources? How to deliver a step change in chip design and manufacturing, and manufacturing as a whole? How to create sustainable ecosystems within

the food–water–energy nexus? AI4OPT focuses on *AI for Engineering*, which raises deep scientific challenges in terms of reliability, robustness, and scalability. Indeed, AI4OPT is driven by high-stake applications that feature physical, engineering, and business constraints, and complex objectives that must balance efficiency, resilience, sustainability, and equity. Moreover, the underlying optimization problems at the core of the grand challenges are of very large scale, many of which are beyond the scope of existing technologies. To address these, AI4OPT is organized around methodology thrusts that focus on specific challenges: they include a new generation of data-driven optimization solvers, decision making under uncertainty, combinatorial and reinforcement learning, end-to-end optimization, and decentralized learning and optimization. In addition, a transversal thrust on Ethical AI ensures that ethics is included in the design of every fundamental and use-inspired project, not as an afterthought. The complementarity of the end-use cases and the methodology thrusts creates a virtuous cycle of innovation, both in foundational research and industrial impact.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.

The research mission of AI4OPT is complemented by its educational vision which is

*to create longitudinal pathways for AI in engineering, from high school to graduate education, using a “teach the teachers” philosophy to maximize impact.*

The pathways start in middle and high schools (through summer camps and engineering practices), move to undergraduate education through the Faculty Training Program, and the creation of graduate programs. The “teaching the teachers” philosophy is pervasive across the institute: its goal is to empower teachers at every education level to create programs in their own institutions, for example, minors and majors in artificial intelligence. AI4OPT also includes mentorship programs to help students take leadership roles and reinforce their understanding of the material over time. The institute has a strong focus on historically black high schools, Historically Black Colleges and Universities (HBCUs), and Minority Serving Institutions (MSIs), to develop talent and increase the diversity of the AI workforce.

AI4OPT is led by the Georgia Institute of Technology in collaboration with universities in California (UC Berkeley, USC, UC San Diego), Texas (UT Arlington), and Georgia (Clark Atlanta University). The Institute is creating a vibrant nexus in AI and optimization, bringing together academic institutions, industrial partners, international collaborators, and educators. In Atlanta, AI4OPT is located on the 12th floor of the Coda building in Midtown, providing a prime space for faculty, research scientists, and students that encourages knowledge cross-fertilization. AI4OPT brings together world-leading researchers in artificial intelligence, operations research and their integration, as well as domain experts in its use cases, creating a unique environment for innovation.

The Industrial Partner Program (IPP) of AI4OPT features novel internship programs to facilitate research collaborations between academia and industry. It assembles some of the most innovative companies in supply chains, manufacturing, and energy systems, with the goal of maximizing the impact emerging from the fusion of AI and optimization.

AI4OPT has metrics of success that go beyond the traditional academic indicators (e.g., publications, keynotes, awards). On the research side, success is measured by *demonstrating how AI-assisted optimization can solve applications that are outside the realm of existing technology*. Real-time optimization and human-in-the-loop optimization are two classes of applications where the Institute has already achieved success. On the education side, success is about creating minor and major programs in AI and building AI research capacity at HBCUs and MSIs,

as well as developing camps and courses for high school students.

The rest of this article is organized as follows. “Optimization Proxies” section illustrates how the institute contributes scientific breakthroughs that the two fields cannot achieve independently. “End-Use Cases” section describes some of the societal and technological challenges that are driving AI4OPT research. “Methodology Thrusts” section briefly reviews the methodology thrusts driven by the end-use cases. “Workforce Development” and “AI4OPT as a Nexus” sections outline some workforce development activities and how AI4OPT acts as a nexus at the intersection of AI and optimization. It is impossible to do justice to all the activities of the institute in a short article, but the hope is that this presentation will encourage readers to learn more about AI4OPT.

## OPTIMIZATION PROXIES

At its core, optimization models are used in decision-making applications to map problem inputs into optimal solutions. Optimization solvers have seen dramatic progress over the last decades, producing optimal solutions to many industrial problems. For instance, optimization models quite literally keep the lights on by committing generators and dispatching electricity in real time every five minutes. Optimization models run end-to-end supply chains, which involve aspects such as the scheduling of manufacturing plants, load consolidation for middle-mile logistics, and the design of e-commerce networks, to name a few. Yet recent developments are challenging even the best solvers: optimization models have grown even larger, are expected to capture the realities of an increasingly uncertain and volatile world, and are blurring the distinction between planning and operations. As a result, optimization solvers have become too slow in many contexts: they include real-time applications, large-scale Monte-Carlo simulations that are based on optimization, and environments where humans interact with optimization technology.

In those circumstances, it is natural to explore whether machine learning can replace optimization, moving most of the computational burden offline. A machine-learning model can then learn the input/output mapping of the optimization model, producing a first approximation to the concept of an *optimization proxy*. The challenge, however, comes from applying this idea to *AI for Engineering*. Indeed, many of the end-use cases of the institute feature optimization problems with hard physical, engineering, and business constraints. For instance, in an electrical grid, the load (demand) and the generation (supply) must be equal at all times. In supply chains, shipments must fit within the vehicle capacity. In addition, optimization



proxies will be deployed in high-stake applications, which means that they must cater to a wide variety of instances, deliver high-quality solutions with performance guarantees, and learn instances with millions of input parameters and make hundreds of thousands of predictions.

To address these considerations, the *end-to-end optimization* thrust explores the science and engineering of optimization proxies. Its research led to novel architectures, which postulate that an optimization proxy is the composition of a machine-learning layer that produces a high-quality approximation of the optimization model, followed by a feasibility layer that repairs the prediction to deliver a feasible, near-optimal, solution. This *Learning and Repair* architecture can be trained end-to-end, using a self-supervised framework that backpropagates the loss function through the feasibility layers.

This fusion of AI and optimization can deliver breakthroughs that cannot be achieved by the two fields independently. One such application is real-time risk assessment, which is becoming increasingly pervasive in energy systems and supply chains, including for some of the key partners of AI4OPT. As mentioned earlier, transmission system operators typically solve a market-clearing optimization every 5 min to balance load and generation. Given the increased volatility in net load due to intermittent renewable resources, grid operators are interested in real-time risk assessment tools. Such tools run a large number of Monte-Carlo simulations, using scenarios from probabilistic forecasts, to quantify the system-wide risk. A single simulation may take up to 15 min, given the computational complexity and the sheer number of optimization problems, making it impractical to assess risk in real time. Work by the institute has shown that optimization proxies make it possible to run these simulations in 5 s (two to three orders of magnitude faster), identifying risk with very high accuracy (Chen, Tanneau, and Van Hentenryck 2023). Real-time risk assessment represents one of the early successes of AI4OPT, demonstrating the true potential of fusing AI and optimization. Such orders of magnitude improvements are one of the key success indicators of the Institute. *Observe that the engineering applications raise challenges in speed, reliability, scalability, robustness, and performance guarantees that are pushing AI in directions that are largely unexplored.* They require fundamental scientific and engineering advances to ensure feasibility, near-optimality, and reliability.

## END-USE CASES

As mentioned, already, at AI4OPT, challenges from end-use cases inspire foundational research, which then delivers

innovations to address them. Here is a brief description of these end-use cases.

### *Supply chains*

Supply-chain management used to be an arcane topic, discussed by a few and invisible to the general public. This has changed after the pandemic: the public is now aware of a topic that has become top-of-mind in many corporate boards. Supply chains have become larger, and e-commerce has proliferated, imposing significant environmental costs to meet new customer expectations. At the same time, many customers and suppliers, especially in rural regions, face increasing difficulties in procuring or delivering specific products. *What is needed is a paradigm change, a new vision for supply chains that complements efficiency with resilience, sustainability, and equity goals.* Research in supply chains at AI4OPT is centered around end-to-end supply chains, with scalability, resilience, sustainability, and equity as core challenges. AI4OPT has assembled a consortium of partners that cover (almost) all aspects of supply chains. It leverages novel forecasting methods, optimization proxies, decision making under uncertainty, and automation to meet these challenges.

### *Energy systems*

The challenge for energy systems is clear: *how to reinvent the planning and operations of a grid powered by renewable energy sources and storage.* Energy systems are transitioning from the century-old “generation follows the load” organization of the grid to a paradigm centered on *risk assessment and risk management.* This end-use case helps reinvent energy systems by pursuing four overarching themes: (1) probabilistic forecasting to quantify uncertainty; (2) stochastic and risk-aware optimization to capture this uncertainty in decision processes; (3) optimization proxies to perform real-time risk assessment and risk-aware optimization; and (4) decentralized optimization to address the massive proliferation of distributed energy resources.

### *Chip design and manufacturing*

Each generation of chips is becoming more expensive to design, requiring numerous cycles between expert designers and simulators. It is no longer possible or desirable to separate the various phases of the design, for example, circuit synthesis, placement, and routing. *What is needed is a new generation of tools that help engineers design circuits more holistically.* Machine learning (including RL and inverse learning) has been shown to have significant potential in this area; it is the goal of this end-use case to explore its role in circuit optimization. Chip manufacturing has also become increasingly complex and subject to uncertainty in supply, demand, and the complexity of the

bill of materials. In conjunction with supply chains, this end-use case also aims at transforming the optimization of the chip manufacturing process.

### *Sustainable systems*

The food–energy–water nexus is identified as one of the key grand challenges of the 21st century and AI has demonstrated early potential to address complex problems in this space. This end-use case conducts three interconnected projects on biogas, water, and food to reduce greenhouse emissions and boost food production.

## METHODOLOGY THRUSTS

The methodology thrusts carry out foundational AI research inspired by the end-use cases. Here is a brief description of the methodology thrusts of AI4OPT.

### *End-to-end optimization*

The end-to-end optimization thrust primarily focuses on the *science and engineering of optimization proxies* that were described in “Optimization Proxies” section. Recent contributions include the concepts of self-supervised primal-dual learning (Park and Van Hentenryck 2023), compact learning (Park et al. 2023), and end-to-end learning and repair (Chen, Tanneau, and Van Hentenryck 2023). The thrust draws inspiration from the end-use cases in energy systems and supply chains, and also explores topics in decision-focused learning, learning to optimize, verification, explanation, and formal guarantees.

### *New generation solvers*

The solvers thrust works on a new generation of highly tunable optimization solvers that use machine learning and historical data to dramatically improve performance in settings where an optimization model is used repeatedly. Recent results apply the *Learning to Optimize* paradigm to mixed-integer programming (Huang, Ferber, et al. 2023), mixed-integer nonlinear programming (Ferber et al. 2023), and AI planners (Huang, Shivashankar, et al. 2023). The end-use cases contribute problem instances to benchmark solver performance.

### *Decision making under uncertainty*

The energy and supply-chain end-use cases clearly indicate the need for advances in decision making under uncertainty. This includes probabilistic forecasting, uncertainty quantification, scenario generation, and detection of rare events in presence of spatial-temporal correlations (Xu and Xie 2023). Of particular interest is the fundamental and applied research on conformal predictions. The thrust also explores solution techniques for specific classes of multistage stochastic optimization problems (Lan and Shapiro 2023) and new Bayesian risk-sensitive

and distributionally robust optimization models (Ju and Lan 2023).

### *Reinforcement learning*

The RL thrust focuses on the end-use cases of the institute, which are much larger and more complex than environments in which RL has been successful so far. It contributes foundational advances to deep RL to handle such complex environments (Laskin et al. 2022), and expands RL research to include societal and ethical considerations. The thrust will be increasingly focused on *offline RL* (Chen and Maguluri 2022) to make the technology safer and more amenable to industrial use.

### *Combinatorial learning*

This thrust studies machine learning in the context of combinatorial and highly constrained applications with the goal to improve generalization and interpretability and reduce errors. Recent results include highly specific strong convex models with structured sparsity (Atamtürk and Gómez 2022), pairwise-based optimization algorithms for counteracting learning bias (Hochbaum, Liu, and Goldschmidt 2023), and meta-algorithms to automatically select the best solver (Asín-Achá et al. 2022).

### *Distributed and multi-agent learning and optimization*

This thrust explores decentralized solutions to manage a large number of agents, motivated by applications in energy systems and automated warehouses. Recent results focus on AI-based decentralized path planning and execution (Xu et al. 2022), on how agents can help each other learn through communication (Zhang, Pananjady, and Romberg 2022), and distributed learning algorithms that are robust against communication imperfections (Zeng, Doan, and Romberg 2023).

### *Ethical AI*

The ethical AI thrust is transversal: it draws from, and informs, every thrust and end-use case in the institute. It leverages the fusion of optimization and machine learning to ensure *ethical and socially conscious design* of large scale deployments. Projects include creating new theoretical foundations for ethics in practice (Gupta, Moondra, and Singh 2023), including fairness into supply chains and energy networks (Hettle, Gupta, and Molzahn 2021), technological progress for high-impact policy changes (Gillani et al. 2023), and the incorporation of IEEE Well-being Metrics into the design and deployment of AI and optimization research.

## WORKFORCE DEVELOPMENT

The education and workforce development initiatives of AI4OPT are presented in detail by Pierre et al. (2022) and are



only briefly mentioned here. Perhaps the most distinctive feature of these programs is the “teaching the teachers” *philosophy* that permeates the initiatives. AI4OPT is reaching middle and high school students through the Seth Bonder summer camps that are delivered, not only to students, but also to high school teachers. The camps are organized at Georgia Tech, at UC Berkeley, and online (in collaboration with Kids Teach Tech), and attract a diverse range of participants, with a large proportion of minority students and young women. Students who successfully complete the camp are invited to become mentors in the following year.

The AI4OPT Faculty Training Program (FTP) provides faculty members from HBCUs and MSIs with courses in AI, data science, and course design to create minors and majors in AI at their own institutions. This 3-year program includes a yearly 3–4-week visit to Georgia Tech and online courses throughout the year. The program started in June 2022, and multiple FTP participants are already creating minors and majors, and are working with AI4OPT to expand their AI offerings.

## AI4OPT AS A NEXUS

AI4OPT acts as a nexus for AI and optimization with applications in supply chains, energy systems, manufacturing, and sustainability. The institute nexus is organized around its research and educational programs, its IPP, its national and international collaborations, and its outreach activities. In outreach, AI4OPT is pursuing additional partnerships, for example, with Jackson State University in Mississippi and Huston-Tillotson University in Austin, Texas. The IPP continues to grow and covers the entire spectrum of activities in end-to-end supply chains and the planning and operations of electrical power systems. AI4OPT features novel longitudinal internships that are piloted by the Institute at Georgia Tech, and strong collaborations with DOE national laboratories, Independent Systems Operators (ISO), and peer international institutions. A particularly exciting development is the collaboration with the AI Institute ICICLE around decentralized collaborative multimodal food supply chains with a focus on tribal communities.

## CONCLUSION

This article presented AI4OPT, an NSF AI Institute carrying out foundational research on the fusion of AI and optimization. The methodology research emerging from AI4OPT is inspired by fundamental societal challenges in supply chains, energy systems, chip design and manufacturing, and sustainable food systems. The institute is

demonstrating that this fusion can deliver tools for these applications that are orders of magnitude faster and that can be used in real time, using the concept of optimization proxies. AI4OPT expects this fusion to expand in scope, both in methodology and application domains. In addition, the Institute will continue building its longitudinal pathways for democratizing AI education in high schools and minority serving institutions.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National AI Research Institutes program supported by NSF and Intel under Award No. 2112533. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Pascal Van Hentenryck  <https://orcid.org/0000-0001-7085-9994>

Kevin Dalmeijer  <https://orcid.org/0000-0002-4304-7517>

## REFERENCES

- Asín-Achá, Roberto, Olivier Goldschmidt, Dorit S. Hochbaum, and Isaías Huerta. 2022. “Fast Algorithms for the Capacitated Vehicle Routing Problem Using Machine Learning Selection of Algorithm’s Parameters.” In *International Conference on Knowledge Discovery and Information Retrieval*, 29–39. <https://doi.org/10.5220/0011405400003335>.
- Atamtürk, Alper, and Andrés Gómez. 2022. “Supermodularity and Valid Inequalities for Quadratic Optimization with Indicators.” *Mathematical Programming* 201(1-2): 295–338. <https://doi.org/10.1007/s10107-022-01908-2>.
- Chen, Wenbo, Mathieu Tanneau, and Pascal Van Hentenryck. 2023. “End-to-End Feasible Optimization Proxies for Large-Scale Economic Dispatch.” *IEEE Transactions on Power Systems*. <https://doi.org/10.1109/TPWRS.2023.3317352>.
- Chen, Zaiwei, and Siva Theja Maguluri. 2022. “Sample Complexity of Policy-Based Methods under Off-Policy Sampling and Linear Function Approximation.” In *International Conference on Artificial Intelligence and Statistics*, 11195–214.
- Ferber, Aaron, Taoan Huang, Daochen Zha, Martin Schubert, Benoit Steiner, Bistra Dilkina, and Yuandong Tian. 2023. “SurCo: Learning Linear SURrogates for Combinatorial Nonlinear Optimization Problems.” In *International Conference on Machine Learning*, 10034–52.
- Gillani, Nabeel, Doug Beeferman, Christine Vega-Pourheydarian, Cassandra Overney, Pascal Van Hentenryck, and Deb Roy. 2023. “Redrawing Attendance Boundaries to Promote Racial and Ethnic Diversity in Elementary Schools.” *Educational Researcher* 52(6): 348–64. <https://doi.org/10.3102/0013189X231170858>.
- Gupta, Swati, Jai Moondra, and Mohit Singh. 2023. “Which Lp Norm is the Fairest? Approximations for Fair Facility Location Across

- All “p.” *ACM Conference on Economics and Computation*, 817. <https://doi.org/10.1145/3580507.3597664>.
- Hettle, Cyrus, Swati Gupta, and Daniel K. Molzahn. 2021. “Fair and Reliable Reconnections for Temporary Disruptions in Electric Distribution Networks using Submodularity.” *CoRR, abs/2104.07631*.
- Hochbaum, Dorit S., Zhihao Liu, and Olivier Goldschmidt. 2023. “A Breakpoints Based Method for the Maximum Diversity and Dispersion Problems.” In *SIAM Conference on Applied and Computational Discrete Algorithms*, 189–200. <https://doi.org/10.1137/1.9781611977714.17>.
- Huang, Taoan, Aaron Ferber, Yuandong Tian, Bistra Dilkina, and Benoit Steiner. 2023. “Searching Large Neighborhoods for Integer Linear Programs with Contrastive Learning.” In *International Conference on Machine Learning*, 13869–90.
- Huang, Taoan, Vikas Shivashankar, Michael Caldara, Joseph Durham, Jiaoyang Li, Bistra Dilkina, and Sven Koenig. 2023. “Deadline-Aware Multi-Agent Tour Planning.” In *International Conference on Automated Planning and Scheduling*, 33, 189–97. <https://doi.org/10.1609/icaps.v33i1.27194>.
- Ju, Caleb, and Guanghui Lan. 2023. “Dual Dynamic Programming for Stochastic Programs Over an Infinite Horizon.” <https://api.semanticscholar.org/CorpusID:257353385>.
- Lan, Guanghui, and Alexander Shapiro. 2023. “Numerical Methods for Convex Multistage Stochastic Optimization.” *arXiv:2303.15672*.
- Laskin, Michael, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. 2022. “Unsupervised Reinforcement Learning with Contrastive Intrinsic Control.” In *Neural Information Processing Systems Conference*, 34478–91.
- Park, Seonho, and Pascal Van Hentenryck. 2023. “Self-Supervised Primal-Dual Learning for Constrained Optimization.” *Proceedings of the AAAI Conference on Artificial Intelligence* 37(4): 4052–60. <https://doi.org/10.1609/aaai.v37i4.25520>.
- Park, Seonho, Wenbo Chen, Terrence W. K. Mak, and Pascal Van Hentenryck. 2023. “Compact Optimization Learning for AC Optimal Power Flow.” *IEEE Transactions on Power Systems*, forthcoming.
- Pierre, Charles, Pascal Van Hentenryck, Kevin Dalmeijer, and Tuba Ketenci. 2022. “The Longitudinal Education Programs of AI4OPT.” *OR/MS Today* 49(4). <https://doi.org/10.1287/orms.2022.04.12>.
- Xu, Chen, and Yao Xie. 2023. “Sequential Predictive Conformal Inference for Time Series.” In *International Conference on Machine Learning*, 38707–27.
- Xu, Qinghong, Jiaoyang Li, Sven Koenig, and Hang Ma. 2022. “Multi-Goal Multi-Agent Pickup and Delivery.” In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 9964–71. <https://doi.org/10.1109/iroso47612.2022.9981785>.
- Zeng, Sihan, Thinh T. Doan, and Justin Romberg. 2023. “Connected Superlevel Set in (Deep) Reinforcement Learning and its Application to Minimax Theorems.” *arXiv:2303.12981*.
- Zhang, Sheng, Ashwin Pananjady, and Justin Romberg. 2022. “A Dual Accelerated Method for a Class of Distributed Optimization Problems: From Consensus to Decentralized Policy Evaluation.” *2022 IEEE 61st Conference on Decision and Control (CDC)*, Cancun, Mexico, 2022, pp. 5220–25. <https://doi.org/10.1109/CDC51059.2022.9993276>.

**How to cite this article:** Van Hentenryck, P., and K. Dalmeijer. 2024. “AI4OPT: AI Institute for Advances in Optimization.” *AI Magazine* 45: 42–47. <https://doi.org/10.1002/aaai.12146>

## AUTHOR BIOGRAPHIES

**Pascal Van Hentenryck** is the director of AI4OPT. He is an A. Russell Chandler III Chair and Professor in industrial and systems engineering at Georgia Tech. A Computer Scientist by training, Van Hentenryck currently focuses on the integration of machine learning and optimization for engineering applications. He is a fellow of AAAI and INFORMS.

**Kevin Dalmeijer** is the Managing Director of AI4OPT and a Senior Research Associate in industrial and systems engineering at Georgia Tech. Dalmeijer has a background in operations research and works on challenging problems in mobility, transportation, and logistics.





SPECIAL TOPIC ARTICLE

# AI Institute in Dynamic Systems: Developing machine learning and AI tools for scientific discovery, engineering design, and data-driven control

J. Nathan Kutz<sup>1</sup> | Steven L. Brunton<sup>1</sup> | Krithika Manohar<sup>1</sup> | Hod Lipson<sup>2</sup> | Na Li<sup>3</sup>

<sup>1</sup>University of Washington, Seattle, Washington, USA

<sup>2</sup>Columbia University, New York, New York, USA

<sup>3</sup>Harvard University, Cambridge, Massachusetts, USA

## Correspondence

J. Nathan Kutz, University of Washington, Seattle, WA 98195, USA.  
Email: [kutz@uw.edu](mailto:kutz@uw.edu)

## Funding information

United States National Science Foundation, Directorate of Engineering, Grant/Award Number: 2112085

## Abstract

The mission of the AI Institute in Dynamic Systems is to develop the next generation of advanced machine learning (ML) and AI tools for controlling complex physical systems by discovering physically interpretable and physics-constrained data-driven models through optimal sensor selection and placement. The research effort is anchored by a common task framework (CTF) that evaluates the performance of ML algorithms, architectures, and optimization schemes for the diverse tasks required in engineering applications. The aim is to push beyond the boundaries of modern techniques by closing the loop between data collection, control, and modeling, creating a unique and cross-disciplinary architecture for learning physically interpretable and physics constrained models of complex dynamic systems from time series data. The CTF further supports sustainable and open-source challenge datasets, which are foundational for developing interpretable, ethical, and inclusive tools to solve problems fundamental to human safety, society, and the environment.

The emergence of machine learning (ML) and AI algorithms is transforming every scientific and engineering discipline. The widespread adoption of ML/AI algorithms represents the second computational revolution. As in the early 1990s, the widespread availability of computers allowed for significant advancements in science and engineering by allowing for numerical proxies for modeling physical systems. Similarly, ML/AI is allowing for the modeling and control of physical systems directly from the quantity and quality of data acquired from emerging sensor technologies. The goal of the AI Institute in Dynamic Systems is to develop the next generation of advanced ML

tools for controlling complex physical systems by discovering physically interpretable and physics-constrained data-driven models through optimal sensor selection and placement. The work is anchored by a common task framework (CTF), which is a suite of challenge data sets with a diverse set of objectives relevant to engineering design and control. Thus the institute evaluates the performance of ML algorithms and allows us to push beyond the boundaries of modern techniques by closing the loop between data collection, control, and modeling, thus creating a unique and cross-disciplinary architecture for learning physically interpretable and physics constrained models of complex

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.

dynamic systems from time series data. The CTF further supports sustainable and open-source challenge datasets, which are foundational for developing interpretable, ethical, and inclusive tools to solve problems fundamental to human safety, society, and the environment.

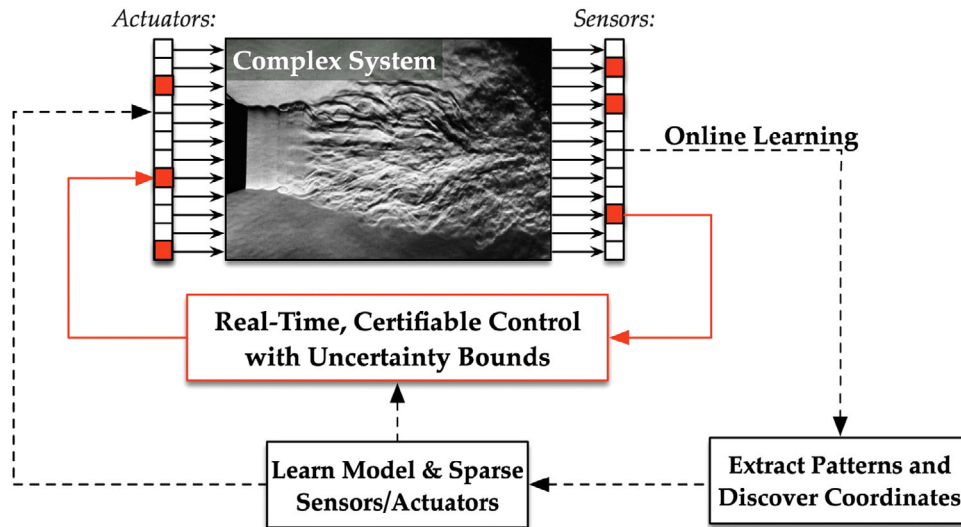
As such, the AI Institute develops the fundamental mathematical and computational tools required to empower AI for real-time sensing (Manohar et al. 2018), learning (Chen et al. 2022; Fasel et al. 2022), decision making (Hu et al. 2023), and prediction capabilities that lead the way towards safe, reliable, efficient, explainable, and ethical data-enabled engineering and science systems. The institute is organized around three key thrusts that are used in synergistic partnership for solving today's grand-challenge problems. Thrust 1 is centered around sensing and data-acquisition, which is required of all AI agents engineered to perform in real-world applications. This thrust also builds fundamental optimization tools, which underpin all ML and AI learning algorithms (Manohar et al. 2018). Thrust 1 enables Thrust 2, which aims to build a diversity of dynamic models directly from sensor data (Chen et al. 2022; Fasel et al. 2022; Gao and Kutz 2022). With the development of dynamic models, data-driven control protocols are developed in Thrust 3 (Hu et al. 2023). The understanding of models and controls then re-integrates with sensing where better sensing strategies can be constructed with knowledge from Thrusts 2 and 3. This creates a virtuous cycle of data-driven integration strategies for broad application across science and engineering. Critically, all thrusts integrate around the CTF, which allow the institute to test, evaluate, and compare diverse algorithms for executing Thrusts 1–3 in application areas, which are relevant to the science and engineering community. The centrality of the CTF to the institute allows the broader community access to data, algorithms, and innovations. As such, it creates a nexus point for scientific activity.

Based at the University of Washington in Seattle, the AI Institute team is geographically centered across the greater Pacific Northwest region, while maintaining key connections to east coast technology hubs. This includes in the Northwest the University of Nevada Reno, the University of Hawaii Manoa, Montana State University, Portland State University, the University of Alaska, and Boise State University. In addition, we are partnered with Harvard University and Columbia University. The geographic connections help facilitate active collaboration among team members through a shared sense of purpose, while remaining connected to other hubs. A long term goal for the institute is to serve as a nexus point for facilitating the rapid transfer of intellectual ideas and cutting edge educational opportunities in AI for dynamic systems from established technology hubs.

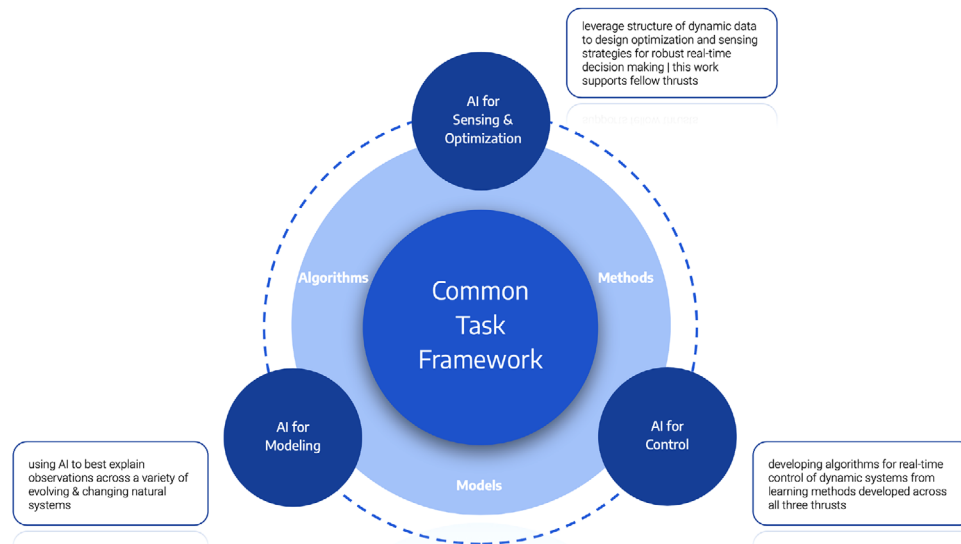
More broadly, the institute is developing the three critical efforts necessary for driving forward AI for dynamic systems: (i) the mathematical foundations of AI, (ii) grand challenge applications for AI, and (iii) a transformational educational and workforce development infrastructure for AI Engineering. Physics-informed ML is emerging as a leading paradigm for bringing together AI and engineering dynamic systems, providing new capabilities in real-time sensing, learning, decision making, and predictions that are safe, reliable, efficient, ethical, and imbued with uncertainty quantification (UQ). The AI Institute team is developing a general and flexible ML framework to rapidly learn new physics, enforce known physical constraints, and discover them directly. Methods are continuously evaluated through demonstrations of real-world grand challenge applications, so that the methods are physically motivated and catalyze advances across multiple applied domains. To facilitate reproducibility and robustness across disciplines, the institute has developed the CTF that leverages SageBionetworks, a nonprofit company dedicated to hosting challenge data sets, for evaluating ML/AI algorithms on challenge data sets in physics and engineering, thus providing a broad service to the engineering ML community. All methods and datasets are made available to the public through actively maintained open source software packages.

**Research:** The foundations of physics-informed ML are rooted in four key disciplines: (i) control theory, (ii) probability and statistics, (iii) optimization, and (iv) dynamical systems (modeling). The integration of all four of these disciplines is critical for the development of ML algorithms that can be leveraged by engineered systems, as illustrated in the integration of methods in Figure 1. Establishing rigorous mathematical connections between these disciplines is a driving inspiration for the institute efforts in reframing the foundations of ML for the engineering sciences. A key step in moving the field forward is to address the current lack of educational infrastructure around training, which integrates these four core background areas, especially as they are applied to pressing challenges in engineering dynamic systems. The institute trains students and postdocs to have deep expertise in both the core mathematical architectures as well as domain applications through the development of new open-source and multimodal undergraduate and graduate curricula.

Given the institute objectives, efforts are organized around three key thrusts that are used in synergistic partnership. Figure 2 shows how each thrust is critical for all other thrusts. Thrust 1 is built around sensing and data-acquisition, which is required of all AI agents engineered to perform in real-world applications. This thrust also builds fundamental optimization tools, which underpin all ML and AI learning algorithms. Thrust 1 enables Thrust 2,



**FIGURE 1** Overview of methodological innovations proposed for physics-informed machine learning for dynamic systems by the AI Institute in Dynamic Systems.



**FIGURE 2** Integration and partnering of thrusts. Each thrust is integral and critical to each other thrust in the overall sensing, modeling, control paradigm for AI integration.

which aims to build a diversity of dynamic models directly from sensor data. With the development of dynamic models, data-driven control protocols are developed in Thrust 3. The understanding of models and controls then re-integrates with sensing where better sensing strategies can be constructed with knowledge from Thrusts 2 and 3. This creates a virtuous cycle of data-driven integration strategies for broad application across science and engineering. Critically, all thrusts integrate around the CTF, which allow us to test, evaluate, and compare diverse algorithms for executing Thrusts 1–3 in application areas which are relevant to the science and engineering community.

**Thrust 1: AI for sensing and optimization:** Much of the power of modern AI derives from our ability to fit highly expressive models to large sets of high-dimensional data, by numerical optimization. Physics and dynamics-informed AI poses additional challenges beyond the standard issues of scale and dimensionality: the goal is to identify accurate physical models that support downstream control and decision making with guaranteed performance. This goes well beyond the standard demands of ML/AI, where the optimization goal is often just to accurately label a given training set. Physical data are structured: individual signals are often sparse in an

appropriate basis, batches of signals are often low-rank, while data generated by systems with a few important underlying degrees of freedom typically concentrate near low-dimensional manifolds (Champion et al. 2019; Chen et al. 2022). From a computational perspective, correctly inferring these low-dimensional models in noisy, incomplete, or otherwise unreliable data typically requires us to solve a high-dimensional nonconvex optimization problem, to global optimality. This requires new perspectives and computational schemes to address the emerging challenges posed for optimization in the context of complex dynamical systems. Thus, the institute has leveraged the structure of dynamic data to design guaranteed optimization and sensing strategies for safe, efficient, robust real-time decision making (Hu et al. 2023; Manohar et al. 2018).

**Thrust 2: AI for physics-informed models:** A significant strength of the institute team is in learning physically interpretable models of dynamical systems from off-line and/or on-line streaming data (Baddoo et al. 2023; Brunton, Proctor, and Kutz 2016; Champion et al. 2019; Chen et al. 2022; Gao and Kutz 2022; Schmidt and Lipson 2009). Physics informed learning is of growing importance for scientific and engineering problems. Physics-informed simply refers to the ability to constrain the learning process by physical and/or engineering principles. For instance, conservation of mass, momentum, or energy can be imposed in the learning process. In the parlance of ML, the imposed constraints are referred to as regularizers. Thus, physics informed learning focuses on adding regularization to the learning process to impose or enforce physical priors. Known physics (e.g., invariances, symmetries, conservation laws, constraints, etc.) may be incorporated at various stages of the learning process. For example, rotational invariance is often incorporated by augmenting the training data with rotated copies, and translational invariance is often captured using convolutional neural network architectures. In kernel-based techniques, such as Gaussian process regression and support vector machines, symmetries can be imposed by means of rotation-invariant, translation-invariant, and symmetric covariance kernels. Additional physics and prior knowledge may be incorporated as additional loss functions or constraints in the optimization problem. Thus, the institute leverages data and AI to automatically discover the underlying mechanisms and mathematical representations that best explain observations across a variety of evolving/changing natural systems. These mathematical formulations are being used to gain insights, make predictions, and discover governing equations and coordinate systems in a variety of application areas already (Baddoo et al. 2023; Champion et al. 2019; Chen et al. 2022; Fasel et al. 2022; Gao and Kutz 2022).

**Thrust 3: AI for data-driven control:** The remarkable successes of ML can be also leveraged towards the control of modern complex dynamical systems. Reinforcement learning (RL) is a class of ML that addresses the problem of learning to control physical systems by explicitly considering their inherent dynamical structure and feedback loop. To date, the successes of RL have been limited to very structured or simulated environments, and its successes in real-world systems are few. RL has significant challenges involving: (i) Scalability: How to develop scalable RL methods for large-size network multiagent (MARL) dynamical systems? (ii) Robustness: How to maintain the performance (efficiency and safety) of the learned policies even when there is a model class mismatch? (iii) Safety: How to guarantee RL maintains stability and stays in the safe constraint while still learning efficiently? The institute is leading the developing of critically enabling mathematical and computational architectures to overcome these challenges which are present in a diverse number of applications involving complex dynamical systems. Thus, the institute is innovating algorithms for real-time control of dynamic systems for their safe, reliable, and efficient operation leveraging principles from model-based and model-free control, optimization, and learning methods developed across all the three thrusts of modeling, control, and sensing/optimization. Institute partners are already establishing the theoretical foundations of policy optimization for learning control policies, thus revolutionizing how modern RL control can be embedded in engineering systems (Hu et al. 2023).

**Education in AI:** The institute has a significant and focused effort on open-source educational courses and algorithms in AI for science and engineering, including the development of open source packages for engineering AI applications (de Silva et al. 2021; Kaptanoglu et al. 2022; Pan et al. 2023; Van Breugel et al. 2022). This serves as the foundation for nurturing and developing the next generation of talent in the sciences. Through these initiatives, we aim to support the growth and learning of individuals interested in this field, ensuring a continuous pipeline of skilled professionals, and advance innovations broadly in academia and industry. To efficiently disseminate the knowledge and tools that the institute is developing, the institute will provide the mathematical and software foundations necessary for diverse students to contribute to new AI for dynamic systems research. The CTF helps promote an active and broad engagement, with code and data openly available to student, postdocs, and professionals in the field overall. More broadly, to reach a broad and diverse community, our education plan integrates active learning classroom experiences with state-of-the-art online lectures, bootcamps, and workshops to teach students from K-12 through graduate school both fundamentals and



practical skills. All members of the institute are involved with the educational and workforce development efforts, providing educational materials that leverage each AI Institute members strengths. This institute effort enriches the broader community along with providing the best and most complete open-source educational materials in the engineering and physical sciences for graduate students, undergraduate students, and postdocs alike. Indeed, the overarching goal of the institute is to design and implement a comprehensive education and training plan that integrates ML and artificial intelligence seamlessly into undergraduate and graduate engineering curriculum. This involves identifying existing and desired skills and objectives and developing modular curriculum content to include in existing and new courses. This content is actively deployed within the institute and made available more broadly for other institutions and industry partners.

In addition to academic research efforts and partnerships, the institute is committed to building bridges to industry and national laboratories whose futures depend on the successful integration of ML/AI algorithms in the design, execution, and control of their emerging technologies. The institute has a large number of existing partnerships with traditional engineering companies to modern tech companies, balancing engineering design with state-of-the-art ML. Partnerships also exist with national laboratories who have developed significant efforts around physics-informed ML with applications to national interest.

**Summary:** From autonomy to turbulence, the design of computational algorithms capable of reliably controlling today's modern, high-dimensional complex systems remain a grand challenge scientific endeavor. Such systems are characterized by the manifestation of physical phenomena that can span multiple spatial and temporal scales and involve coupling between multiple types of physics, which are often unknown and unmeasured. Indeed, our current mathematical approaches to actuate such systems are insufficient to establish reliable and assured control paradigms. Emerging ML and artificial intelligence (AI) methods are transforming many technological landscapes due to their ability to flexibly and adaptively integrate measurement data into the modeling framework. In regards to engineered dynamic systems, ML and AI strategies are now being engineered by the AI Institute in Dynamic Systems to form the fundamental underpinnings of *physics-informed ML* strategies, which account for the constraints imposed by the underlying physics of the system, while providing greater flexibility to represent more complex physical phenomena. The institute's development of physics-informed ML holds tremendous promise for a new modeling paradigm to control modern dynamic systems. This integration also is offering trans-

formative impact in education and training by bridging the gap between traditional engineering disciplines and modern ML methods.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National AI Research Institutes program supported by NSF and DHS under Award No. 2112085. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

J. Nathan Kutz  <https://orcid.org/0000-0002-6004-2275>

## REFERENCES

- Baddoo, P. J., B. Herrmann, B. J. McKeon, J. Nathan Kutz, and S. L. Brunton. 2023. "Physics-Informed Dynamic Mode Decomposition." *Proceedings of the Royal Society A* 479(2271): 20220576.
- Brunton, S. L., J. L. Proctor, and J. N. Kutz 2016. "Discovering Governing Equations from Data by Sparse Identification of Non-linear Dynamical Systems." *Proceedings of the National Academy of Sciences* 113(15): 3932–37.
- Champion, K., B. Lusch, J. N. Kutz, and S. L. Brunton. 2019. "Data-Driven Discovery of Coordinates and Governing Equations." *Proceedings of the National Academy of Sciences* 116(45): 22445–51.
- Chen, B., K. Huang, S. Raghupathi, I. Chandratreya, Q. Du, and H. Lipson. 2022. "Automated Discovery of Fundamental Variables Hidden in Experimental Data." *Nature Computational Science* 2(7): 433–42.
- de Silva, B. M., K. Manohar, E. Clark, B. W. Brunton, J. N. Kutz, and S. L. Brunton. 2021. "PySensors: A Python Package for Sparse Sensor Placement." *Journal of Open Source Software* 6(58): 2828.
- Fasel, U., J. N. Kutz, B. W. Brunton, and S. L. Brunton. 2022. "Ensemble-SINDy: Robust Sparse Model Discovery in The Low-Data, High-Noise Limit, with Active Learning and Control." *Proceedings of the Royal Society A* 478(2260): 20210904.
- Gao, L., and J. N. Kutz. 2022. "Bayesian Autoencoders for Data-Driven Discovery of Coordinates, Governing Equations and Fundamental Constants." *Proceedings of the Royal Society A*. arXiv preprint arXiv:2211.10575.
- Hu, B., K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar. 2023. "Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies." *Annual Review of Control, Robotics, and Autonomous Systems* 6: 123–58.
- Kaptanoglu, A. A., B. M. de Silva, U. Fasel, K. Kaheman, A. J. Goldschmidt, J. Callahan, C. B. Delahunt, et al. 2022. "PySINDy: A Comprehensive Python Package for Robust Sparse System Identification." *Journal of Open Source Software* 7(69): 3994.
- Manohar, K., B. W. Brunton, J. N. Kutz, and S. L. Brunton. 2018. "Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating The Benefits of Exploiting Known Patterns." *IEEE Control Systems Magazine* 38(3): 63–86.

- Pan, S., E. Kaiser, B. M. de Silva, J. N. Kutz, and S. L. Brunton. 2023. "PyKoopman: A Python Package for Data-Driven Approximation of the Koopman Operator." arXiv preprint arXiv:2306.12962.
- Schmidt, M., and H. Lipson. 2009. "Distilling Free-Form Natural Laws from Experimental Data." *Science* 324(5923): 81–85.
- Van Breugel, F., Y. Liu, B. W. Brunton, and J. N. Kutz, 2022. "PyNumDiff: A Python Package for Numerical Differentiation of Noisy Time-Series Data." *Journal of Open Source Software* 7(71): 4078.

**How to cite this article:** Kutz, J. N., S. L. Brunton, K. Manohar, H. Lipson, and N. Li. 2024. "AI Institute in Dynamic Systems: Developing machine learning and AI tools for scientific discovery, engineering design, and data-driven control." *AI Magazine* 45: 48–53.  
<https://doi.org/10.1002/aaai.12159>

## AUTHOR BIOGRAPHIES

**J. Nathan Kutz** is the Robert Bolles and Yasuko Endo Professor of Applied Mathematics and Electrical and Computer Engineering at the University of Washington. He is also Director of the AI Institute in Dynamic Systems. His research interests include dynamical systems and machine learning with applications to diverse

scientific disciplines that include optics, neuroscience, and fluid dynamics.

**Steven L. Brunton** is the James Morrison Professor of Mechanical Engineering at the University of Washington. He is also an Associate Director of the AI Institute in Dynamic Systems. His research interests include machine learning, fluid dynamics, dynamical systems, and control.

**Krithika Manohar** is an Assistant Professor of Mechanical Engineering at the University of Washington. Her research interests include sensing and sensor placement, model reduction, and dynamical systems.

**Hod Lipson** is the Director of Columbia University's Creative Machines Lab. Lipson's work focuses on evolutionary robotics, design automation, rapid prototyping, artificial life, and creating machines that can demonstrate some aspects of human creativity.

**Na Li** is the Winokur Family Professor of Electrical Engineering and Applied Mathematics in the School of Engineering and Applied Sciences (SEAS) at Harvard University. Her research lies in control, learning, and optimization of networked systems, including theory development, algorithm design, and applications to real-world cyber-physical systems, such as energy systems, buildings, multirobots, and bio-medical systems.



## SPECIAL TOPIC ARTICLE

# The TILOS AI Institute: Integrating optimization and AI for chip design, networks, and robotics

Andrew B. Kahng<sup>1</sup> | Arya Mazumdar<sup>2</sup> | Jodi Reeves<sup>3</sup> | Yusu Wang<sup>2</sup>

<sup>1</sup>Departments of CSE and ECE, UC San Diego, La Jolla, California, USA

<sup>2</sup>Halıcıođlu Data Science Institute, UC San Diego, La Jolla, California, USA

<sup>3</sup>School of Technology and Engineering, National University, San Diego, California, USA

## Correspondence

Andrew B. Kahng, Departments of CSE and ECE, UC San Diego La Jolla, California, USA.

Email: [abk@ucsd.edu](mailto:abk@ucsd.edu)

Yusu Wang, Halıcıođlu Data Science Institute, 3234 Matthews Ln, La Jolla, CA 92093, USA.

Email: [yusuwang@ucsd.edu](mailto:yusuwang@ucsd.edu)

## Funding information

National AI Research Institutes program; NSF and Intel: CCF., Grant/Award Number: 2112665; Directorate for Computer and Information Science and Engineering

## Abstract

Optimization is a universal quest, reflecting the basic human need to *do better*. Improved optimizations of energy-efficiency, safety, robustness, and other criteria in engineered systems would bring incalculable societal benefits. But, fundamental challenges of scale and complexity keep many such real-world optimization needs beyond reach. This article describes The Institute for Learning-enabled Optimization at Scale (TILOS), an NSF AI Research Institute for Advances in Optimization that aims to overcome these challenges in three high-stakes use domains: chip design, communication networks, and contextual robotics. TILOS integrates foundational research, translation, education, and broader impacts toward a new nexus of optimization, AI, and data-driven learning. We summarize central challenges, early progress, and futures for the institute.

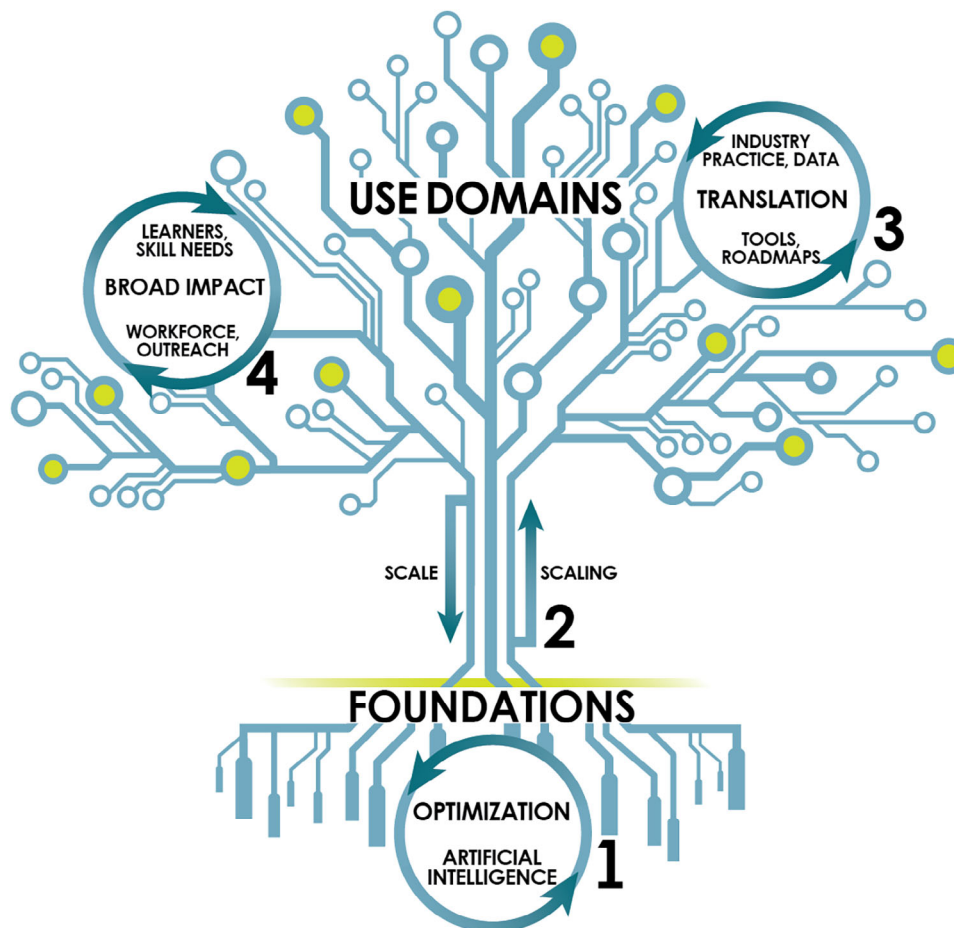
## INTRODUCTION

TILOS, The Institute for Learning-enabled Optimization at Scale, is an NSF AI Institute partially supported by Intel Corporation. The institute is a partnership of six universities: UC San Diego (lead), the Massachusetts Institute of Technology, National University, the University of Pennsylvania, the University of Texas at Austin, and Yale University. It began operations in November 2021, with a mission to “make impossible optimizations possible, at scale and in practice”. TILOS aims to discover a new integration of AI, optimization, and the leading edge of practice for three high-stakes use domains: chip design, communication networks, and contextual robotics.

These domains collectively underpin future innovations of information and communication technology, along with cyberphysical systems. Advancements in these domains critically hinge on better optimization: all these areas involve complex engineered systems with many pieces that need to be optimized individually and also holistically. For example, how can more circuits fit in a smaller region on the chip while using less energy? With the scale and complexity of these systems skyrocketing in the modern era, the real-world optimization needs have surpassed the reach of traditional methods. TILOS aims to use the power of AI to significantly accelerate optimization in these three use domains. At the same time, optimization is a key component of many AI frameworks: the

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Toward integration of AI/ML and optimization: four virtuous cycles in The Institute for Learning-enabled Optimization at Scale (TILOS).

training of a deep learning neural network fundamentally involves optimizing a highly non-convex system. Hence, TILOS will develop innovations that can fuse AI and optimization, propelling their symbiotic advancement in both foundations and applications.

The approach proposed by TILOS is reflected in Figure 1, which shows four *virtuous cycles* in the institute. First, *mutual* advances of AI and optimization provide our foundations. Second, the challenges of scale in practical optimization contexts, alongside the scaling breakthroughs achieved by new AI/ML and optimization methods, bind together foundations and the use domains of chip design, networks, and robotics. Third, the cycle of translation and impact must steadily bring research and the leading edge of practice closer together. Fourth, a cycle of research, education, and broadening participation must be established in order to grow a new AI-optimization-use nexus and its workforce. This is an ambitious agenda for just under 30 faculty, with a similar number of Ph.D. students and post-doctoral scholars. Thus, as TI-LOS members create these virtuous cycles, an added mindset is required: *How can we amplify our efforts, and optimize our impacts?* This mindset

shapes many basic processes, from collaboration mechanisms to curriculum creation for workforce development, to culture change (reproducibility, benchmarking, data-sharing, etc.) across entire research communities.

## FUNDAMENTAL RESEARCH

Fundamental research in TILOS is balanced across the foundations of AI/ML and optimization, and the three use domains.

### New foundations of learning and optimization

New, “modern” vistas for foundational research in optimization have opened up in the past decade, with the confluence of areas, emerging computational resources, and consequences of automation. The TILOS *Foundations* team, with 12 faculty from four institutions, focuses on five core research thrusts at the interface of AI and





optimization: (1) bridging continuous and discrete optimization; (2) parallel, distributed and federated optimization; (3) optimization on manifolds; (4) dynamic decision-making in uncertain environments; and (5) nonconvex optimization in deep learning. These thrusts are respectively motivated by five aspects of the modern context for optimization, as exemplified by our three core use-domains: (1) continuous models and methods are being deployed for inherently discrete problems; (2) distributed and federated models of data storage and algorithms are replacing centralized ones; (3) spaces with richer geometric structures than Euclidean (e.g., the configuration space of an articulated robot can be better modeled by a manifold) are increasingly used; (4) optimization must increasingly be performed in unknown and dynamic environments; and (5) nonconvex models and methods are required to explain and approach modern machine learning.

## Interplay at the interface of foundations and use

The use domains of chip design, networks, and robotics bring diverse optimization challenges but inspire shared solutions with commonalities such as physical embeddedness, hierarchical-system context, underlying graphical models, safety, and robustness as first-class concerns, and the bridging of human-guided and autonomous systems. Moreover, practical optimizations in each of these domains bring further, common challenges: (1) instances have enormous scale; (2) representations and abstractions are crucial to success; (3) objectives are hazy, particularly with multi-stage optimizations and dynamic settings; (4) optimization tools must provide reliability and generalization; and (5) scaling of productivity increasingly demands new ways to learn and optimize using modern compute fabrics.

Addressing the above challenges requires us to identify the right representations, develop the machine learning methods suitable for those representations, and leverage the interplay between learning and optimization. Indeed, representations should incorporate domain knowledge, structure in data (e.g., low-dimensional, non-Euclidean), as well as the mathematical structures behind the problems at hand. The right models and representations are fundamental to optimization performance and generalization. As optimization is a fundamental component of modern machine learning, while conversely learning can help solve difficult optimization problems, we need to better understand and leverage this close interplay, form new connections and interactions, and co-evolve both.

For example, TILOS team members' investigations have advanced fundamental understanding of the capacity of

graph neural networks (GNNs) in terms of representation learning and optimization (Jegelka, 2022). Examples of these advances include articulating the precise classes of functions that can be represented by GNNs, studying robustness via suitable graph distances, understanding how to better capture long-range interactions in large, sparse (hyper)graphs (critical for netlists in chip design), and developing new sign- and basis-invariant GNNs (which are crucial given the importance of meaningful position encoding in graph learning). TILOS team members are further developing more effective graph learning and optimization models for, for example, chip design applications (Kahng et al., 2024; Luo et al., 2024).

TILOS researchers have made fundamental progress on non-Euclidean optimization and sampling from Riemannian manifolds, providing a key step toward building the theory of computational complexity of solving stochastic differential equations on manifolds. Other highlights include a new framework for continuous neural set extensions to facilitate learning/optimization with discrete functions, progress on first-order methods for min-max optimization, and understanding of optimization dynamics of neural networks and the convergence of stochastic gradient descent. These are just a very small sample of examples: references of the aforementioned results, and many other research outcomes from TILOS can be found at (TILOS).

## Use-domain challenges

Fourteen faculty are engaged in TILOS use-domain research: six in robotics and five each in networks and chips, with two faculty jointly pursuing both networks and chips research. There is no shortage of unsolved problems and challenges; indeed, the institute's 5-year strategic plan includes nearly 60 distinct projects across 35 research topics.

*Robotics* exemplifies how AI and optimization connect to the physical world. TILOS robotics researchers pursue optimization and AI-based methodologies for manufacturing autonomous, adaptive, heterogeneous teams of sensors, robots, and intelligent machines that can work with humans. Key challenges include concurrency, online processing, and minimizing the need for labeled data. This gives rise to three major research thrusts: (1) metric, semantic, and dynamic artificial perception in a physical world; (2) hierarchical semantic mapping at the edge; and (3) multi-robot perception, planning, and communication. Current research outcomes of the institute (see e.g., publications at [TILOS]) show ties to optimization foundations (e.g., solving continuous and discrete optimization problems at scale to facilitate perception, mapping, planning,

communication, and reconfiguration), networks (autonomous networking provides the backbone for perception, control, and learning in teams), and chips (hardware design to support computation, communication, and sensing on SWaP-constrained autonomous robots).

As an example, a new collaboration has been formed between the Robotics and Foundations teams on the topic of learning dynamics from trajectory data that respect the underlying structure and physical laws of dynamical systems. In particular, mobile and articulated robots often have Lie groups (viewed as a manifold) as configuration spaces. It is therefore highly desirable to formulate learning and control of robot dynamics that evolve on Lie groups. In a joint effort across the Robotics and Foundations teams, TILOS members develop a new structure-preserving deep learning architecture that can learn controlled Lagrangian or Hamiltonian dynamics on Lie groups, either from position-velocity or position-only data (Duruiseaux et al., 2023). They are further developing safe control synthesis methods with provable safety guarantees for problems with uncertain dynamics and constraints.

*Communication networks* also depend on AI and optimization, for example, for distributed optimization of a large number of discrete and continuous variables over nonconvex geometric structures induced by interference. A core challenge is to manage at scale radio resources across many devices distributed in space, with comprehension of the physics of signal propagation and information-theoretic limits. This gives rise to three major research thrusts: (1) multi-scale network optimization; (2) automated network fine-tuning; and (3) integration of human experts and physics. Current research outcomes of the institute (see e.g., publications at [TILOS]) highlight the ties to modern optimization foundations (e.g., sequential sampling, federated, and deep learning methods), as well as commonalities with other use domains (e.g., physical embeddedness, underlying graphical models) that enable cross-disciplinary bridges.

One particular work exemplifying the collaboration between the Networks and Foundations teams is the deployment of federated learning in a common three-tier IoT network architecture (Yu et al., 2023). Federated learning of multi-agent systems has been used here in an asynchronous setting with varying network delays, and shown to be effective with highly heterogeneous datasets. TILOS researchers also show their method to be highly resilient to system heterogeneity and dropouts (Vardhan, Ghosh, & Mazumdar, 2023; Yu et al., 2023).

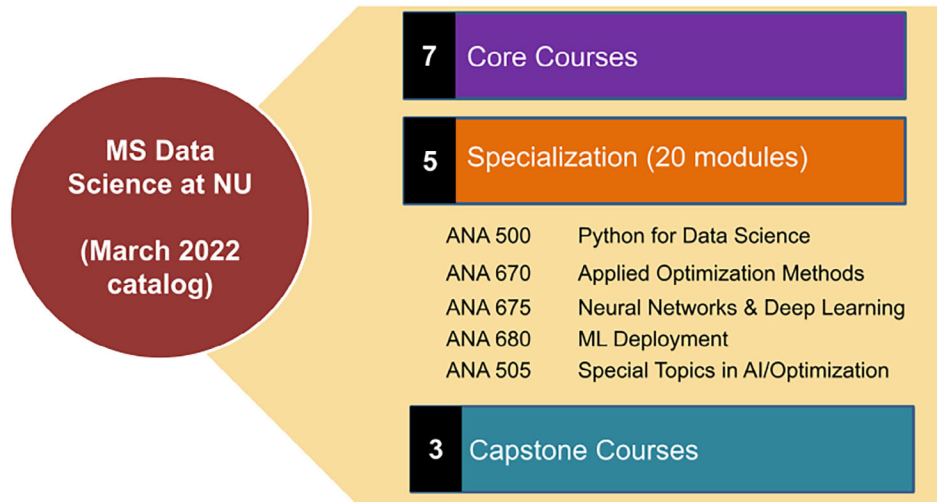
*Chip design* brings challenges that include hierarchical-system context, extreme cost, and sensitivity of training data, “multi-everything” (physics, objectives) constrained optimization, and pervasive security aspects. Given its decades-long history as a driver of applied optimization

and automation, chip design also highlights augmenting rather than rediscovering domain expertise, by encoding expert knowledge and intuition to serve optimization and decision-making agents. This gives rise to four major research thrusts: (1) direct generation of layout from circuit descriptions; (2) breakthrough scaling of verification methods; (3) quantifying the intrinsic cost of robustness in optimization and learning, with respect to aspects such as data anonymity, data integrity, and privacy in federated and distributed settings; and (4) data, benchmarking, and road mapping to improve reproducibility and relevance of research, along with translation into real-world contexts (Kahng, 2022). Research outcomes of the institute include several multi-organizational collaborative efforts seen in the TILOS organizational GitHub (TILOS Organization GitHub), ranging from new machine learning contests to open research enablements to breakthrough results for the classic hypergraph partitioning optimization problem.

## EDUCATION AND WORKFORCE DEVELOPMENT

New pathways to lifelong learning for diverse students, along with development of shareable and scalable curricula, are overarching objectives of TILOS efforts in Education and Workforce Development. The six TILOS institutions have a shared goal of (1) making education in optimization, AI, computing, robotics, networking, and chip design more accessible to a diverse group of students; and (2) providing opportunities for those already in the workforce to keep current with the latest developments. These aims are well-aligned to national needs, such as the revival of U.S. semiconductor technology leadership and a diverse domestic workforce, per the 2022 CHIPS and Science Act (CHIPS for America; CHIPS & Science Act, H.R.4346).

TILOS institute members work closely with corporate partners to understand the latter's workforce training needs. This guides creation of new curriculum modules, courses, and programs for university education (both in-person and online, undergraduate and graduate) as well as professional certificates, tutorials, short courses, and summer camps. San Diego-based National University (NU) is a key driver and motivating force for these efforts in TILOS. NU is a 53-year-old, non-profit university that educates a diverse group of students from across the U.S. with over 230,000 alumni and approximately 29,000 active students. NU student demographics include 60% female, 38% male, 14% active-duty military, 14% veteran, and an average age of 33 years. NU is also a Hispanic Serving Institution and a member of the Hispanic Association of Colleges and Universities.



**FIGURE 2** Five courses, comprising 20 modules, in the new AI/Optimization specialization of the M.S. Data Science degree program at National University.

A highlight of the past 2 years has been the development, teaching, and assessment of 20 new curriculum modules for a new five-course AI/Optimization specialization in NU's M.S. Data Science degree program (see Figure 2). Enrollment in the specialization has increased from 10 students in the calendar year 2022 to 21 students in Q1 2023 to 32 students in Q2 2023, with 56% of enrolled students being military-affiliated (veteran, retired military, active duty, and active reserve). More recently, a deep collaboration with Intel has enabled National University to integrate several of Intel's "AI for Workforce Development" curriculum modules into newly developed courses in the brand new B.S. Data Science degree at NU, with concentrations in AI/ML, cybersecurity analytics, and bioinformatics.

## KNOWLEDGE TRANSFER

Recall from Figure 1 that *translation* at the interface between industry and academia is the third virtuous cycle of TILOS. TILOS has built strong ties with various industries, facilitated by events such as TILOS Industry Day. In an idealized life cycle of translation, real-world practitioners supply problems and data, researchers bridge foundations and use domains to discover new methods, and these results go back into the real world. Unfortunately, today this picture is complicated by various technical and cultural obstacles. In addition to the many fruitful interactions TILOS has with our industry partners, TILOS also aims to help mitigate, if not remove, such obstacles. Four examples of obstacles to knowledge transfer and translation, along with potential mitigations from TILOS, are as follows.

- (1) Relevant (real) datasets may be proprietary and shared (if at all) only with very few researchers. TILOS research seeks new democratizations: Can we develop a science of "data virtual reality", enabling the generation of shareable, proxy research data that is artificial but indistinguishable from real *from the perspective of optimization methods and real-world practitioners*? A complementary need is to develop trusted tests for, for example, identity leakage.
- (2) Real data may be extremely scarce and expensive, for example, a single execution of the chip design flow may take several weeks, using tool licenses that cost millions of dollars. The research need is to learn to optimize with less real data, and to improve the reliability of methods for data augmentation and transfer learning.
- (3) Research may be inherently irreproducible. (i) Some fields have not yet gone through the stages of introspection (Hutson, 2018) and subsequent adoption of "papers with code" as a cultural norm. Here, TILOS directly aims to change long-standing cultures. A recent TILOS effort (Cheng et al., 2023; TILOS MacroPlacement GitHub) provides new open-source (code, data) research enablement, toward transparent assessment of a deep learning method for chip placement. (ii) Irreproducibility may also be due to proprietary tool scripting languages, or report and log-file formats, which cannot be published. Here, TILOS efforts elicited policy changes from major electronic design automation (EDA) tool vendors during the second half of 2022 (Junkin, 2022).
- (4) Benchmarking can be forbidden by suppliers of optimization software tools. At the same time, progress of optimization methods requires a clearly illuminated

leading edge, as even very well-studied optimizations may be far from well-solved. (For example, the state of the art for the classic hypergraph min-cut partitioning optimization remained static for a quarter-century, until results last year in Bustany et al. (2022)). The resulting challenge for both research and culture change is to develop new principles and mechanisms that provide a foundation for fair benchmarking.

Through the mitigation of the above obstacles, ongoing research, and efforts to change community culture, TILOS aims to scale *people* in addition to optimization in practice, by pioneering new democratizations, new cultural and scientific or technical norms, and principled bases for looking forward and investing resources.

## CONCLUSION

Our world is at the brink of an era where AI becomes an integral part of our daily lives, enhancing our capabilities and shaping a brighter future for humanity. Modern challenges for optimization include helping existing systems to learn from data and adapt to changing circumstances, so as to achieve improved accuracy, speed, and efficiency. In their collective pursuit of optimization in and via AI, TILOS researchers aim to advance foundational mathematics and domain sciences, fueling innovation and enabling us to unlock the full potential of AI technologies. We invite readers to learn more at the institute website, [tilos.ai](https://tilos.ai).


## ACKNOWLEDGMENTS

This material is based upon work supported by the National AI Research Institutes program supported by NSF and Intel under Award No. 2112665. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Andrew B. Kahng  <https://orcid.org/0000-0002-4490-5018>

## REFERENCES

Bustany, I., A. B. Kahng, Y. Koutis, B. Pramanik, and Z. Wang. 2022. "SpecPart: A Supervised Spectral Framework for Hypergraph Partitioning Solution Improvement." In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*.

- Cheng, C.-K., A. B. Kahng, S. Kundu, Y. Wang, and Z. Wang. 2023. "Assessment of Reinforcement Learning for Macro Placement." In *Proceedings of the ACM/IEEE International Symposium on Physical Design*, 158–66.
- CHIPS and Science Act, H.R.4346, 117th Congress (2021–2022), Public Law No: 117-167 (08/09/2022). <https://www.congress.gov/bill/117th-congress/house-bill/4346/text>
- CHIPS for America, National Institute of Standards and Technology (NIST), U.S. Department of Commerce. <https://www.nist.gov/chips>
- Duruiseaux, V., T. Duong, M. Leok, and N. Atanasov. 2023. "Lie Group Forced Variational Integrator Networks for Learning and Control of Robot Systems." *PMLR Learning for Dynamics and Control Conference (L4DC)*.
- Hutson, M. 2018. "Artificial Intelligence Faces Reproducibility Crisis." *Science* 359(6377): 725–26.
- Jegelka, S. 2022. "Theory of Graph Neural Networks: Representation and Learning." In *Proceedings of the International Congress of Mathematicians (ICM)*.
- Junkin, D. 2022. "Supporting the Scientific Method for the Next Generation of Innovators." <https://open-source-eda-birds-of-a-feather.github.io/doc/slides/BOAF-Junkin-DAC-Presentation.pdf>.
- Kahng, A. B. 2022. "Leveling Up: A Trajectory of Open-ROAD, TILOS and Beyond." In *Proceedings of the ACM/IEEE International Symposium on Physical Design*, 73–79.
- Kahng, A. B., R. R. Nerem, Y. Wang., and C. Yang. 2024. "NN-Steiner: A Mixed Neural-Algorithmic Approach for the Rectilinear Steiner Minimum Tree Problem." In *Proceedings of 38th Annual AAAI Conference on Artificial Intelligence (AAAI)*.
- Luo, Z., T. Hy, P. Tabaghi, M. Defferrard, E. Rezaei, R. Xarey, R. Davis, R. Jain, and Y. Wang. 2024. "DE-HNN: An effective neural model for Circuit Netlist representation." In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- TILOS. <https://www.tilos.ai/publications/>.
- TILOS MacroPlacement GitHub. <https://github.com/TILOS-AI-Institute/MacroPlacement>.
- TILOS Organization GitHub. <https://github.com/TILOS-AI-Institute/>.
- Vardhan, H., A. Ghosh, and A. Mazumdar 2023. "A Convergent Federated Clustering Algorithm without Initial Condition." In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.
- Yu, X., L. Cherkasova, H. Vardhan, Q. Zhao, E. Ekaireb, X. Zhang, A. Mazumdar, and T.Š. Rosing. "Async-HFL May 2023. Efficient and Robust Asynchronous Federated Learning in Hierarchical IoT Networks." In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, 236–48.

**How to cite this article:** Kahng, A. B., A. Mazumdar, J. Reeves, and Y. Wang. "The TILOS AI Institute: Integrating optimization and AI for chip design, networks, and robotics." *AI Magazine* 45: 54–60. <https://doi.org/10.1002/aaai.12165>

## AUTHOR BIOGRAPHIES

**Andrew B. Kahng**, University of California, San Diego, is the Founding PI and director (2021–2023), Chips team co-lead. He is Distinguished Professor of CSE and ECE in the Jacobs School of Engineering at UCSD, where he holds the endowed chair in high-performance computing. He is a Fellow of ACM and IEEE, and the 2019 Ho-Am Prize laureate in Engineering. His interests span many aspects of IC design automation, including fundamental algorithms, machine learning, and open source. ORCID: 0000-0002-4490-5018

**Arya Mazumdar**, University of California, San Diego, is the Deputy Director and Associate Director for Research at TILOS, Co-PI and Foundations team co-lead, Associate Professor, The Halicioğlu Data Science Institute. He is a Distinguished lecturer of the IEEE Information Theory Society and a recipient of the NSF CAREER Award, Jack K. Wolf Paper Award, and a European Association for Signal Processing Best Paper Award among others. Mazumdar's cur-

rent research interests include information theory, statistical machine learning and estimation, and distributed/federated optimization. ORCID: 0003-4605-7996

**Jodi Reeves**, National University, is the Associate Director for Education, Diversity, and Outreach and chair of the Education and Workforce Development Committee, Professor and Department Chair of Data Science, Academic Program Director for M.S. Data Science and B.S. Data Science. ORCID: 0001-9240-1722

**Yusu Wang**, University of California, San Diego, is the PI and Director (2023-present), Foundations team and formerly Associate Director, Research (2021–2023), Professor at The Halicioğlu Data Science Institute, Faculty co-lead for Foundations of Data Science CoP at Translational Data Analytics Institute at the Ohio State University from 2018 to 2020. NSF CAREER Award and DOE ECPI Award. Member of the Computational Geometry Steering Committee, AATRN Advisory committee, and SIGACT CATCS Committee. ORCID: 0000-0001-7950-4348



## SPECIAL TOPIC ARTICLE

# From learning optimization to learner flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (iSAT)

Sidney K. D'Mello<sup>1</sup> | Quentin Bidy<sup>1</sup> | Thomas Breideband<sup>1</sup> | Jeffrey Bush<sup>1</sup> |  
Michael Chang<sup>2</sup> | Arturo Cortez<sup>3</sup> | Jeffrey Flanigan<sup>4</sup> | Peter W. Foltz<sup>1</sup> |  
Jamie C. Gorman<sup>5</sup> | Leanne Hirshfield<sup>1</sup> | Mon-Lin Monica Ko<sup>1</sup> |  
Nikhil Krishnaswamy<sup>6</sup> | Rachel Lieber<sup>1</sup> | James Martin<sup>1</sup> | Martha Palmer<sup>7,8</sup> |  
William R. Penuel<sup>1,3</sup> | Thomas Philip<sup>2</sup> | Sadhana Puntambekar<sup>9</sup> |  
James Pustejovsky<sup>10</sup> | Jason G. Reitman<sup>1</sup> | Tamara Sumner<sup>1</sup> |  
Michael Tissenbaum<sup>11</sup> | Lyn Walker<sup>4</sup> | Jacob Whitehill<sup>12</sup>

<sup>1</sup>Institute of Cognitive Science, University of Colorado Boulder, Boulder, Colorado, USA

<sup>2</sup>Graduate School of Education, University of California Berkeley, Berkeley, California, USA

<sup>3</sup>School of Education, University of Colorado Boulder, Boulder, Colorado, USA

<sup>4</sup>Jack Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, California, USA

<sup>5</sup>Polytechnic School, Arizona State University, Mesa, Arizona, USA

<sup>6</sup>Computer Science, Colorado State University, Fort Collins, Colorado, USA

<sup>7</sup>Linguistics, University of Colorado Boulder, Boulder, Colorado, USA

<sup>8</sup>Computer Science, University of Colorado Boulder, Boulder, Colorado, USA

<sup>9</sup>Department of Educational Psychology, University of Wisconsin-Madison, Madison, Wisconsin, USA

<sup>10</sup>Computer Science Department, Brandeis University, Waltham, Massachusetts, USA

<sup>11</sup>College of Education, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

<sup>12</sup>Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, USA

## Correspondence

Sidney K. D'Mello, Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA.

Email: [sidney.dmello@colorado.edu](mailto:sidney.dmello@colorado.edu)

## Abstract

The Institute for Student-AI Teaming (iSAT) addresses the foundational question: *how to promote deep conceptual learning via rich socio-collaborative learning experiences for all students?*—a question that is ripe for AI-based facilitation and has the potential to transform classrooms. We advance research in speech, computer vision, human-agent teaming, computer-supported collaborative learning, expansive co-design, and the science of broadening participation to design and study next generation AI technologies (called AI Partners) embedded in student collaborative learning teams in coordination with teachers. Our institute

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.



ascribes to theoretical perspectives that aim to create a normative environment of widespread engagement through responsible design of technology, curriculum, and pedagogy in partnership with K–12 educators, racially diverse students, parents, and other community members.

## INTRODUCTION

*What should the classroom of the future look like? And what is the role of AI in these classrooms of the future?*

Research on how people learn has converged toward a perspective of learning as fundamentally interactive, collaborative, and supported by tasks that are authentic to students' identities and interests (NASSEM, 2018). Research has also documented the conditions that promote inclusive learning and honor diversity (Langer-Osuna, 2017). This rich body of research has profoundly influenced national standards in math, science, and educational reform efforts aimed at developing skills for the 21st century workforce (Fiore, Graesser, & Greiff, 2018).

Yet, the dominant approach to incorporating artificial intelligence in education (i.e., AIED) has primarily focused on an entirely different vision where students individually interact with technology that “optimizes” learning (Grandbastien et al., 2016). This 50+ year-old vision has been implemented via traditional adaptive computer-based learning, intelligent tutoring systems, recommender systems, and more recently, in teacher dashboards. Generative AI, including large language models like Chat-GPT, GPT-4, and Bard, are very attractive to this vision because they can potentially address several persistent problems including authoring content, assessment of open-ended responses, question answering, and adaptive coaching (D’Mello & Graesser, in press). However, doing so risks reinforcing a 20th century vision of learning centered around *individual optimization* (i.e., helping individual students achieve mastery in narrow domains) rather than a 21st century vision focused on *collaborative flourishing* (co-constructing knowledge using disciplinary practices and 21st century skills across domains).

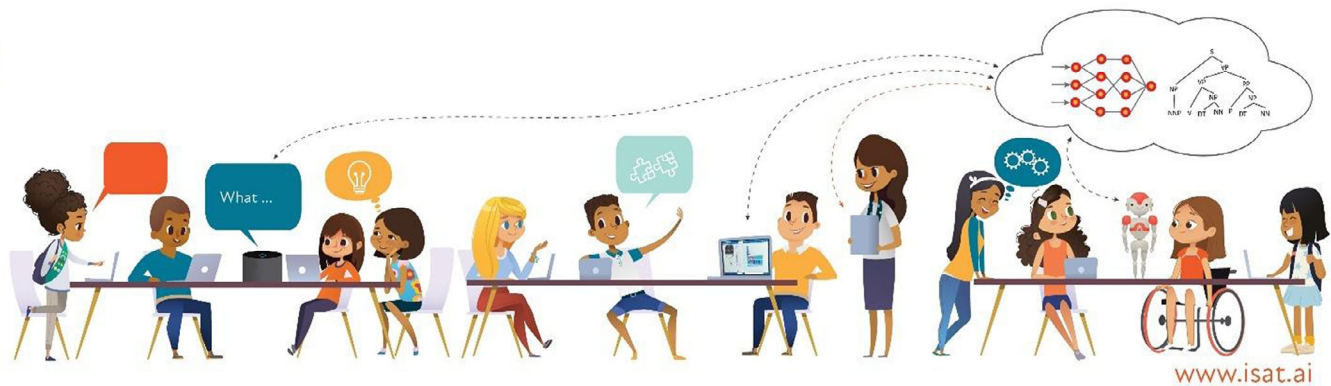
Accordingly, the Institute for Student-AI Teaming (iSAT) aims to reframe the role of AI in education, expanding from a current emphasis on intelligent tools supporting personalized learning through unimodal, individualized, unidimensional, instruction toward a future where AI is viewed as a social, collaborative AI Partner (Figure 1) that collaborates with students and teachers to make learning more effective, engaging, and equitable. Given the

importance of maintaining U.S. leadership in AI, and the excitement and concerns ushered forth by generative AI, we have selected AI-education as the focal domain for our research; our AI-enhanced curricula support diverse students to learn about, create, and critique AI technologies, and to think critically about the role of AI in society.

## OUR VISION: AI PARTNERS HELPING TEACHERS ORCHESTRATE CLASSROOMS FOR COLLABORATIVE LEARNING

In our envisioned future, classrooms have been transformed into knowledge-building communities, where student-AI teams engage in inquiry, critical thinking, and collaborative problem-solving as they investigate a scientific phenomenon, solve real-world problems, or develop solutions to a design challenge. Distinguishing characteristics of these communities are the ways in which teachers, students, and AI Partners systematically construct conversations that probe deep and sustained reasoning, enable all students to share and build on each other's ideas, and collaboratively solve challenging problems. In some cases, the AI Partner intelligently participates in conversations among small groups of students, facilitating their sense making and supporting them when they might get stuck. Other times, the students are teaching the AI Partner or are engaging it in peer learning. In all cases, the AI Partner communicates naturally by understanding students' speech, facial expressions, eye gaze, and gestures. Its algorithms extract meaningful information from these signals in real-world classrooms, while avoiding the pitfalls of bias and inequity. And it is socially sensitive so that students do not feel like they are being surveilled or monitored, but instead have a trusting relationship with the AI Partner.

In these future classrooms, AI Partners have not replaced the teacher. Rather, the AI Partners have been co-designed with educators (and students) to complement and augment the teachers to do what they are best at—the care and nurture of their students. Thus, AI helps teachers orchestrate effective learning experiences at the individual, small group, and whole class levels. For example, it extracts insightful nuggets from student small group conversations, such as moments when students are pushing



**FIGURE 1** An AI Partner (disembodied voice on left, embodied virtual agent in middle, robot on the right) collaborates with student teams and helps teachers orchestrate collaborative learning in classrooms.

each others' thinking and provides these to the teacher to facilitate whole-class discussions.

As a result of scaling these student-AI teams across a large number of classrooms in the future, there has been deeper student engagement and persistence in STEM, more inclusive classroom cultures, and significant increases in student learning outcomes, practices, and 21st century skills of collaboration and critical thinking.

## WHY, WHO, & SO WHAT?

### Rationale

The rationale for our Institute is five-fold.

*Foundational AI to power AI Partners.* Advances for understanding collaborative discourse, such as speech recognition, natural language processing, computer vision, and multimodal integration, form the basic building blocks for the AI Partners. However, foundational AI research is needed to enable these technologies to robustly address challenges unique to the classroom context where communication is noisy, multiparty, multimodal, and situated in the real-world, something that existing generative AIs do not address.

*Integrative knowledge on student-AI teaming.* The education and learning sciences have developed extensive knowledge on human-human collaborative learning, whereas the human-computer interaction, AI, and related communities have been exploring the foundations of human-agent teaming. There is a need to integrate and extend these knowledge bases to advance foundational research on the new science of student-AI teaming.

*New methods for design.* The field of AI currently lacks methods and processes to ensure that AI technologies reflect the needs, interests, and values of diverse community stakeholders. Accordingly, new methods are needed

to empower students with diverse identities to envision, co-create, critique, and apply AI learning technologies.

*Ethical & responsible AI.* Given the broad societal impacts of AI, it is imperative that developers of AI innovations, especially for use by students and other vulnerable populations, embody a culture of ethical and responsible AI. But too often the term “ethics” and “responsible” serve as useful talking points and design afterthoughts rather than foundational design principles. There is a need to develop and study methods for enacting responsible and ethical AI in real-world AI technology design.

*Emergence through convergence.* Our goal of “reimagining AI” implies a type of emergence—or the birth of something new where the whole is greater than the sum of its parts—through the convergence of seemingly disparate lines of research and expertise, an effort that requires major initiatives in multidisciplinary integration.

### Team & organization

Our team integrates more than 80 researchers and students from nine geographically distributed universities with our K–12 partners, including diverse students, teachers, and parents from two school districts and nonprofits with expertise in working with diverse youth. The team is organized as follows.

*Strand 1 (Understanding & facilitating collaborations)* is advancing foundational AI research in speech processing, natural language understanding, computer vision, and multimodal processing to develop AI models that can monitor and support collaborative learning at multiple levels including the content, the conversational dynamics, gestures, and social signals.

*Strand 2 (Orchestrating interactions with AI)* is developing the nascent science of student-AI teaming, including novel conceptual frameworks and interaction paradigms which specify how to orchestrate effective student and





teacher interactions with AI grounded in team science, collaborative practices, and learning outcomes.

*Strand 3 (Broadening participation with co-design)* is developing new methods for engaging culturally, ethnically, and gender diverse students and educators in AI-technology design, along with co-designing, implementing, and studying middle and high school STEM curriculum materials supporting broadening participation in AI education.

*Institute-wide (nurturing convergence research)* is integrating research across the strands, providing cross-strand services (data, annotation, technology), and coordinating the development and testing of the AI Partners.

The *Community & Outreach Hub* is promoting collaboration, knowledge sharing, and integration across the Institute, with its external partners, and the public at large.

## Anticipated outcomes

Our anticipated outcomes are in eight areas as shown in Table 1.

## SELECTIVE EXAMPLES OF ADVANCES IN FOUNDATIONAL, USE-INSPIRED RESEARCH, & BROADER IMPACTS

### Understanding and mitigating the impact of automatic speech recognition (ASR) errors

We addressed whether contemporary ASR systems, which are benchmarked on adult speech in idealized conditions, can be used to transcribe child speech in classroom settings. We found that state-of-the-art models (Google Speech and Whisper ASR) have very high word error rates on classroom data. However, downstream natural language understanding models that rely on embedding-based semantic representations have a much higher tolerance for ASR errors than those that also analyze semantic structure (Cao et al., 2023). Further, fine-tuning large ASR models on a combination of different child speech datasets resulted in improvements in ASR accuracy.

### Advances in integration of gesture and content analysis with collaboration constructs

A significant fraction of what is said cannot be understood without seeing what students are doing with

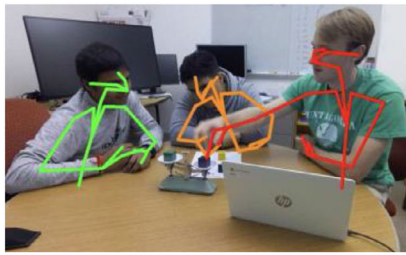
**TABLE 1** Anticipated outcomes of the Institute.

Research outcomes: AI Partners that embody	Strategic impacts
Foundational AI for multimodal, multiparty, multicurricular collaboration in classrooms	Over 5000 culturally, ethnically, and gender diverse K–12 students with new capacity to participate in AI learning and innovation
A new science of student-AI teaming including new frameworks for collaboration and classroom orchestration	A multiorganizational, multidisciplinary research community with new AI research capacity
New methods for broadening participation in the design of AI systems including new ethical design frameworks	Knowledge transfer to interdisciplinary research communities and communities of practice
Innovative AI-enabled curricula that enable middle and high school educators to integrate AI education into their classrooms	National nexus point for responsible design of AI technologies with diverse stakeholders

their hands and their bodies (e.g., “that one there” [pointing]—Figure 2A). To address this, we extended the Abstract Meaning Representations (AMR) linguistic annotation scheme to incorporate the meanings of gesture in Gesture-AMR (GAMR) (Brutti et al., 2022). In parallel, we developed NICE—*Nonverbal Interactions in Collaborative-Learning Environments*—a coding scheme to analyze nonverbals signals at a higher level of abstraction (e.g., are teammates engaged even if they are not speaking much?). We unified GAMR and NICE to analyze nonverbal behaviors grounded in collaboration constructs.

### Development of the collaborative learning and teaming framework

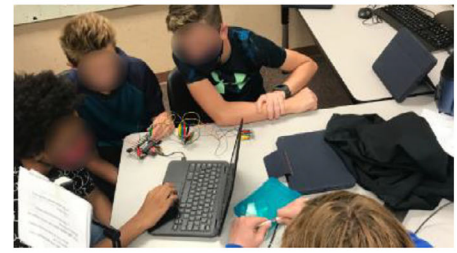
We organized the collaborative learning literature into a framework that incorporates different levels of interaction ranging from the individual to the team. We further advanced basic research aimed at addressing two major gaps in the literature: (1) first-person event-level coding and segmentation of collaboration for the purpose of imbuing coding schemes with users’ perceptions of the unfolding collaboration; and (2) development of real-time metrics for tracking how team interactions both influence and are influenced by student team members, teachers, and a potential AI Partner.



(A) Situated grounding in small group collaborations



(B) Students voices in design of the AI Partners



(C) Students collaborating in the Sensor Immersion curriculum unit

**FIGURE 2** Examples of research and education activities.

## New design methodology for escaping institutional gravity in expansive design

Co-design to (re)-imagine institutional contexts (and the technologies embedded within them) are impeded when participants' frame of reference is constrained by institutional realities. To address this, we developed a three-step design process to temporarily escape institutional gravity in expansive design by: (1) experiencing familiar alternate spaces; (2) creating speculative spaces with inexpensive materials (Figure 2B); and (3) bringing speculative spaces into existing institutions. This methodology created the conditions for youth to propose novel designs for AI Partners.

## New instructional materials to facilitate high-quality AI learning opportunities for nearly 4000 diverse students to develop, use, and critique AI systems

We co-designed (with educators) three curriculum units (Sensor Immersion, Self-driving Cars, and AI in Games unit) where students explored and programmed environmental sensors (Figure 2C), analyzed and visualized complex data streams, trained interactive bots, critiqued game design, and built models of neural networks. We then trained dozens of teachers to orchestrate high-quality learning opportunities with these units for nearly 4000 students, the majority of whom are from historically under-represented groups. Data from these implementations is used to train the machine-learning models underlying the AI Partners and learn how to embed them in the curricula.

## RESPONSIBLE & ETHICAL AI

We adopt the responsible innovation framework (RIF), which “means taking care of the future through collective stewardship of science and innovation in the present” (Stil-

goe, Owen, & Macnaghten, 2013). As elaborated below, this framework is reflected in all our work.

## Building shared values

We held a virtual kick-off retreat with goals of developing a sense of how our disciplinary lenses, positionalities, and life experiences shape what we notice in classrooms and how we see the potential of AI for learning, with a particular focus on the lens of AI justice.

## Learning Futures Workshops (LFWs)

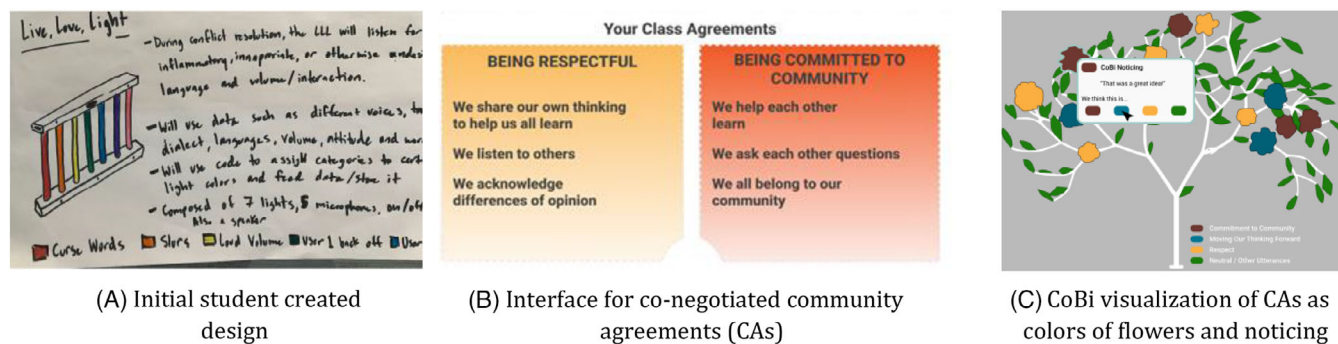
These intense, multi-day or multi-week engagements provide a space for diverse people to come together to envision possible futures for learning and to explore the role of AI in those futures. We have held three workshops with both students and teachers.

## Participatory design & co-design

We use participatory design methods that empower school and community stakeholders with diverse identities to participate in AI-related research and development. Further, we co-design (with educators) and study innovative middle and high school units supporting AI education.

## Adaptive conjecture mapping

We co-developed and used adaptive conjecture mapping (Chang & Dickler, 2023) which is graphical representation that connects the back-end decisions (e.g., what AI can sense), front-end decisions (e.g., how to interact with users), the mediating processes (what cognitive, behavioral, and affective processes arise from the interactions), and the intended outcomes, including both positive and negative.



**FIGURE 3** Design sketches of the Community Builder AI Partner (CoBi).

## Design sprints

In multiple-day in-person/hybrid meetings, our interdisciplinary teams organized design activities centered around incorporating the inputs from the various stakeholders (especially students) into the design of the AI Partners.

## Reflexivity

Lastly, we investigate the extent to which our Institute lives up to its commitment to the principles of responsible innovation by examining our engagement with diverse stakeholders, our uptake of their ideas in design, and the values of our own members.

## PUTTING IT ALL TOGETHER: COBI—THE COMMUNITY BUILDER AI PARTNER

We originally imagined a key role for the AI Partner was to help keep students together and on track when working in small groups, thereby addressing a chief challenge of teachers who assign small group collaborative work. Initially, youth themselves said they wanted something that could keep other students on task, until asked whether they would want such a partner to intervene in their own actions. After some probing, we learned they wanted a partner that could affirm their ideas, to recognize their contributions, and support them during collaborations. However, it was challenging for youth to imagine what good collaboration could look like beyond what they had experienced in school. Hence, we took them to a housing coop where the membership turned over on a regular basis, but where a common community feeling existed. They met a person who had the role of “community builder,” and the youth got really interested in the idea of shifting the narrative from AI policing their behavior to supporting their self-adherence to mutually agreed community agreements. They wondered if the AI Partner could help them

to generate and maintain such agreements and developed a design sketch to embody their ideas (Figure 3A).

Across many subsequent design sessions, including interviews with teachers and students, conjecture mapping, storyboarding, and prototyping, we developed our first AI Partner—the Community Builder or CoBi. CoBi helps students and teachers to co-negotiate classroom agreements along four dimensions: being respectful, being equitable, being committed to community, and moving thinking forward (Figure 3B). As students engage in collaborative learning, CoBi analyzes student discourse for evidence, or “noticings” of the agreement categories using our fine-tuned models for speech diarization, speech recognition, and discourse classification; eventually the models will also incorporate nonverbal information. The results are aggregated across student groups (to protect student privacy), and then visualized at the classroom level. Figure 3C shows a sketch of a qualitative and expansive visualization by way of a growing tree animation where the noticings are shown as flowers that bloom. Teachers use CoBi to guide students to reflect on the extent to which their collaborative discourse was aligned with their co-negotiated community agreements and to discuss any discrepancies. In future versions, students will have the opportunity to interrogate the underlying NLU models as a means for transparency, trust building, fostering agency, and to understand the strengths and limitations of AI.

In addition to CoBi, we are also developing AI Partners that support and engage in collaborative conversations with students (Cao et al., 2023). Our next steps with these partners involve user testing and refinement, testing for evidence of their effectiveness in promoting collaboration and learning, and scaling more broadly to classrooms across the nation.

## ACKNOWLEDGMENTS

This research was supported by NSF DRL 2019805.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Sidney K. D'Mello  <https://orcid.org/0000-0003-0347-2807>

## REFERENCES

- Brutti, R., L. Donatelli, K. Lai, and J. Pustejovsky. 2022. "Abstract Meaning Representation for Gesture." In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France.
- Cao, J., R. Dickler, M. Grace, J. Bush, A. Roncone, L. Hirshfield, M. Walker, and M. Palmer. 2023. "Designing an AI Partner for Jigsaw Classrooms." *Workshop on Language-Based AI Agent Interaction with Children (AIAIC'2023)*, Los Angeles, CA.
- Cao, J., A. Ganesh, J. Cai, R. Southwell, M. Perkoff, M. Regan, K. Kann, J. Martin, M. Palmer, and S. K. D'Mello. 2023. "A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse." In *Proceedings of the ACM International Conference on User Modeling, Adaptation and Personalization (UMAP 2023)*, 250–62, ACM.
- Chang, M. A., and R. Dickler. 2023. *A Conjecture Mapping Primer for Computer Scientists: Merging Learning Theories and Technical Research* (CIRCLS Rapid Community Report Series, Issue).
- D'Mello, S. K., and A. C. Graesser. 2023. "Intelligent Tutoring Systems: How Computers Achieve Learning Gains That Rival Human." In *Handbook of Educational Psychology*. 4th Edition.
- Fiore, S. M., A. Graesser, and S. Greiff. 2018. "Collaborative Problem-Solving Education for the Twenty-First-Century Workforce." *Nature Human Behaviour* 2(6): 367–369.
- Grandbastien, M., R. Luckin, R. Mizoguchi, and V. Aleven. 2016. "Preface to the IJAIED 25th Anniversary Issue." *International Journal of Artificial Intelligence in Education* 26: 1–3.
- Langer-Osuna, J. M. 2017. "Authority, identity, and collaborative mathematics." *Journal for Research in Mathematics Education*, 48(3): 237–47.
- NASEM. 2018. *How People Learn II: Learners, Contexts, and Cultures*. Washington, DC, National Academies Press.
- Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42(9): 1568–80.

**How to cite this article:** D'Mello, S. K., Q. Bidy, T. Breideband, J. Bush, M. Chang, A. Cortez, J. Flanigan, P. W. Foltz, J. C. Gorman, L. Hirshfield, M.-L. Monica Ko, N. Krishnaswamy, R. Lieber, J. Martin, M. Palmer, W. R. Penuel, T. Philip, S. Puntambekar, J. Pustejovsky, J. G. Reitman, T. Sumner, M. Tissenbaum, L. Walker, and J. Whitehill. 2024. "From Learning Optimization to Learner Flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (iSAT)." *AI Magazine* 45: 61–68.  
<https://doi.org/10.1002/aaai.12158>

## AUTHOR BIOGRAPHIES

**Sidney D'Mello** is a Professor in the Institute of Cognitive Science (ICS) and Department of Computer Science at CU Boulder. He directs the Institute for Student-AI Teaming (iSAT).

**Quentin Bidy** is an Assistant Research Professor at ICS at CU Boulder.

**Thomas Breideband** is a Research Scientist at ICS at CU Boulder.

**Jeffrey Bush** is a Research Scientist at ICS at CU Boulder.

**Michael Alan Chang** is a Postdoctoral Researcher at the University of California, Berkeley.

**Arturo Cortez** is an Assistant Professor of Learning Sciences and Human Development at CU Boulder.

**Jeffrey Flanigan** is an Assistant Professor of Computer Science and Engineering at the UC Santa Cruz.

**Peter Foltz** is a Research Professor at ICS at CU Boulder.

**Jamie Gorman** is a Professor of Human Systems Engineering at ASU.

**Leanne Hirshfield** is an Associate Research Professor at ICS at CU Boulder.

**Monica Ko** is an Assistant Research Professor at ICS at CU Boulder.

**Nikhil Krishnaswamy** is an Assistant Professor of Computer Science at CSU.

**Rachel Lieber** is the Outreach Coordinator for iSAT at CU Boulder.

**James Martin** is a Professor of Computer Science and at ICS at CU Boulder.

**Martha Palmer** is a Professor of Computer Science and Linguistics at CU Boulder. She is an ACL and AAAI Fellow.



**William R. Penuel** is a Distinguished Professor at ICS at CU Boulder.

**Thomas Philip** is a Professor at in the School of Education at the University of California Berkeley. He is an AERA Fellow.

**Sadhana Puntambekar** is a Sears-Bascom Professor of Learning Sciences at UW–Madison.

**James Pustejovsky** is the TJX Feldberg Chair in Computer Science at Brandeis University.

**Jason G. Reitman** is a Research Scientist at ICS at CU Boulder.

**Tamara Sumner** is a Professor at ICS at CU Boulder.

**Michael Tissenbaum** is an Associate Professor of Curriculum & Instruction and Educational Psychology at UIUC.

**Marilyn Walker** is a Professor of Computer Science and Engineering at the University of California, Santa Cruz.

**Jacob Whitehill** is an Associate Professor of Computer Science at Worcester Polytechnic Institute.



## SPECIAL TOPIC ARTICLE

# The AI Institute for Engaged Learning

James Lester<sup>1</sup> | Mohit Bansal<sup>2</sup> | Gautam Biswas<sup>3</sup> | Cindy Hmelo-Silver<sup>4</sup> |  
Jeremy Roschelle<sup>5</sup> | Jonathan Rowe<sup>1</sup>

<sup>1</sup>Computer Science, North Carolina State University, Raleigh, North Carolina, USA

<sup>2</sup>University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

<sup>3</sup>Computer Science, Vanderbilt University, Nashville, Tennessee, USA

<sup>4</sup>Center for Research on Learning and Technology, Indiana University, Bloomington, Indiana, USA

<sup>5</sup>Learning Sciences Research, Digital Promise, Washington, District of Columbia, USA

### Correspondence

James Lester, Computer Science, North Carolina State University, Raleigh, NC, USA.

Email: [lester@ncsu.edu](mailto:lester@ncsu.edu)

### Funding information

National Science Foundation, Grant/Award Number: DRL-2112635

### Abstract

The EngageAI Institute focuses on AI-driven narrative-centered learning environments that create engaging story-based problem-solving experiences to support collaborative learning. The institute's research has three complementary strands. First, the institute creates narrative-centered learning environments that generate interactive story-based problem scenarios to elicit rich communication, encourage coordination, and spark collaborative creativity. Second, the institute creates virtual embodied conversational agent technologies with multiple modalities for communication (speech, facial expression, gesture, gaze, and posture) to support student learning. Embodied conversational agents are driven by advances in natural language understanding, natural language generation, and computer vision. Third, the institute is creating an innovative multimodal learning analytics framework that analyzes parallel streams of multimodal data derived from students' conversations, gaze, facial expressions, gesture, and posture as they interact with each other, with teachers, and with embodied conversational agents. Woven throughout the institute's activities is a strong focus on ethics, with an emphasis on creating AI-augmented learning that is deeply informed by considerations of fairness, accountability, transparency, trust, and privacy. The institute emphasizes broad participation and diverse perspectives to ensure that advances in AI-augmented learning address inequities in STEM. The institute brings together a multistate network of universities, diverse K-12 school systems, science museums, and nonprofit partners. Key to all of these endeavors is an emphasis on diversity, equity, and inclusion.

## INTRODUCTION

AI holds significant transformative potential for improving K-12 education. Narrative-centered learning, which features story-based learning experiences, has long been recognized for the significant promise it holds for creating learning experiences that are both effective and engag-

ing for a broad range of student populations and subject matters (Mott et al. 1999; Saleh et al. 2022). The NSF AI Institute for Engaged Learning (EngageAI) conducts (1) use-inspired AI research on AI-driven narrative-centered learning environments, and (2) foundational AI research on natural language processing, computer vision, and machine learning. Inspired by a student-centered vision

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.

of AI-driven learning, the EngageAI Institute is creating AI-driven narrative-centered learning environments designed to promote student engagement. Advances in core AI technologies will enable new levels of interactivity and multimodal engagement, as well as support the creation of powerful predictive models of student learning.

The EngageAI Institute focuses on AI-driven narrative-centered learning environments that create engaging story-based problem-solving experiences to support collaborative inquiry learning (Saleh et al. 2022). The institute's AI-driven narrative-centered learning environments feature interactions with virtual characters capable of engaging in conversation while playing a variety of different roles. Generating interactive narratives dynamically tailored to the needs and interests of individual students in specific learning contexts requires a deep level of awareness of students' learning processes, as well as deep contextual understanding of what is happening in individual, small-group, and classroom (or museum) settings. Multimodal learning analytics utilizing new advances in natural language processing, computer vision, and machine learning will provide this awareness and understanding. By augmenting social interactions via embodied conversational agents that leverage multimodal data streams, AI-driven narrative-centered learning environments will deeply engage students in story-based learning experiences. These narrative-centered experiences will immerse students in storylines that drive their efforts, deepen their understanding of STEM concepts, facilitate their experience of STEM as a process of collective inquiry, and help them see STEM as addressing key societal challenges.

## RESEARCH STRANDS

The institute's research has three complementary strands (Figure 1).

### Narrative-centered learning

The institute is creating narrative-centered learning environments that generate engaging interactive story-based problem scenarios (Saleh et al. 2022). To support interest-driven learning explorations, the learning environments dynamically generate interactive narratives linked to authentic problem-solving scenarios, characters' behaviors and speech, curricular content, and support for learning. These capabilities are being created based on advances in generalizable and robust training of ML models with limited supervision, so as to enable tailored generation of interactive narratives that foster engaged student learning. With an emphasis on supporting narrative scenar-

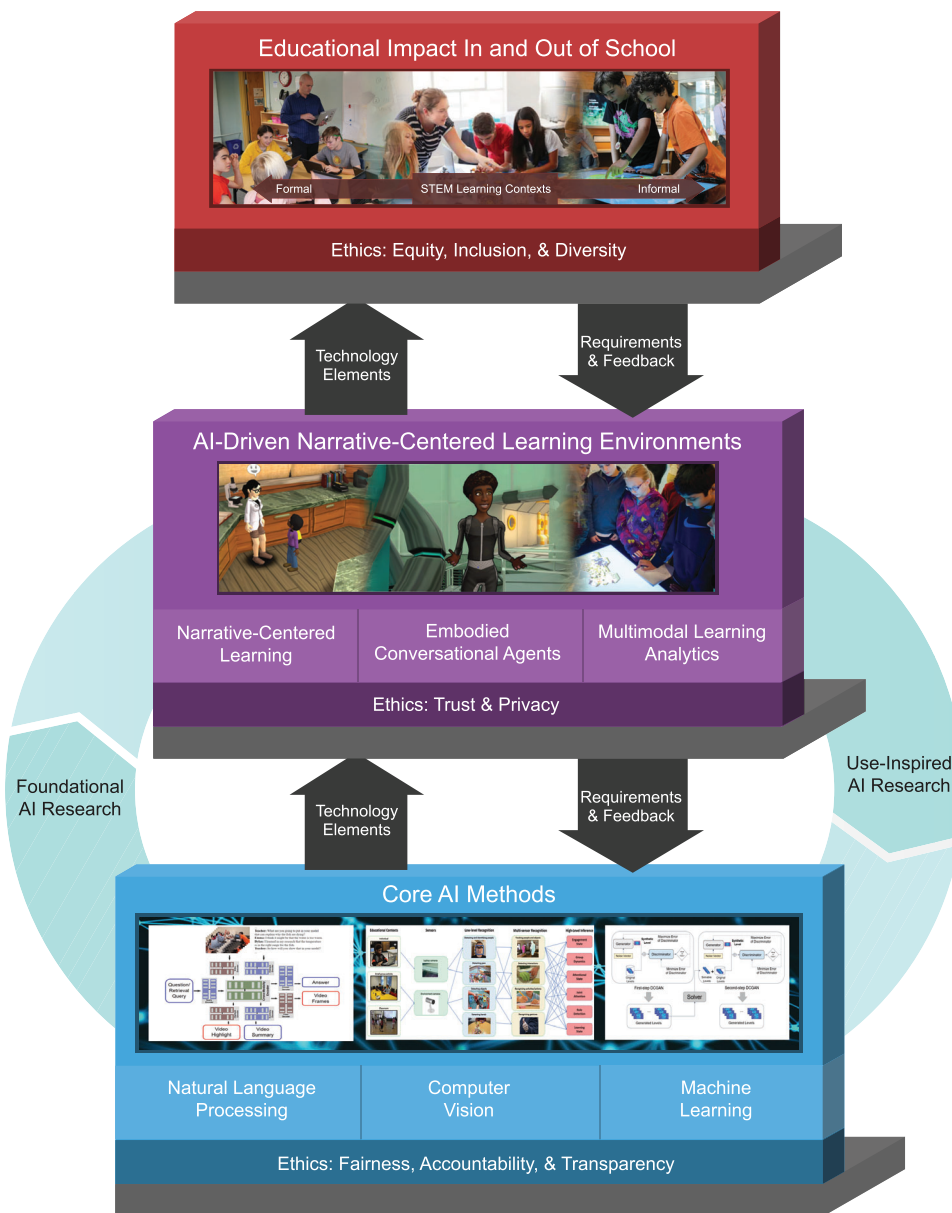
ios that elicit rich communication, require coordination, and spark collaborative creativity, the narrative-centered learning technologies will be driven by advances in multimodal generative machine learning and reinforcement learning.

### Embodied conversational agents

The institute is creating embodied conversational agent technologies (i.e., virtual agents) with multiple modalities for communication (speech, facial expression, gesture, gaze, and posture) to support engaging learning interactions (Johnson and Lester 2018). Embodied conversational agents are driven by advances in (1) natural language understanding (multiparty dialog structure, language models for long-form dialog, query-based video retrieval, low-resource automatic speech recognition for nonadults, multiple speakers, and noisy classrooms), (2) natural language generation (video and dialog summarization, question-answering, question-generation, explanation generation, paraphrasing for stylistic alignment, and controllable text-to-speech synthesis), (3) computer vision (gaze estimation, joint attention, attention and motion tracking, gesture, and action recognition), and (4) student modeling (predictive models of students' goals, plans, problem-solving strategies, cognitive states, and learning outcomes). The agents will support a broad range of roles including (1) providing students with cognitive, motivational, and affective support, (2) serving as virtual learning companions in collaborative learning, and (3) serving as cognitive assistants to teachers.

### Multimodal learning analytics

The institute is creating an innovative framework for multimodal learning analytics to support students, researchers, teachers, and informal STEM educators (Hutchins and Biswas 2023). It seeks to greatly expand teacher awareness and support with innovations in natural language processing, computer vision, and machine learning methods. Rich streams of multimodal data derived from students' conversations, gaze, facial expressions, gesture, and posture as they interact with each other, with teachers, and with embodied conversational agents, will support comprehensive student modeling. The institute is developing multimodal interfaces with visualization, summarization, and query-based retrieval capabilities for students and teachers with a special focus on multimodal learning analytics for narrative-centered learning. Explanatory features of multimodal interfaces will significantly enhance teachers' capacity to



**FIGURE 1** EngageAI Institute Research.

dynamically analyze and support the orchestration of rich student learning activities.

The institute's design of AI-driven narrative-centered learning environments is guided by educators' needs and students' interests and competencies. For example, middle school science teachers provide high-level guidance on narrative scenario generation to teach particular science concepts. By analyzing prior student performance inferred from student models, which are being driven by advances in machine learning, narrative-centered learning environments should anticipate student difficulties and need to adaptively support problem solving. For example, they should provide collaborative problem-solving advice delivered by conversational agents to

a group of middle school students solving a science problem.

### Illustrative scenario

To illustrate how AI-driven narrative-centered learning environments developed by the institute leverage advances in natural language processing, computer vision, and machine learning, consider a hypothetical middle school science class that is using an AI-enhanced version of a narrative-centered learning environment, EcoJourneys. When using EcoJourneys in the classroom, students work in small, co-located groups where they can





easily converse and interact using a shared instance of the learning environment. Early in the EcoJourneys storyline, students travel to a fictional island in the Philippines where they encounter a dilemma: a local fish farm is reporting that tilapia are falling sick at an alarming rate. This science inquiry-based problem-solving scenario centers on ecosystems, moving from understanding relevant components, processes and mechanisms in a system, before advancing to more complex ecosystem interactions. To support students through the collaborative inquiry process, each student can be assigned a unique path in EcoJourneys to explore the virtual environment and progress through the problem scenario.

Leveraging advances in foundational and use-inspired AI research by the institute, problems such as the sick fish scenario can be generated and dynamically tailored to foster engaged learning experiences that are meaningful and adapted to local relevance, students' interests, collaboration requirements, and STEM knowledge. Narrative-centered learning environments can feature interactive storylines linked to authentic science problems that shape how the scenarios unfold as students solve them. The primary elements of interactive narratives such as sequences of plot events and character behaviors can be synthesized by generative models and reinforcement learning. Virtual environments can be generated procedurally based upon several factors, including socio-cultural factors to prioritize inclusion and relevance across a diverse range of learners. To create learning experiences that are both effective and engaging, dynamically generated adaptive support (e.g., hints, prompts, and feedback) can be discreetly embedded within the narrative scenarios to support learners in articulating how the data they collect can support claims about the algal bloom and its impact on the aquatic ecosystem.

Adaptive scaffolding can be delivered by virtual characters deeply infused with next-generation conversational agent technologies that combine dynamically generated dialog, facial expression, gaze, gesture, and body movement for each character. The characters can engage with students in dialogs by serving as (1) virtual mentors, (2) teachable agents, (3) virtual learning companions, (4) virtual collaborators, and (5) virtual facilitators. Students can engage in rich multiparty dialogs with virtual characters; engage in exchanges involving multimodal question generation, answering, and summarization; and communicate naturally through verbal and nonverbal behavior.

As students work together on EcoJourneys in the classroom, multimodal learning analytics can track students' learning and the difficulties they face, as well as students' eye gaze, facial expression, posture, and verbal and nonverbal interactions, with each other, and with their teacher. This data can be analyzed using natural language processing and computer vision to drive run-time narrative

generation and pedagogical decisions. Multimodal learning analytics can also track students' contributions to the collaborative learning process, and adaptive scaffolding functionalities delivered through embodied conversational agents can support student collaboration and reasoning about the problem scenario.

Throughout students' collaborative learning experiences, the teacher can be informed of students' progress and difficulties, and they can use this information to engage each group through conversations, providing resources, and further facilitating dialog among the students to reinforce ideas and perspectives from within the student's assigned roles in supporting the inquiry process.

## INSTITUTE TEAM

The Institute brings together a deeply interdisciplinary team spanning five organizations with expertise in AI and education: four universities (North Carolina State University, University of North Carolina at Chapel Hill, Vanderbilt University, and Indiana University) and Digital Promise, a nonprofit intermediary organization, which serves as a "nexus" role for the institute. Headquartered in the Research Triangle of North Carolina and with a multi-state network of school and museum partners, the Institute is led by PI James Lester (North Carolina State University) and Managing Director Jonathan Rowe (North Carolina State University) and co-PIs Mohit Bansal (University of North Carolina at Chapel Hill), Gautam Biswas (Vanderbilt University), Cindy Hmelo-Silver (Indiana University), and Jeremy Roschelle (Digital Promise). They are joined by colleagues in natural language processing, computer vision, machine learning, AI in education, learning analytics, the learning sciences, STEM education, and teacher education.

## ETHICAL AI IN EDUCATION

The transformative potential of AI comes with significant responsibility to look beyond its prospective benefits and recognize the challenges and potential risks inherent in AI-augmented engaged learning. The institute thus identifies strategies for addressing challenges, managing the risks, and assessing the areas where caution is necessary to ensure that AI's impacts are beneficial for learning and fairly distributed. Equally important to the careful design and execution of the Institute's ethical AI is transparent communication among all partners in AI-augmented engaged learning: researchers, educators, learners, their families, and the public. **Promoting equity, inclusion, and diversity in AI-augmented**

**engaged learning:** The institute engages with a diverse range of learners and educators throughout the design and development of its AI-driven narrative-centered learning environments, and it is advancing computational models of narrative toward producing story-centric problem scenarios and virtual worlds that embody socio-culturally relevant settings and contexts, cultivating feelings of personal identification and relevance across a broad range of learners. **Maintaining privacy and trust:** The institute's research, development, and dissemination efforts are infused with an emphasis on the use of privacy-aware techniques throughout the design and implementation of the AI-driven narrative-centered learning environments. Additionally, the institute is focused on designing, developing, and investigating AI models that are *trustworthy*, particularly among key stakeholders: students, educators, researchers, administrators, parents, and policy makers (Tam et al. 2023). **Fairness, accountability, and transparency in AI-empowered education:** Recent years have seen growing recognition of the important role of algorithmic bias in AI systems. The institute is formulating principles and methods to detect and mitigate the potential for codifying implicit bias into the AI models that drive its AI-driven narrative-centered learning environments.

## ENGAGEAI R&D MODEL

The institute's research is driven by progression through (1) fundamental breakthroughs in narrative-centered learning environment technologies, (2) integrated narrative-centered learning environments, and (3) scalable narrative-centered learning environments. First, the institute targets fundamental breakthroughs in narrative-centered learning environment technologies, which span foundational AI and use-inspired AI technologies. Second, the institute's research on narrative-centered learning environments includes the design and development of integrated narrative-centered learning environments that integrate all of the functionalities required for narrative-centered learning. Third, the institute's research on narrative-centered learning environments will include the design and development of scalable narrative-centered learning environments that can be implemented "in the wild" at scale. The outcome of this work will be narrative-centered learning environments that can be used by students and educators at scale.

The institute has established a hybrid top-down/bottom-up research and development model that supports interconnections between the institute's use-inspired AI research and foundational AI-research activities and enables advances toward the institute's targeted research outcomes (Figure 2). The R&D model maps out a set

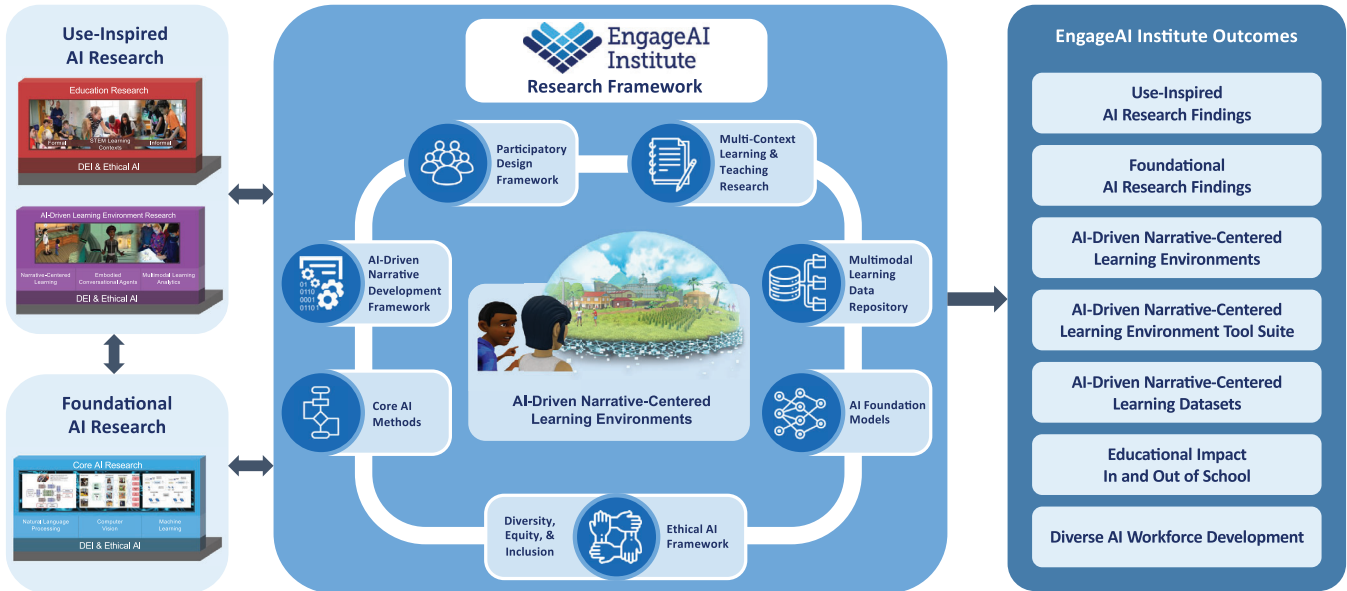
of critical shared resources that are continuously created, refined, and used by the institute in service to the accomplishment of its strategic and tactical objectives. The institute employs a system-integration approach to bring together the key enabling technologies—narrative-centered learning environments, embodied conversational agents, and multimodal learning analytics—into a single interoperable system that functions as a whole. The institute's AI-driven narrative-centered learning environments utilize a layered system architecture. The top layer, a dynamic narrative environment with multiple embodied conversational agents, is being designed to support teachers and students as they participate in interactive story-based problem-solving scenarios. Supporting the top layer is a lower layer consisting of a set of analytic and interpretation engines powered by multimodal analytics.

## BROADER IMPACTS

The EngageAI Institute emphasizes broad participation and diverse perspectives to ensure that advances in AI-augmented learning address inequities in STEM and computer science education. The institute brings together a multistate network of universities, diverse K-12 school systems, science museums, and nonprofit partners. Digital Promise, a nonprofit intermediary organization, serves a nexus role for the institute. Digital Promise leads the League of Innovative Schools, a network of 114 districts nationwide, serving over three million students. Through Digital Promise, the institute engages educators, related research and development communities, industry and start-ups, and philanthropy to solve broad-ranging problems in education and STEM workforce development. The institute provides a robust infrastructure to support at-scale implementations of AI-driven narrative-centered learning environments. A key emphasis in all of these endeavors is on diversity, equity, and inclusion. In addition to workforce development for undergraduates, graduate students, and postdocs, as well as Institute Convergence Workshops and partnerships with Boys & Girls Clubs, the institute partners with organizations committed to broadening participation in computer science, including STARS Computing Corps, Code.org, the Computer Science Teachers Association, and CSforAll.

## HIGHLIGHTS OF ACCOMPLISHMENTS

During its first 2 years, the EngageAI Institute has seen the launch of more than 25 active research projects, including 12 multi-institutional multidisciplinary projects, spanning use-inspired AI research and foundational AI research that



**FIGURE 2** EngageAI R&D Model.

support the design, development, and investigation of AI-driven narrative-centered learning environments. Selected accomplishments include the following:

1. The institute designed and developed its first prototype AI-driven narrative-centered learning environment that serves as an AI system integration demonstration for AI-driven narrative planning, embodied conversational agents, and multimodal learning analytics functionalities integrated into a single system. The prototype features a run-time AI-driven narrative planner that can dynamically redirect the flow of interactive narrative events unfolding in real-time in a narrative-centered learning environment (Wang et al. 2018). The prototype's narrative planning capabilities are guided by a set of decision-making policies induced using deep reinforcement learning techniques. The prototype AI-driven embodied conversational agent provides a text-based natural language interface, which utilizes pretrained language models (e.g., T5, S-BERT) to drive natural language understanding functionalities in support of question-answer dialogs between the student and agent relevant to the learning environment's science problem-solving scenario. The prototype was also integrated with a multimodal sensor-based data capture module that provides functionalities for capturing student facial expression, posture tracking, and interaction trace logs that record problem-solving actions in the narrative-centered learning environment.
2. A joint team from NC State University and Indiana University investigated deep learning-based multiparty dialog act recognition using text chat data from student

- collaborative learning interactions in the EcoJourneys narrative-centered learning environment (Saleh et al. 2022). The analysis utilized text utterances from student collaborative problem-solving chat, which were manually annotated using a theoretically grounded coding scheme. Using this dataset, the team examined the efficacy of transfer learning techniques with pretrained LLMs (i.e., BERT, T5) on dialog analysis of student and facilitator chat messages. Results showed LLM-based methods outperformed baseline models on automated classification of topic- and epistemic-based labels for the utterances. The results of this analysis show significant promise for informing AI-driven models for adaptively scaffolding student collaborative problem-solving by embodied conversational agents.
3. A team at the University of North Carolina investigated the performance of state-of-the-art video-and-language retrieval and summarization models on publicly available classroom video datasets. Specifically, the team focused on text-to-video retrieval, video-moment retrieval, video-moment segmentation, and video-moment captioning models, which retrieve the most relevant video from a set of candidate videos according to an input text query and caption the small step/moment-based contents of the video. For text-to-video retrieval, the team evaluated CLIP ViTB/32 (zero-shot/fine-tuned) and the HiREST joint model (zero-shot/fine-tuned) that was developed in Co-PI Bansal's lab (Zala et al. 2023). For moment retrieval, moment segmentation, and moment captioning, the team focused on evaluating the HiREST joint model. Results from experiments indicated that

current models, such as CLIP and the HiREST joint model, are able to perform reasonably well on all four of the tasks (text-to-video retrieval and moment retrieval/segmentation/captioning) in this domain, but there is still significant room for improvement, especially for text-to-video retrieval and moment captioning.

4. A team led by Co-PI Biswas at Vanderbilt University designed and developed a prototype multimodal learning analytics pipeline that implements a modular, component-based, distributed data collection networking architecture that supports the collection, alignment, and archiving of data from multiple sensing modalities as well as a flexible multiprocessor computational architecture that supports the use of compute-intensive machine learning algorithms for multimodal data analysis (Hutchins and Biswas 2023). While still in its infancy, the implementation supports high-throughput communication in a distributed environment, where the artifacts from sensors (different modalities) can be collected, aligned, and archived in a centralized fashion.
5. The institute's Nexus team, led by Digital Promise, convened a pair of EngageAI Institute Forums at the Computer History Museum in Mountain View, California to engage researchers, practitioners, and developers in reflective discussion about artificial intelligence in education. A national group of 150 attendees have participated in the Forums' highly interactive programs, including panels, poster sessions, and roundtable discussions encouraging conversation about the present and future of AI-augmented learning. A series of blogs summarizing key findings from the forums, as well as other pieces on AI-augmented learning, has been published on the EngageAI Institute website and disseminated through social media.

## CONCLUSION AND FUTURE PLANS

Driven by a vision in which AI supports and extends the intelligence of teachers and learners, the EngageAI Institute is designing, developing, and investigating AI-driven narrative-centered learning environments that create engaging story-based, collaborative problem-solving experiences. The EngageAI Institute will continue to pursue an ambitious research agenda consisting of foundational AI research in natural language processing, computer vision, and machine learning, as well as use-inspired AI research on narrative-centered learning environments with rich AI-driven virtual agents and powerful multimodal learning analytics to understand how students learn and collaborate in story-based problem scenarios. The

institute will develop a robust infrastructure to support at-scale implementations of AI-driven narrative-centered learning environments. It will serve as a nexus for distinctive innovations in in-school and out-of-school STEM education, and empower diverse learners to become the next-generation STEM workforce by creating generative, collaborative AI-driven narrative-centered learning environments that deeply engage learners in schools, at museums, and within their own communities. This vision is being informed by connections with diverse stakeholders to ensure that the institute's learning environments are ethically designed and promote diversity, equity, and inclusion. The EngageAI Institute stands at the forefront of innovation in STEM education, combining the power of AI-augmented learning with immersive storytelling to foster engaged and effective STEM learning experiences for all learners.

## ACKNOWLEDGMENTS

The material is based upon work supported by the National Science Foundation under Grant No. DRL-2112635. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

James Lester  <https://orcid.org/0000-0003-1481-6601>

Gautam Biswas  <https://orcid.org/0000-0002-2752-3878>

Cindy Hmelo-Silver  <https://orcid.org/0000-0003-2275-5212>

Jeremy Roschelle  <https://orcid.org/0000-0003-2219-0506>

Jonathan Rowe  <https://orcid.org/0000-0003-2038-9239>

## REFERENCES

- Hutchins, N. M., and G. Biswas. 2023. "Co-designing Teacher Support Technology for Problem-Based Learning in Middle School Science." *British Journal of Educational Technology*.
- Johnson, L., and J. Lester. 2018. "Pedagogical Agents: Back to the Future." *AI Magazine* 39(2): 33–44.
- Mott, B. W., C. B. Callaway, L. S. Zettlemoyer, S. Y. Lee, and J. C. Lester. 1999. "Towards Narrative-Centered Learning Environments." In *Proceedings of the 1999 AAAI Fall Symposium on Narrative Intelligence*, 78–82.
- Saleh, A., T. M. Phillips, C. E. Hmelo-Silver, K. D. Glazewski, B. W. Mott, and J. C. Lester. 2022. "A Learning Analytics Approach Towards Understanding Collaborative Inquiry in a Problem-Based Learning Environment." *British Journal of Educational Technology* 53(5): 1321–42.
- Tam, D., A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel. 2023. "Evaluating the Factual Consistency of Large



Language Models Through News Summarization.” In *Findings of the Association for Computational Linguistics: ACL 2023*, 5220–55.

Wang, P., J. Rowe, W. Min, B. Mott, and J. Lester. 2018. “High-Fidelity Simulated Players for Interactive Narrative Planning.” In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3884–90. Stockholm, Sweden.

Zala, A., J. Cho, S. Kottur, X. Chen, B. Oguz, Y. Mehdad, and M. Bansal. 2023. “Hierarchical Video-Moment Retrieval and Step-Captioning.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23056–65.

**How to cite this article:** Lester, J., M. Bansal, G. Biswas, C. Hmelo-Silver, J. Roschelle, and J. Rowe. 2024. “The AI Institute for Engaged Learning.” *AI Magazine* 45: 69–76.  
<https://doi.org/10.1002/aaai.12161>

## AUTHOR BIOGRAPHIES

**James Lester** is the Goodnight Distinguished University Professor in Artificial Intelligence and Machine Learning at the North Carolina State University, and he is the Director of the National Science Foundation AI Institute for Engaged Learning.

**Mohit Bansal** is the John R. & Louise S. Parker Professor and the Director of the MURGe-Lab (UNC-NLP Group) in the Department of Computer Science at the University of North Carolina at Chapel Hill.

**Gautam Biswas** is a Cornelius Vanderbilt Professor of Engineering, and Professor of Computer Science and Computer Engineering at Vanderbilt University, and he is a Senior Research Scientist at the Institute for Software Integrated Systems.

**Cindy Hmelo-Silver** is a Distinguished Professor of Learning Sciences and the Barbara B. Jacobs Chair in Education and Technology at Indiana University, and she is also a Director of the Center for Research on Learning and Technology.

**Jeremy Roschelle** is the Executive Director of Learning Sciences Research at Digital Promise.

**Jonathan Rowe** is a Senior Research Scientist in the Center for Educational Informatics at the North Carolina State University and a Managing Director of the NSF AI Institute for Engaged Learning.



## SPECIAL TOPIC ARTICLE

# AI-ALOE: AI for reskilling, upskilling, and workforce development

Ashok Goel<sup>1</sup> | Chris Dede<sup>2</sup> | Myk Garn<sup>3</sup> | Chaohua Ou<sup>3</sup>

<sup>1</sup>Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA

<sup>2</sup>Harvard University, Cambridge, Massachusetts, USA

<sup>3</sup>Georgia Institute of Technology, Atlanta, Georgia, USA

### Correspondence

Ashok Goel, Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA.

Email: [goel@cc.gatech.edu](mailto:goel@cc.gatech.edu)

To appear in AAAI's AI Magazine Special Issue on National AI Institutes, Ashok Goel & Chaohua Ou (editors), Spring 2024.

### Funding information

National Science Foundation, Grant/Award Numbers: 2112531, 2247790

### Abstract

The National AI Institute for Adult Learning and Online Education (AI-ALOE) develops AI learning and teaching assistants to enhance the proficiency of adult reskilling and upskilling, and thereby transform workforce development. The AI assistants both address known problems in online education for reskilling/upskilling and help personalize adult learning for workforce development. AI-ALOE develops new AI models and techniques for self-explanation, machine teaching, and mutual theory of mind to make the AI assistants usable, learnable, teachable, and scalable. AI-ALOE is also developing a data architecture for deploying and evaluating the AI assistants, collecting and analyzing data, and personalizing learning at scale.

## VISION AND MISSION

We live in the age of AI, the era when AI became ubiquitous and impactful. As AI becomes increasingly powerful and pervasive, it likely will lead to the disruption of many skills and the displacement of many workers. The National AI Institute for Adult Learning and Online Education (AI-ALOE) envisions that AI can also help reskill and upskill workers throughout their lifetimes to remain in or rejoin the workforce. Thus, AI-ALOE develops AI theories, techniques, and tools to enhance the scale and proficiency of adult reskilling and upskilling.

Adult learning is different from other types of learning, cognitively, emotionally, and socially (National Academy of Sciences, 2018). For example, most K-12 children go to day schools for their education. However, many adults can-

not move to places of education due to age, health, jobs, families, finances, etc. Online education offers a scalable platform for reaching adult learners where they live and work. The distributed and frequently asynchronous nature of online education allows adults to learn at their own place and pace. Although online education is now common and rapidly growing, it has well-known drawbacks such as low *cognitive presence* (learners construct meaning through sustained interactions with the educational materials), low *social presence* (learners construct social connectedness with one another and with the teacher), and low *teaching presence* (design and facilitation of educational experiences including planning learning activities, communicating requirements, and guiding discussions) (Garrison, Anderson, & Archer, 2000). Completion rates for Massive Open Online Courses often are in single digits.

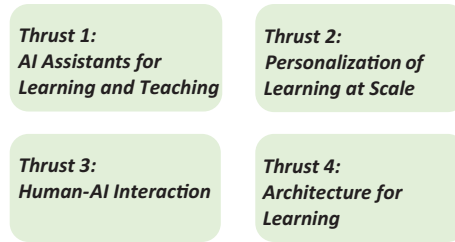
This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.

Thus, in its first thrust on *AI Assistants*, AI-ALOE develops AI learning assistants such as conversational textbooks and interactive videos to enhance cognitive engagement, as well as social assistants to improve teacher-learner and learner-learner interactions. It also develops AI teaching assistants to offload some of teachers' work, for example, through automated assessment and automated question answering.

Online education also offers access to much more data on learners and learning (and teachers and teaching) than is typically available for in-person learning. This large amount of data opens a path to enhance the quality of online education for adult learners through personalization of learning. Personalization of learning has long been a major goal of educational research (Bloom, 1984) and AI has made significant progress on personalized learning for well-defined problems in K-12 education. However, while K-12 education tends to focus on well-defined problems, adults often need to address ill-defined problems and adult learning often is self-directed and open-ended. Moreover, adults and adult learning vary across a vast range due to differences in age, health, education, competency, experience, goals, etc. Customization of learning to individual characteristics, strengths, and motivations is even more important for adult reskilling and upskilling than for K-12 students (Dede and Richards, 2020). Thus, in its second thrust on *Personalization of Learning*, AI-ALOE develops AI interactive environments for personalization of learning at scale in the context of ill-defined problems and self-directed learning. This entails the development of new analytical techniques that can monitor and analyze the behavior of individual learners as well as new methods of coaching that can nudge learners toward more productive behaviors.

Given that AI-ALOE's AI learning and teaching assistants and environments are used by humans (learners, teachers, other stakeholders), it is critical that they are: (i) Usable (learners and teachers of varying degrees of AI literacy should be able to easily use them), (ii) Learnable (learners and teachers of varying degrees of familiarity should be able to easily learn how to use them), (iii) Teachable (teachers of varying expertise should be able to train and tune the assistants to their individual preferences), and (iv) Scalable (developers should be able to efficiently scale the deployment of the assistants to a multitude of learners, teachers, and educational contexts) as shown in Figure 1. Thus, in its third thrust on *Human-AI Interaction*, AI-ALOE develops new AI models and techniques for (1) Self-explanation (an agent's ability to explain its reasoning), (2) Information visualization (the ability to make data comprehensible to humans), (3) Machine teaching (the ability to interactively teach an AI agent), and (4) Mutual theory of mind (the ability of humans and AI agents to ascribe mental states to each other).



**FIGURE 1** Four interdependent research thrusts of National AI Institute for Adult Learning and Online Education (AI-ALOE).

In its fourth thrust, AI-ALOE is developing an *Architecture for AI-Augmented Adult Learning (A4L)*. The A4L technology and data architecture enables deployment and evaluation of the AI assistants for learning and teaching in real educational contexts, collection and analysis of large-scale data on learning, and personalization of learning at scale. It also addresses the issues of data privacy, data security, and data sharing.

## AI-ALOE partner institutions

The AI-ALOE Institute consists about twenty computer, cognitive, learning, and education scientists from Georgia Institute of Technology, Georgia State University, Harvard University, Technical College System of Georgia (TCSG), University of North Carolina at Greensboro, and Vanderbilt University; the nonprofit organization, 1EdTech; and industrial partners Accenture, Boeing, IBM, and Wiley. The Institute, headquartered at Georgia Tech, includes about 50 students, and research and administrative staff.

## RESEARCH METHODOLOGY

The theory, design, development, deployment, and evaluation of the AI assistants and environments is an iterative process, with each cycle feeding into the next and leading to continual and sustained improvement in learning. At the inception of AI-ALOE, we made a strategic decision to deploy extant AI techniques and tools in real educational contexts as early as possible so subsequent iterations would be data-driven and evidence based. In its first two years, AI-ALOE has conducted its experiments in two different educational contexts to ensure robust generalizations: courses at the TCSG and classes in the Online Master of Science in Computer Science (OMSCS) at Georgia Tech. TCSG consists of 22 2-year technical colleges that serve a very diverse group of more than 300,000 students throughout the State of Georgia. OMSCS serves more than 12,000 globally distributed adult learners.

In the first 2 years, AI-ALOE deployed AI technologies in real classes at Georgia Tech and TCSG with thousands

of students, designed the A4L architecture to collect large amounts of data on adult learning, and established baseline learning results for these cohorts. Now that we have preliminary versions of the key AI technologies and baseline results on learning, we are expanding from basic, general education classes at TCSG to courses on occupational skills, and from classes in higher education at Georgia Tech and TCSG to courses for workforce development in industry. Further, in its first 2 years, AI-ALOE focused mostly on exploratory research, with initial deployments and observations in the first year paving the way for preliminary outcomes and formative assessments in the second year. As the Institute advances into the third year, it is poised to move to comparative research, and deploy A/B experiments and randomized controlled trials examining cognitive, teaching, and social engagement, student retention and satisfaction, and learning efficiency and effectiveness. In Years 4 and 5, we plan quasi-experimental and longitudinal studies for examining proficiency in learning first for preparation for advanced courses and then for preparation for the workforce.

## Responsible AI

AI-ALOE views the AI assistants it develops as part of a socio-technical system for advancing human goals, interests, and values. Thus, all AI assistants at AI-ALOE are designed to take the well-being of their users and other stakeholders into account. Michael Hoffmann at Georgia Tech leads a team that uses a participatory design model in which teams of users and practitioners evaluate AI technologies using IEEE (2022) rubrics for human well-being. Qualitative data is collected using focus groups and one-on-one discussions with students, instructors, and AI technology developers. This data is then used for co-design of the AI assistants in the next iteration.

## AI ASSISTANTS FOR LEARNING AND TEACHING

AI-ALOE is developing a suite of AI technologies for supporting adult learning through online education. Below we provide short summaries of half dozen AI assistants and then describe one (Jill Watson for conversational courseware) in more detail.

*Apprentice tutors for skill learning:* Skill learning refers to learning of procedures to accomplish a task. For example, in nursing, calculation of the precise amount of a drug to be given to a patient is a skill. Chris MacLellan and his team at Georgia Tech have developed Apprentice Tutors to enable learners to learn about arithmetic operations on fractions. Early results from use of Apprentice Tutors by hundreds of TCSG students across dozens of course sections for a

college algebra course indicate that many learners gained fractional arithmetic skills (MacLellan, Stowers, & Brady, 2022).

*SMART for concept learning:* Concept maps are graphical representations of concepts and relations among them. The Student Mental Model Analyzer for Research and Teaching (SMART) developed by Min Kyu Kim and his team at Georgia State University helps students build concept maps from text and provides feedback to help the students revise their maps. Recent studies at TCSG engaging hundreds of students in multiple classes in English and Biology indicate that SMART helps students build better and deeper concept maps (Kim et al., 2023).

*VERA for model learning:* VERA is an interactive inquiry-based learning environment for enabling a learner to construct conceptual models of ecological phenomena, evaluate the model through agent-based simulation, analyze the results, and revise the model (An et al., 2020). It has been used in both pedagogical contexts at Georgia Tech and TCSG and self-directed learning contexts common to adult learning. The use of VERA at TCSG in classes on natural resource management indicate it helps students learn about models and modeling.

*Ivy for interactive videos:* We have developed interactive videos that present exercises to test problem-solving skills after each lesson and provide adaptive feedback on students' responses to the exercises (Goel, & Joyner, 2017). Many students found the interactive exercises engaging and useful for their understanding of the course content (Ou, Joyner, & Goel, 2019). We are developing an infrastructure called Ivy for empowering instructors of online courses to develop their own interactive video lessons.

*iTELL for intelligent textbooks:* The Intelligent Textbooks for Enhanced Lifelong Learning (iTELL) project led by Scott Crossley at Vanderbilt University asks students to summarize textual documents, uses pretrained Large Language Models to automatically assess the student's summarization, and provides the student with feedback to enhance their comprehension of the document.

*SAMI for social interactions:* SAMI uses learners' self-introductions in an online class to help build connections based on the learners' locations, backgrounds, interests, etc. We have deployed SAMI in several OMSCS classes at Georgia Tech and found that for many adult learners, SAMI can help "break the ice" and foster social interactions (Wang et al., 2022).

## Jill Watson for conversational courseware

Jill Watson is a virtual cognitive assistant for engaging students in extended conversations about courseware including textbooks, video transcripts, presentation slides, class syllabi, and other course materials. The initial





version of Jill in 2016 used IBM's Watson platform and used class syllabi to answer students' questions on online discussions forums (such as Piazza) anytime anywhere (Goel & Polepeddi, 2018). An intermediate version of Jill Watson in 2019 switched to Google's BERT as the platform because it was available as open-source software and thus could be tuned for Jill. This version of Jill operated on online discussion forums (such as EdStem) as well as directly on the Canvas learning management system. It was also embedded in interactive environments such as VERA where it answered questions based on VERA's user guides (Goel et al., 2024).

A new version of Jill Watson uses ChatGPT as the backend to answer students' questions and support conversations with the course materials to enhance cognitive engagement and teaching presence. To reduce ChatGPT's biases and hallucinations, Jill conducts preprocessing on students' questions and creates prompts for ChatGPT, as well as postprocessing on the ChatGPT's answers. In preprocessing, it uses a variation on retrieval-augmented generation based on the courseware (textbook, slides, syllabus, etc.). We have found that while ChatGPT improves the accuracy and precision of Jill's answers to students' question, Jill Watson helps reduce ChatGPT's hallucinations. In Fall 2023, we introduced the new Jill in a Georgia Tech OMSCS class on AI consisting of more than 600 online students as well as in a TCSG class on freshman English Composition consisting of more than 100 online students. Preliminary results from A/B experiment in the OMSCS class on AI indicate that Jill Watson *enhances teaching presence and may also help improve student grades*.

## Human-AI interaction

The development of AI assistants raises foundational AI questions of usability, learnability, teachability and scalability. AI-ALOE is investigating machine teaching for teachability and scalability and theory of mind for usability and learnability of AI assistants.

**Machine teaching:** In machine teaching, a human teaches a machine learning model how to accomplish a task, such as named entity recognition or intent classification, in a manner that minimizes the human cost of teaching the model and the risk of the model being inaccurate. As an example, let us consider the task of building Jill Watson on top of ChatGPT, a zero-shot learner that makes many mistakes. One way of training ChatGPT is through human feedback on its answers, but this can be very costly. We are investigating machine teaching techniques such as active label correction that predict which answers of ChatGPT are likely to be incorrect, send only the fraction of

answers most likely to be incorrect to the human annotator, and train ChatGPT on the limited human feedback, thereby reducing the teaching cost.

In parallel, the Apprentice Tutors project is investigating how human teachers can be empowered to create and adapt the tutors by interactively teaching them. Apprentice Tutor represents knowledge of skills as a Hierarchical Task Network (HTN) and uses ChatGPT for translating a teacher's instructions in English into HTN.

**Theory of mind:** Theory of mind refers to the human ability to ascribe mental states to other humans in the form of goals, plans, knowledge, beliefs, feelings, etc. Mutual theory of mind pertains to the human ability to recognize another human's theory of one's own mind, for example, my theory of the reader's theory of my mind. One finding from the Jill Watson project is that students' perception of Jill's abilities not only change over time, but also that the changes are manifested in the linguistic expressions they use in interacting with Jill (Wang et al., 2021). Thus, we envision a future version of Jill that can analyze the changes in the linguistic expressions used by a student over time to build a theory of the student's current theory of the AI and adapt its responses accordingly. At AI-ALOE we are investigating how humans build a theory of AI agent's mind, how an AI agent may build a theory of a human user's mind, and how we may specify the content, structure, and processes of a mutual theory of mind between humans and AI agents.

## PERSONALIZATION OF LEARNING

Personalization aims to provide individualized optimal learning experiences that help each learner to achieve their maximum potential. AI-ALOE is investigating personalized learning at scale in two different ways. First, the AI learning and teaching assistants can support personalization. The Apprentice Tutors project, for example, is developing a framework for personalization driven by a model of expert problem-solving. The VERA, SMART, and iTELL projects too are building personalization using the AI assistants. We briefly describe personalization for self-directed learning in VERA below. Second, teachers and learners can personalize their teaching and learning if they have access to learning data in an easily comprehensible form. Visualizations can help keep humans in the loop to continuously integrate feedback into the processes of teaching and learning. Alex Endert and his team at Georgia Tech are investigating how visualization of learning using Apprentice Tutors informs learners and the teacher on the student's workflow. A related project focuses on the visualization of AI agents for the purpose of explainability and transparency.

## Personalization in VERA

Adult learning is distinct from K-12 education in many ways. For example, while K-12 education tends to focus on well-defined problems, adult learning typically occurs in the context of ill-defined problems and often is self-directed. The VERA project affords investigation of personalization of learning in the context of ill-defined problems and self-directed learning. VERA is an interactive learning environment for enabling a student of natural resource management to construct answers to questions such as “Why is the population of Brook trouts in Georgia rivers declining?” This problem is ill-defined because the system boundaries and variables are underspecified. VERA is accessible through Smithsonian Institution’s Encyclopedia of Life webportal (eol.org) for use both in pedagogical contexts and for self-directed learning.

AI-ALOE is developing machine learning techniques, such as dimensionality reduction, sequence analysis, hierarchical clustering, and Markov chain modeling, for analyzing the modeling behaviors of self-directed learners addressing ill-defined problems. It is also developing a suite of interactive coaches that can nudge an individual learner toward a more productive behavior based on an analysis of the current behavior. In Fall 2023, we deployed the coaches for personalized feedback in a self-directed laboratory section of a Georgia Tech undergraduate class on ecology.

## Human-AI interaction redux

Personalization of learning in ill-defined problems and self-directed learning raises additional foundational issues for human-AI interaction such as transparency and trust: Why should a learner trust the results of, say, an agent-based simulation of a conceptual model in VERA or the recommendations made by a VERA coach? AI-ALOE is investigating self-explanation as a strategy for making VERA’s reasoning transparent and its results trustworthy.

*Self-explanation:* Self-explanation refers to an agent’s ability to explain its reasoning and decisions. AI-ALOE takes a meta-cognitive stance toward generation of self-explanations. We endow the agent, in this case VERA, with a self-model, that is, a model of its own knowledge and reasoning. When VERA addresses a problem, it keeps track of its chain of thought and the results thereof. When a user asks for an explanation, VERA can introspect on its self-model and chain of thought to produce an answer. We are presently evaluating self-explanation in VERA for enhancing transparency of its reasoning and trust in its recommendations.



**FIGURE 2** Numbers of classes and learners who have used National AI Institute for Adult Learning and Online Education (AI-ALOE) AI assistants as of June 2023.

## ARCHITECTURE FOR AI-AUGMENTED ADULT LEARNING

AI-ALOE is developing an integrated technology and data architecture for A4L. The A4L architecture is intended to support the deployment of AI assistants, collection and analysis of data on adult learners and learning, use of the data for personalization of learning in the AI assistants, and reinforcement of the teacher-learner information feedback loop. Data on learning is collected from multiple sources, including the learning management system, student log information from the AI assistants, student interactions on discussion forums, learning assessments, and class surveys. We have developed data models for the data collected from the various data sources, a data pipeline for processing the data, and a data warehouse for storing the data. AI-ALOE has partnered with 1EdTech to collect data using LTI tools and store it in the standardized form of Caliper Analytics (1EdTech, 2023). To ensure data privacy, AI-ALOE has developed machine learning techniques for anonymizing student data. In the near future, we expect to use the A4L architecture for analyzing the data and feeding the results back to teachers, learners, and the AI assistants for personalization of learning. In the medium term, we hope to share the data on adult learning we are collecting with the larger learning and education research community.

## BROADER IMPACTS

AI-ALOE targets adult learners with a broad variety of age, health, work, family, and life situations. As Figure 2 indicates, in the first 20 months of AI-ALOE, 14,202 students used AI-ALOE's technologies including 3511 students in 145 classes at TCSG, and 10,691 students in 19 online courses in Georgia Tech's OMSCS program. The use of the AI assistants also helps enhance AI literacy. There is growing evidence that research at AI-ALOE may help make online education simultaneously more available (through online learning and use of online educational materials), more affordable (through virtual teaching assistants that offload teachers' work and amplify their reach), and more achievable (through virtual learning assistants that support learners cognitively and socially), and thereby, more equitable to adult learners.

## ACKNOWLEDGMENTS

AI-ALOE is sponsored by US National Science Foundation in partnership with Accenture through Grants 2112531 and 2247790.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Ashok Goel  <https://orcid.org/0000-0003-4043-0614>

Chaohua Ou  <https://orcid.org/0000-0002-3065-2021>

## REFERENCES

- 1EdTech. 2023. Caliper Analytics. August 25, 2023. <https://www.msglobal.org/activity/caliper>.
- An, S., R. Bates, J. Hammock, S. Rugaber, E. Weigel, and A. Goel. 2020. "Scientific Modeling Using Large Scale Knowledge." In *Proceedings of the AIED'2020*, pp. 20–24.
- Goel, A. K. and D. A. Joyner. 2017. "Using AI to Teach AI: Lessons from an Online AI Class." *AI Magazine* 38(2): 48–59.
- Bloom, B. 1984. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Educational Researcher* 13(6): 4–16.
- Dede, C. and J. Richards, eds. 2020. *The 60-Year Curriculum: New Models for Lifelong Learning in the Global Digital Economy*. New York, NY: Routledge.
- Garrison, D., T. Anderson, and W. Archer. 2000. "Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education Model." *The Internet and Higher Education* 2(23): 87–105.
- Goel, A., V. Nandan, E. Gregori, S. An, & S. Rugaber. 2024. Explanation as Question Answering Based on User Guides. In S. Tulli & D. W. Aha (Eds.), *Explainable Agency in Artificial Intelligence: Research and Practice*. CRC Press.

- Goel, A. and L. Polepeddi. 2018. "Jill Watson: A Virtual Teaching Assistant for Online Education." In *Education at Scale: Engineering Online Teaching and Learning*, edited by Dede, C., J. Richards, and B. Saxberg. NY: Routledge.
- IEEE. 2020. IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being. Retrieved August 25, 2023. <https://ieeexplore.ieee.org/document/9084219>.
- Kim, M., N. Kim, G. Haddadian, and A. Heidari. 2023. "A Test of Learning Progress Models Using an AI-Enabled Knowledge Representation System." In *Proceedings of the ICLS/CSCL 2023*.
- MacLellan, C., K. Stowers, and L. Brady. 2022. "Evaluating Alternative Training Interventions Using Personalized Computational Models of Learning." *Advances in Cognitive Systems* 10: 1–18.
- National Academies of Sciences, Engineering, and Medicine. 2018. *How People Learn II: Learners, Contexts, and Cultures*. Washington, DC: The National Academies Press.
- Ou, C., D. Joyner, and A. Goel. 2019. "Designing and Developing Videos for an Online AI Class." *Journal of Online Learning* 23(2): 84–104.
- Wang, Q., K. Saha, E. Gregori, D. Joyner, and A. Goel. 2021. "Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant." In *Proceedings of the CHI 2021*, 384:1-384:14.
- Wang, Q., I. Camacho, S. Jing, and A. Goel. 2022. "Understanding the Design Space of AI-Mediated Social Interaction in Online Learning: Challenges and Opportunities." In *Proceedings of the CSCW1*, 1–26.

**How to cite this article:** Goel, A., C. Dede, M. Garn, and C. Ou. "AI-ALOE: AI for reskilling, upskilling, and workforce development." *AI Magazine* 45: 77–82. <https://doi.org/10.1002/aaai.12157>

## AUTHOR BIOGRAPHIES

**Dr. Ashok Goel** is a Professor of Computer Science at Georgia Tech and the PI and Executive Director of AI-ALOE.

**Dr. Chris Dede** is the Timothy E. Wirth Professor in Learning Sciences at Harvard University and the AI-ALOE Associate Director for Research.

**Dr. Myk Garn** is a Senior Advisor to AI-ALOE for Industrial Partnerships.

**Dr. Chaohua Ou** is the Managing Director of AI-ALOE.



**SPECIAL TOPIC ARTICLE**

# AIFARMS: Artificial intelligence for future agricultural resilience, management, and sustainability

**Vikram S. Adve**<sup>1,2,3,4</sup> | **Jessica M. Wedow**<sup>1,2,5,4</sup>  | **Elizabeth A. Ainsworth**<sup>1,6,7,8,4</sup> | **Girish Chowdhary**<sup>1,2,9,3,4</sup> | **Angela Green-Miller**<sup>1,2,9,4</sup> | **Christina Tucker**<sup>1,2,9,4</sup>

<sup>1</sup>AIFARMS National AI Institute, Urbana, Illinois, USA

<sup>2</sup>Center for Digital Agriculture, Urbana, Illinois, USA

<sup>3</sup>Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois, USA

<sup>4</sup>University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

<sup>5</sup>NCSA, Urbana, Illinois, USA

<sup>6</sup>Department of Crop Science, University of Illinois Urbana-Champaign, Urbana, Illinois, USA

<sup>7</sup>Department of Plant Biology, University of Illinois Urbana-Champaign, Urbana, Illinois, USA

<sup>8</sup>Global Change and Photosynthesis Research Unit, USDA-ARS, Urbana, Illinois, USA

<sup>9</sup>Department of Agriculture and Biological Engineering, University of Illinois Urbana-Champaign, Urbana, Illinois, USA

## Correspondence

Vikram S. Adve, AIFARMS National AI Institute, Urbana, IL, USA.  
Email: [vadve@illinois.edu](mailto:vadve@illinois.edu)

## Funding information

National Institute of Food and Agriculture, Grant/Award Numbers: 2020, 67021, 32799

## Abstract

The AIFARMS Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability national AI institute brings together over 40 world-class AI and agriculture researchers, with the common mission to develop foundational advances in AI and use them to ensure that future agriculture is environmentally friendly, sustainable, affordable, and accessible to diverse farming communities. Since its establishment in 2020, AIFARMS has advanced the state of the art in autonomous farming, cover crop planting, machine learning for improved outcomes from remote sensing, dynamic estimation of yield loss from weeds, and livestock management. The institute has prioritized the creation and utilization of high-quality, openly available data sets for advancing foundational AI and tackling agricultural challenges. AIFARMS leverages a close partnership between UIUC and Tuskegee University to build programming for a skilled and diverse next-generation workforce in digital agriculture. We are expanding the reach of AIFARMS outside of the current partners to collaborate with national AI institutions and international partners.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.



## INTRODUCTION

Traditional agriculture, such as row cropping common throughout the US Midwest, relies on significant external inputs to continually reach record-setting yields without considering the extreme environmental consequences associated with these practices. Current agricultural production causes soil degeneration, nitrogen and chemical runoff, greenhouse gas emissions, animal welfare concerns, and unsustainable labor needs. Technological advances are essential to address major challenges facing world agriculture: meeting growing demands for food, feed, fiber, and fuel under tremendous demographic pressures and climate change while simultaneously maintaining environmental quality. Two critical bottlenecks within agriculture throttle productivity and cause environmental harm:

1. The limited ability of human labor to carry out granular and scalable practices essential for sustainable, environmentally sensitive, and productive crop and livestock management.
2. Tremendous spatial heterogeneity and temporal variability in agriculture.

These critical challenges are difficult to tackle with human capacity and conventional technologies alone. Artificial Intelligence (AI) offers the potential to use machine intelligence for highly efficient seed development and advanced farm management practices, including advanced data-driven decision-making and low-cost autonomous systems capable of precise crop and livestock management activities. Moreover, autonomous reasoning techniques can accommodate the great heterogeneity and spatiotemporal variability of agriculture by analyzing vast amounts of information from diverse sources at varying scales. To be successful, however, grand AI challenges (Feigenbaum 2003) of predicting over varied spatiotemporal scales with little data, learning for low-cost autonomous robots in uncertain environments, transferring learnings between domains, integrating domain knowledge to improve explainability, and enabling effective human interaction with autonomous systems must be addressed.

## AIFARMS MISSION AND TEAM

**AIFARMS: AI for Future Agricultural Resilience, Management, and Sustainability** is a national AI Institute funded by USDA-NIFA and hosted by the Center for Digital Agriculture at UIUC, together with the University of Chicago, the Donald Danforth Plant Science Center, Michigan State University, Tuskegee University (TU), Argonne National Laboratory, and USDA-ARS. *The*

*mission of the AIFARMS Institute is to develop foundational advances in AI and use them to ensure that future agriculture is environmentally friendly, sustainable, affordable, and accessible to diverse farming communities.*

The AIFARMS team includes many AI researchers tackling a wide range of foundational AI challenges, applied to many different domains, including agriculture. The team also includes numerous prominent agriculture researchers, with deep expertise and access to large data sets crucial for success in applying AI-driven techniques to solve difficult agriculture challenges. A key outcome of the creation of AIFARMS has been to create numerous joint projects involving close collaborations centered on AI techniques for agriculture. By organizing these projects into the four research thrusts with strong common themes, supplementing them with the two cross-cutting thrusts (Figure 1), and by channeling funding, computing infrastructure, a carefully organized data management (DM) strategy, and staff support for these projects, AIFARMS is serving to further enhance the already high disciplinary expertise of the team members.

## THE ROLE OF AI: USE-INSPIRED AI RESEARCH FOR AGRICULTURE

Collaborative AI + agriculture research has numerous applications in sustainable agriculture, such as enhancing precision crop and livestock management, reducing labor requirements, high-throughput phenotyping for faster seed breeding, improving livestock health and welfare, reducing environmental impacts, and increasing climate resilience. These goals can benefit greatly from augmenting human capacity via AI-based autonomous technologies such as computer vision (CV), robotics, and recommendation systems. Such technologies can enable greater precision, greater scale, and far more efficient engineering solutions. A key constraint is to preserve low costs through scale-neutral technologies such as low-cost sensors, small robots, and drones, supported by suitable machine learning (ML) algorithms.

Moreover, data in the agriculture domain are rich in spatial and temporal heterogeneity, complex environments, and highly domain-specific characteristics. AI-driven techniques are needed to be able to tackle such data diversity, but foundational AI advances are essential to accomplish these goals.

## FOUNDATIONAL AI

AIFARMS researchers have identified seven categories of foundational AI challenges that are important to agriculture (Figures 1 and 2).

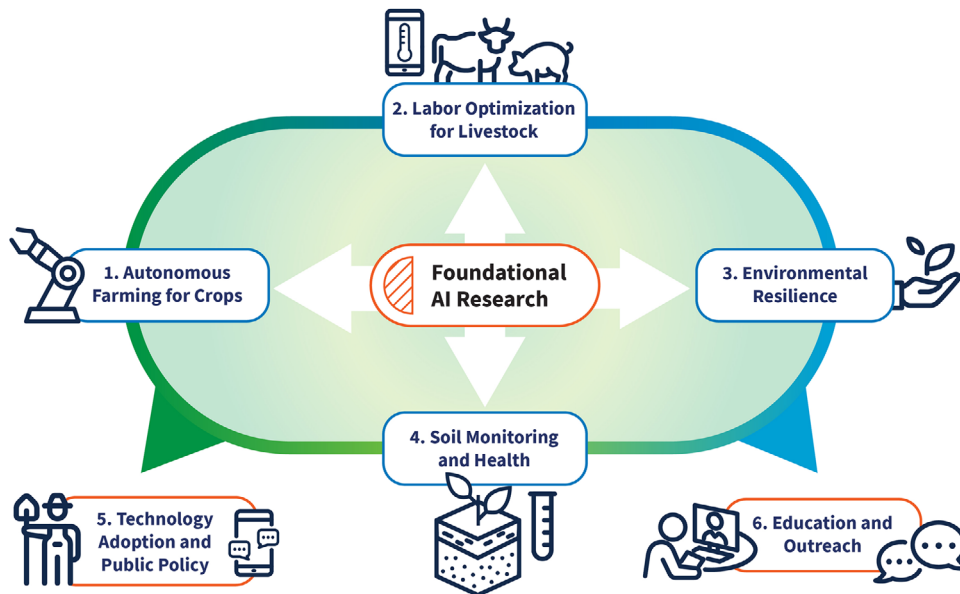


FIGURE 1 AIFARMS thrust structure.

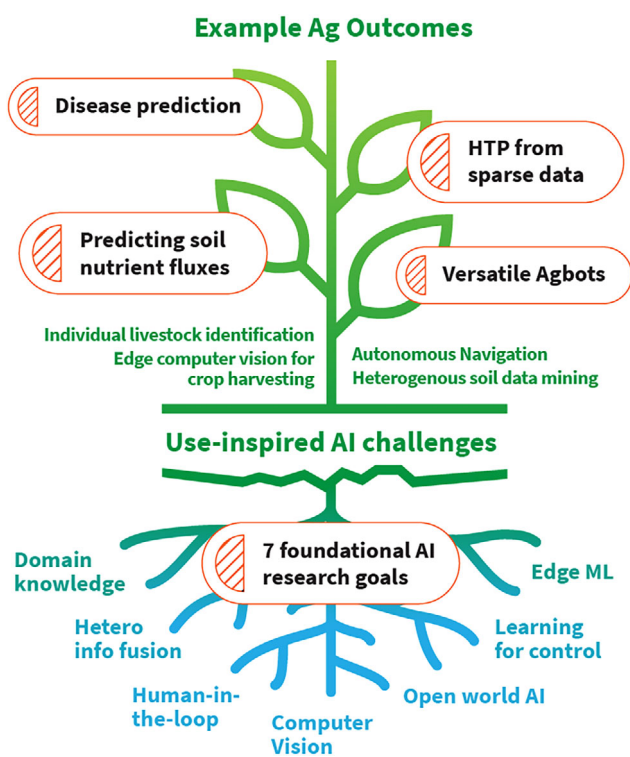


FIGURE 2 Collaborative AI + agricultural research in AIFARMS happens in three tiers. Agriculture goals are the outcomes (leaves), use-inspired AI research tasks deliver those outcomes, and foundational AI research tackles fundamental AI challenges to make use-inspired AI more effective.

CV has been extremely successful in numerous tasks ranging from facial recognition to semantic segmentation, but serious challenges remain for use cases in agriculture.

These include images with severe occlusion (e.g., leaves obscuring fruit), lack of labeled data sets, domain gaps (see below), and generalizable techniques for 3D shape and articulation (for mechanical harvesting).

**Learning for control:** Autonomous systems in agriculture (including large farm machinery, robots, and drones) must operate in uncontrolled, outdoor environments for planting, harvesting, cover crop planting, targeted spraying, plant phenotyping, and so forth. This requires major advances in learning techniques that enable robust control and reduced costs.

**Federated edge ML:** Data-driven agricultural applications, such as those using field robots or soil sensors, must operate under severe constraints on computational capacity and network connectivity. Foundational advances will enable efficient, adaptable learning in the field to support such applications.

**Open world AI:** Agricultural applications experience “domain gaps” due to variability across crops, soils, weather and rainfall patterns, pests, disease, and weed populations. Advanced ML models are needed that are robust to “open world” assumptions, through techniques such as few-shot and zero-shot learning, self-supervised learning, transfer learning, and outlier detection.

**Heterogeneous information fusion:** To optimize the performance of predictive models, it is critical to leverage effective and efficient AI techniques that can fuse information from heterogeneous sources with large spatiotemporal variability. Novel AI approaches are particularly needed for fusion techniques that are robust to intrinsic variability in reliability of sources or channels and for developing foundation models that integrate multimodal data and can be



readily used for downstream tasks such as yield prediction and phenotyping.

**Integrating domain knowledge in ML:** Agriculture has a long history of leveraging scientific research and on-farm experience for enhanced productivity. ML models will be more effective and trustworthy if they can incorporate this vast domain knowledge. While the requisite domain knowledge depends on specific tasks and models, the underlying techniques to incorporate such knowledge into ML models are in their infancy and require extensive foundational advances.

**Human-centered autonomy:** Farming is an intensely “hands-on” profession and AI-driven systems in farming need to be designed to be managed by their human counterparts. Significant technical advances are required to enable and encourage farmers to adopt autonomous systems, such as intelligent farm equipment, robots, drones, and recommendation systems, more intuitively and confidently.

## RESEARCH RESULTS TO DATE

Since its establishment, the institute has generated numerous research results associated with our core mission to make agriculture environmentally friendly, sustainable, affordable, and accessible to diverse farming communities. Several of these are beginning to show real impact.

As one example of technical contributions growing into greater success, AIFARMS teams within our autonomous farming thrust have demonstrated level 2 autonomy, defined as human on-site or nearby (“eyes off”), in field robots using LIDAR + GPS and alternatively using much cheaper and more robust CV-guided models for navigation within crop rows (Baquero et al. 2021).

An important outcome of these foundational advances in robot autonomy is that over 100 acres of cover-crops were planted well before harvest using autonomous under-canopy agricultural robots. Besides large cost reductions from using autonomous robots, cover crops planted prior to harvest have the potential for greater carbon sequestration. Integrated work with the Technology Adoption and Public Policy thrust is investigating farmers’ willingness to plant cover crops with versus without autonomous robots (Wu, Khanna, and Atallah 2023). This technology is being commercialized by EarthSense Co, which is partnering with AIFARMS team members in a new project funded by the Climate-Smart Commodities program to scale this effort up to 10,000 acres in the next 2 years.

Building on this and other successes, UIUC was awarded the USDA-NIFA Farm of the Future testbed, called I-FARM (Illinois-Farming and Regenerative Management), to accelerate the creation, maturation, and adoption of

new management technologies that are fundamentally more sustainable, profitable, affordable, and scale-neutral. This award partners with Tuskegee University and industry and collaborates closely with AIFARMS to test new technologies at the I-FARM.

Segmentation and object tracking are crucial CV tasks for numerous agricultural problems, including crop phenotyping, disease detection, weed detection, and livestock activity recognition, but video labeling for each use case is prohibitively expensive. A novel, integrated framework we have developed enables video segmentation and tracking *without training on task-specific data* (Cheng et al. 2023). It has been shown to be successful with agriculture-based challenges, such as tracking of individual animals using the PigLife data set. This framework greatly reduces the effort and cost of developing computer-vision-based solutions that require image and video segmentation.

Integrated thrust collaborations are fundamental to the work within AIFARMS. Work developed by the environmental resilience and soil monitoring thrust members generated ML algorithms using cross-scale sensing technology to integrate ground measurements, airborne hyperspectral imagery, and satellite Earth Observation. The integrated measurements scale and accurately quantify regional-scale information of management practices including cover crop, aboveground biomass, and tillage practices. This highly accurate and granular management information is important for field-level farming carbon intensity and sustainable agriculture assessment (Wang et al. 2023).

The livestock thrust has generated a unique CV dataset of labeled video and still images representing the complexities of pig production systems across different production phases, housing scenarios, and levels of occlusion (Li et al. 2023). Foundational AI advances were required to generate robust CV for fundamental tasks from tracking to recognition of pigs with various levels of occlusion, and further expansions will include pig activity recognition. The labeled data are being made available publicly as the PigLife data set (<https://data.aifarms.org/view/PigLife>).

More broadly, the AIFARMS team is developing several novel agricultural data sets for enabling AI research for agriculture, including PigLife, soils data for nutrient fluxes and for microbiome activity, open-world object detection and segmentation and open-world hyperspectral data sets for phenotype prediction. The DM team is working to format, store and preserve data sets at various stages of preparation and encouraging open availability, including a publicly available set of best practices for researchers. These set guidelines for data collection, data set and software publication, licensing guidelines, and preservation. Data sets are distributed via the AIFARMS portal (<https://data.aifarms.org/>).

## BROADER IMPACTS: CONTRIBUTION TO EDUCATION AND WORKFORCE DEVELOPMENT

The Education and Outreach thrust contributes to impactful efforts to inspire younger generations to explore digital agriculture and grow a skilled workforce. AIFARMS is dedicated to expanding workforce diversity, through our collaboration with Tuskegee University and additional programs. AIFARMS and CDA have created a successful in-person REU program, targeting students from Minority Serving Institutions, including five Historically Black Colleges and Universities (HBCUs). The program gives students research experience and career mentoring focused on increasing fluency in ML with cross-disciplinary teams. The AI Foundry for Ag Applications, a virtual week-long summer course for grad students, offers lectures and virtual activities on topics focused on AI and ML in agriculture applications, with a hackathon to implement learned skills. To expand technology skills, thrust members are currently hosting their second CS teacher endorsement cohort at UIUC for K-12 teachers across Illinois.

In Year 2, AIFARMS and CDA launched a Master of Engineering in Digital Agriculture degree, and an associated certificate program. The program's online delivery provides continued training and technology skills for working professionals, both domestic and abroad.

## AIFARMS EFFORTS TO CREATE A NEXUS FOR COLLABORATIVE AI RESEARCH

Coordination with the other four USDA-NIFA AI institutes is a vital nexus point of interinstitutional collaboration for AIFARMS. The five institutes have been working on a coordinated public messaging and outreach campaign for the importance of investment in collaborative AI-Ag research. A collaborative DM working group consists of researchers that are engaged with DM practices at the five Institutes and the NSF-funded ICICLE institute for AI cyberinfrastructure. The five USDA-funded institutes are now coordinating an AI-for-Ag Summit, planned for 2024, to showcase the agricultural domain as an ideal platform to address foundational AI challenges.

In 2022, AIFARMS launched a new international collaboration with the PhenoRob Center of Excellence in Germany. Using USDA-NIFA supplemental funding, this effort is laying the groundwork for sustained collaborations in several areas where technology sharing is beneficial. This collaboration has grown into a multi-institution international DigiCrop network (DigiCrop.net) centered

around research to reduce the negative impacts of crop production on our ecosystems without reducing yields.

UIUC is the host institution for three national AI institutes: AIFARMS, MMLI, and INVITE. AI researchers from AIFARMS and MMLI collaborate on the broad topic of generative modeling and literature mining for AI applications. The INVITE institute was just launched in 2023 with a focus on AI for K-12 education, and AIFARMS is collaborating with INVITE on educational and outreach goals common to both institutes. For example, AIFARMS has developed a “digital agriculture in a box” educational kit for K-12 after-school programs to motivate younger generations to consider digital agriculture careers, and we are working with INVITE to expand the reach of such efforts.


## ACKNOWLEDGMENTS

This work is supported by Agriculture and Food Research Initiative (AFRI) Grant No. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Jessica M. Wedow  <https://orcid.org/0000-0001-5410-9026>

## REFERENCES

- Baquero, A, V. A. Higuti, M. V. Gasparino, A. N. Sivakumar, M. Becker, and G. Chowdhary. 2021. “Multi-Sensor Fusion based Robust Row Following for Compact Agricultural Robots.” *Journal of Field Robotics* 2: 1291–319.
- Cheng, H. K., S. W. Oh, B. Price, A. G. Schwing, and J.-Y. Lee. 2023. “Tracking Anything with Decoupled Video Segmentation.” In *Proceedings of the International Conference on Computer Vision*.
- Feigenbaum, E. A. 2003. “Some Challenges and Grand Challenges for Computational Intelligence.” *Journal of the ACM* 50(1): 32–40.
- Li, J., A. R. Green-Miller, P. Senthil, T. Williams, I. C. F. S. Condotta, A. Lucic, X. Hu, et al. 2023. “PigLife: An Open-Source Image and Video Dataset for Pig Identification and Behavior for Benchmarking Computer Vision and Learning Model Applications.” Presented at the 2nd U.S. Precision Livestock Farming Conference.
- Wang, S., K. Guan, C. Zhang, C. Jiang, Q. Zhou, K. Li, Z. Qin, et al. 2023. “Airborne Hyperspectral Imaging of Cover Crops through Radiative Transfer Process-Guided Machine Learning.” *Remote Sensing of Environment* 285: 11386.
- Wu, L., M. Khanna, and S. Atallah. 2023. “Drivers and Barriers to Adopting a Cover Crop Technology Among Midwestern Farmers.” In *2023 Annual Meeting, Agricultural and Applied Economics Association*.





**How to cite this article:** Adve, VS, JM. Wedow, EA. Ainsworth, G. Chowdhary, A. Green-Miller and C. Tucker. 2024. "AIFARMS: Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability." *AI Magazine* 45: 83–88. <https://doi.org/10.1002/aaai.12152>

## AUTHOR BIOGRAPHIES

**Vikram Adve** is a Professor of Computer Science at the UIUC, Co-Founder and Co-Director of the Center for Digital Ag, and PI of AIFARMS. Adve's research aims to improve the programmability and performance of computer systems.

**Jessica Wedow** is the Executive Director for the AIFARMS National AI Institute at the UIUC and is a member of the Center for Digital Agriculture (CDA) and the National Center for Supercomputing Applications (NCSA).

**Elizabeth Ainsworth** is the Research Leader of the Global Change and Photosynthesis Research Unit with the USDA-ARS. Ainsworth's research aims to improve crop responses to global climate change.

**Girish Chowdhary** is an Associate Professor of Robotics at the UIUC. He is a Leading Researcher in algorithms and systems for autonomous field robots. He is also the Co-founder of EarthSense Inc.

**Angela Green-Miller** is an Associate Professor of Agricultural and Biological Engineering at the UIUC. She addresses production challenges, translating technology into animal applications for better management.

**Christina Tucker** is the Executive Director of CDA and in the Agricultural and Biological Engineering department at the UIUC. She is also the Director of Education for the CDA and AIFARMS.



## SPECIAL TOPIC ARTICLE

# The AIFS Institute: Building a better food system through AI

Ilias Tagkopoulos<sup>1,2,3</sup> | Mason J. Earles<sup>1,4,5</sup> | Danielle G. Lemay<sup>1,6,7</sup> | Xin Liu<sup>1,2</sup> | Nitin Nitin<sup>1,8</sup> | Aaron D. Smith<sup>1,9</sup> | Tarek I. Zohdi<sup>1,10</sup> | Stephen F. Brown<sup>1</sup>

<sup>1</sup>USDA-NIFA/NSF AI Institute for Next-Generation Food Systems (AIFS), Davis, California, USA

<sup>2</sup>Department of Computer Science, University of California, Davis, California, USA

<sup>3</sup>Genome Center, University of California, Davis, California, USA

<sup>4</sup>Department of Viticulture & Enology, University of California, Davis, California, USA

<sup>5</sup>Department of Biological and Agricultural Engineering, University of California, Davis, California, USA

<sup>6</sup>Agricultural Research Service, U.S. Department of Agriculture, Western Human Nutrition Research Center, Davis, California, USA

<sup>7</sup>Department of Nutrition, University of California, Davis, California, USA

<sup>8</sup>Department of Food Science and Technology, University of California, Davis, California, USA

<sup>9</sup>Department of Agricultural and Resource Economics, University of California, Davis, USA

<sup>10</sup>Department of Mechanical Engineering, University of California-Berkeley, Berkeley, California, USA

### Correspondence

Ilias Tagkopoulos, USDA-NIFA/NSF AI Institute for Next-Generation Food Systems (AIFS), Davis, California, USA.  
Email: [itagkopoulos@ucdavis.edu](mailto:itagkopoulos@ucdavis.edu)

### Funding information

AFRI Competitive, Grant/Award Number: 2020-67021-32855; National Institute of Food and Agriculture, Grant/Award Number: 1024262

### Abstract

Our food system is complex, multifaceted, and in need of an upgrade. Population growth, climate change, and socioeconomic disparities are some of the challenges that create a systemic threat to its sustainability and capacity to address the needs of an evolving planet. The mission of the AI Institute of Next Generation Food Systems (AIFS) is to leverage the latest advances in AI to help create a more sustainable, efficient, nutritious, safe, and resilient food system. Instead of using AI in isolation, AIFS views it as the connective tissue that can bring together interconnected solutions from farm to fork. From guiding molecular breeding and building autonomous robots for precision agriculture, to predicting pathogen outbreaks and recommending personalized diets, AIFS projects aspire to pave the way for infrastructure and systems that empower practitioners to build the food system of the next generation. Workforce education, outreach, and ethical considerations related to the emergence of AI solutions in this sector are an integral part of AIFS with several collaborative activities aiming to foster an open dialogue and bringing closer students, trainees, teachers, producers, farmers, workers, policy makers, and other professionals.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.

## INTRODUCTION

Throughout human existence, food has been much more than a necessity. It has been an art, a way of expression, a part of our cultural inheritance, and evolution. It is not a coincidence that almost every culture in antiquity deified food or its process: from Demeter and Ceres, to Freyr and Inari, agriculture and food have been revered and adored, with sacred ceremonies, such as the Eleusinian mysteries in Ancient Greece, celebrating agriculture and the cycle of life (Mark, 2012).

As our world became bigger and more interconnected, so did the challenges that our food system faced. How do you feed a planet of 10 billion people, eliminating distribution and access inequalities, while keeping production sustainable, less wasteful, and resilient to disruptions such as climate change and loss of biodiversity? Concomitantly, the waves of agricultural revolution over the past couple of centuries with industrialization, monocultures, fertilizers, and advanced technology for crop management, coupled with globalization of food supply and interconnected global markets, have enabled unprecedented growth opportunities. Artificial Intelligence (AI) is regarded as a unique opportunity to accelerate innovation in the broader food system in a way that can address inefficiencies and bridge gaps in our food supply chain.

AIFS shares this view, and pioneers efforts to apply AI technologies to each stage of the food system, leveraging data and models to improve efficiencies of the companies and workers, and the safety and health of the consumers of our food system as a whole. At the front end of the food system, we are using molecular breeding to accelerate desirable crop traits including nutrition. In agricultural production, we are leveraging large agricultural datasets and building models to advance precision agriculture, and are also building models to optimize indoor farming. For processing, we are developing machine learning models to enhance the inactivation of pathogens and improve process validation at processing facilities. In the knowledge discovery and synthesis realm, we are using Deep Learning models and the latest in Natural Language Processing to build knowledge bases that ingest knowledge from a plethora of sources and millions of published papers so that the food, ingredients, chemicals, and health areas are intimately and interoperably interlinked, mined, and used in applications such as diet recommendation systems and bioactives discovery.

## RESEARCH CLUSTERS

AIFS has six research clusters (see Figure 1). Four of these (molecular breeding, agricultural production, food pro-

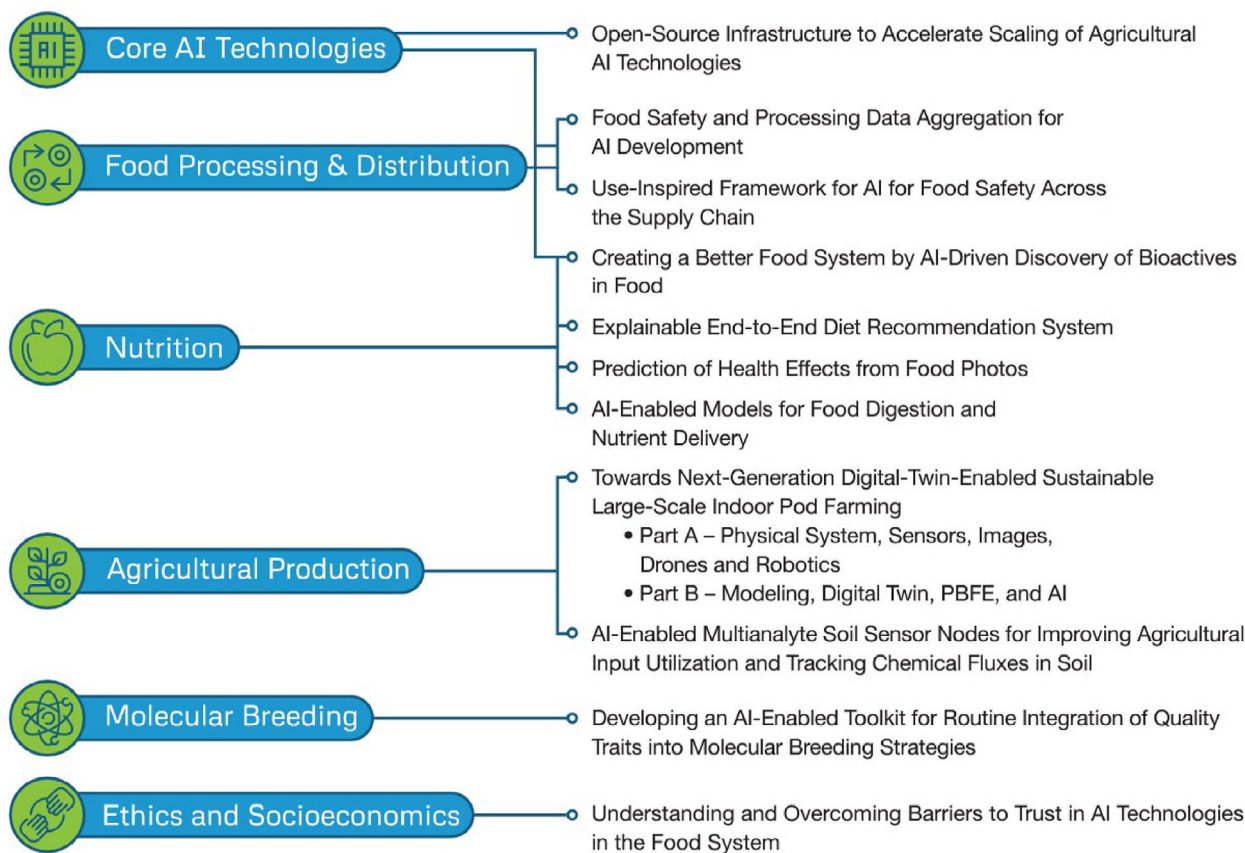
cessing, and nutrition) are focused on discrete areas of the food system, whereas two more research clusters, core AI, and ethics are overarching. AIFS provides competitive funding for approximately 40 researchers that work in interdisciplinary projects to apply AI methodologies that advance AI infrastructure in food systems, promote human health, and help create a more resilient supply chain. Below are some of the project highlights for each cluster.

*Molecular breeding* is a technology that can accelerate the rate at which a new line of a cultivar can arise from precise selection of which crosses to make during plant breeding. With an aim toward improvements in plant traits such as nutrition and climate adaptations, the AIFS team is developing a toolkit for integrating nutritional quality and aroma traits into molecular breeding strategies. Furthermore, the molecular breeding teams develop high-throughput assays and screening large cultivar collections via imaging and rapid sensors, to screen targeted cultivars for targeted molecules, leading to models to predict molecular composition and quality, among other traits.

The *agricultural production* cluster aims to deliver low-cost and easy to access sensing and data solutions. It develops and tests in real-use conditions novel inexpensive wireless AI-enabled sensors for accurately measuring agriculturally relevant soil signals such as nitrate, ammonium, phosphate, potassium, and moisture. The cluster also develops an open-source Python library (AgML), that enables access to agricultural-specific machine learning datasets, benchmarks, pretrained models, and workflows (Choi et al., 2023). Researchers also adapt a highly sophisticated 3D biophysical crop modeling tool (HELIOS) as a machine learning generator for synthetic sensor data. Several teams are investigating the use of reinforcement learning and model-based controls for autonomous navigation of ground- and aerial-robots in complex 3D environments like vineyards and almond orchards.

In the area of *food processing*, our team has been developing AI-based framework and models to enhance food safety, reduce food spoilage, and improve understanding of the role of food structure in enhancing nutrition and health benefits of food. In the area of food safety, the research projects have focused on enhancing the detection of pathogens (Yi et al., 2023) (Ma et al., 2023), improving food processing technologies using digital twins and machine learning models to enhance the inactivation of pathogens and improve process validation. The results of these studies demonstrate that AI models for image detection and classification and spectral data analysis can improve the speed and specificity of detecting target bacteria in complex food systems and aid in validation of novel processing technologies for the inactivation of target pathogens. Other work includes the development

# AIFS Year 3 Projects / Research Thrusts



**FIGURE 1** The six AIFS research clusters and the current projects that span across them. All projects are around the three AIFS thrusts of building the AI infrastructure for the food system, a resilient supply chain, and improving human health through food. AI, Artificial Intelligence; AIFS, AI Institute of Next Generation Food Systems.

of agent-based models to identify risk of the spread of pathogens in food facilities, an AI-based 3D printing model for bioinks for micronutrient release, and a digital twin for indoor farming systems (Meng et al., 2023).

The *nutrition* cluster works in two main thrusts: developing AI methods to use food photos to track and assess what we eat in real-time, and to deliver dietary recommendations based on user preferences and nutritional status. The team is working toward a food “photo-to-ingredient” module and an “ingredients to health outcome” module, each of which will be needed in most computational systems in nutrition. The team has developed key resources such as the SNAPMe Benchmark Database for the evaluation of algorithms for computer vision in dietary assessment and the Food Atlas, a KnowledgeBase that uses deep learning, natural language processing, large language models, and other techniques to connect foods, ingredients, compounds, and health effects. Applications include the prediction of food composition after processing (Naravane & Tagkopoulos, 2023), algorithmic food matching and

tolerance methods (Eetemadi & Tagkopoulos, 2023), automated ontology construction for food (Youn, Naravane, & Tagkopoulos, 2020), and causal prediction of dietary intervention efficacy in digestive diseases (Eetemadi & Tagkopoulos, 2021).

In *application-inspired and foundational AI*, the teams are researching applications in Federated Learning (FL) which is a distributed model training paradigm where clients collectively train a model while keeping their local data private. AIFS can serve the role of a central server because it has no direct conflict of interests with (industrial) entities. Explainability is crucial for the acceptance and adoption of AI-based solutions as in order to confidently use an AI system, it needs to be trusted, and in order to be trusted, it needs to be explainable. Concomitantly, the team is working in the field of AI-driven indoor farming, focusing on developing digital-twin technologies for the optimal optical design of sustainable large-scale indoor farming “pods,” which consist of enclosed trailers with hydroponically grown plants, with energy supplied



by carefully controlled LEDs, whereby a continuous and constant supply of nutrient-rich water helps plants grow more consistently and while reducing inputs. The objective is to develop simulation tools to drive innovative solutions involving vertical farming panels of plants, utilizing LED light strip “walls,” which provide crops with light from only the part of the spectrum required for growth.

The *workforce and ethics* team has developed an ethics framework around the principle that rules and compliance regulations are unlikely to solve the most difficult ethical challenges in AI. It is important that AI researchers and developers follow rules to, for example, protect the privacy of their data and honestly represent their products. Our ethics framework is built around a decisions and recommendations log in which researchers record their decisions and actions and their reasons for taking them with the goal that researchers make mindful decisions (Alexander et al, 2023). The framework emphasizes transparency, vigilance, and clear communication. In studying the ethical and socioeconomic challenges presented by AI in the food system, we ask three questions: (i) who wins and who loses from the technology? (ii) who bears the risks of bad outcomes? and (iii) who decides the answers to these questions? Workers on farms and in food processing plants will need to develop new skills to thrive in the labor market. Helping current workers and communities thrive is an important challenge we are working on.

## EDUCATION AND PUBLIC ENGAGEMENT

Training and outreach is at the core of AIFS’ mission, with the institute sponsoring various fellowships and training initiatives. One such initiative that targets researchers that are not from computer science is AIBridge, where researchers with limited prior programming experience get educated and then apply what they learn on AI projects related to food and agriculture. Over the past three years, hundreds of students have been trained on how to use Python, object oriented concepts, and machine learning libraries in practice, with some of them moving on to pursue this interdisciplinary field for their careers. The AIFS Career Exploration Fellowship Program is another initiative where companies mentor undergraduate students with internships, where students gain hands-on experience in an industrial setting. More broadly, AIFS regularly participates in conferences and competitions, creates various programs from K12 to postgraduate/professional level with various partners, and sponsors various events, such as the Apps for Food & Ag Hackathon and the Farm Robotics Challenge in which 19 teams from 12 universities built robotic solutions to farm tasks. All these initiatives have

resulted in bringing together and training thousands of students and professionals from various industries and paths in life to collectively think and design a better, faster, and more resilient food system, with AI at its center.

## COLLABORATIONS AND SYNERGIES

It takes a village to transform a system, and the food system is no different. Over the years, we have entered into close partnerships and joined activities across the Atlantic to co-sponsor DigiCrop 2022 with PhenoRob in Germany, and held joint activities with Fraunhofer ISE. Through the USDA’s programs for funding 1890 (historically black land-grant) institutions, we have established collaborations with Delaware State University, Tennessee State University, Florida A&M University, Prairie View A&M University, and West Virginia State University. In terms of industrial partnerships, AIFS has also engaged with over 50 companies regarding the challenges facing the agriculture sector and opportunities for AI. From small startups to large CPG and ingredient companies, AIFS has partnered and co-sponsored projects in various topics that range from understanding the molecular composition of milk to creating AI tools related to knowledge organization and molecular discovery of key compounds such as polyphenols and terpenes, by using omics, flavor, and health effect predictors.

## CONCLUSION

To conclude, through pioneering projects and transformative collaborations, AIFS is driven by a vision that transcends the ordinary. We aim to sculpt a society where wellness takes precedence over illness, a world where we nurture our health through accessible, nourishing food that confronts the very origins of diseases, notably chronic inflammation. Our commitment extends from the fertile fields to the very plates we eat from, all while honoring the sanctity of sustainable practices that safeguard our precious planet and its delicate ecosystems. We strive to be stewards of our inheritance, preserving these invaluable resources for the generations that will follow in our footsteps.


## ACKNOWLEDGEMENTS

This work is supported by AFRI Competitive Grant no. 2020-67021-32855/project accession no. 1024262 from the USDA National Institute of Food and Agriculture. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their affiliated institution or agency.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Ilias Tagkopoulos  <https://orcid.org/0000-0003-1104-7616>

## REFERENCES

- Alexander, C. S., Yarborough, M., & Smith, A. (2023). Who is responsible for “responsible AI”? Navigating challenges to build trust in AI agriculture and food system technology. *Precision Agriculture* 25(1): 146–85. <https://doi.org/10.1007/s11119-023-10063-3>
- Choi, T., D. Guevara, G. Bandodkar, Z. Cheng, C. Wang, B. Bailey, M. Earles, and X. Liu. 2023. “DAVIS-Ag: A Synthetic Plant Dataset for Developing Domain-Inspired Active Vision in Agricultural Robots.” arXiv: 2303:05764.
- Eetemadi, A., and I. Tagkopoulos. 2023. “Algorithmic Lifestyle Optimization.” *Journal of the American Medical Informatics Association* 30(1): 38–45.
- Eetemadi, A., and I. Tagkopoulos. 2021. “Methane and Fatty Acid Metabolism Pathways Are Predictive of Low-FODMAP Diet Efficacy for Patients with Irritable Bowel Syndrome.” *Clinical Nutrition* 40(6): 4414–21. <https://doi.org/10.1016/j.clnu.2020.12.041>.
- Ma, L., J. Yi, N. Wisuthiphaet, M. Earles, and N. Nitin. 2023. “Accelerating the Detection of Bacteria in Food Using Artificial Intelligence and Optical Imaging.” *Applied and Environmental Biology* 89(1), <https://journals.asm.org/doi/full/10.1128/aem.01828-22>.
- Mark, J. “The Eleusinian Mysteries: The Rites of Demeter.” *The World History Encyclopedia*, January, 2012.
- Mengi, E., Becker, C. J., Sedky, M., Yu, S. Y., & Zohdi, T. I. (2023). A digital-twin and rapid optimization framework for optical design of indoor farming systems. *Computational Mechanics*: 1-13.
- Naravane, T., and I. Tagkopoulos. 2023. “Machine Learning Models to Predict Micronutrient Profile in Food after Processing.” *Current Research in Food Science* 6: 100500.
- Yi, J., N. Wisuthiphaet, P. Raja, N. Nitin, and J. Earles. 2023. “AI-Enabled Biosensing for Rapid Pathogen Detection: From Liquid Food to Agricultural Water.” *Water Research* 242: 120258. <https://doi.org/10.1016/j.watres.2023.120258>.
- Youn, J., T. Naravane, and I. Tagkopoulos. 2020. “Using Word Embeddings to Learn a Better Food Ontology.” *Frontiers in Artificial Intelligence* 3: 584784. Sec. AI in Food, Agriculture and Water. <https://doi.org/10.3389/frai.2020.584784>.

**How to cite this article:** Tagkopoulos, I., M. J. Earles, D. G. Lemay, X. Liu, N. Nitin, A. D. Smith, T. I. Zohdi, and S. F. Brown. 2024. “The AIFS Institute: Building a better food system through AI.” *AI Magazine* 45: 89–93. <https://doi.org/10.1002/aaai.12164>

## AUTHOR BIOGRAPHIES

**Ilias Tagkopoulos**, PhD, is the AIFS Director, AIFS Principal Investigator (PI), and a professor of Computer Science and the UC Davis Genome Center at the University of California-Davis. He has founded various companies that develop or leverage AI for healthy snacking, molecular discovery, and industrial optimization.

**Mason J. Earles** is assistant professor in the Departments of Viticulture & Enology and Biological and Agricultural Engineering at the University of California-Davis, and is AIFS co-PI and agricultural production cluster co-lead.

**Danielle G. Lemay**, PhD, is a Research Scientist at the USDA Western Human Nutrition Research Center in Davis, California. She is also an Associate Adjunct Professor with the Department of Nutrition and faculty member of the Genome Center at University of California-Davis and the AIFS nutrition cluster lead.

**Xin Liu**, PhD, is a Professor of Computer Science at the University of California-Davis and a co-PI and core AI cluster lead with AIFS.

**Nitin Nitin**, PhD, is vice-chair of the department of Food Science and Technology at the University of California-Davis, and is a co-PI and the food processing and distribution cluster co-lead with AIFS.

**Aaron D. Smith**, PhD, is the DeLoach Professor of Agricultural and Resource Economics at the University of California Davis, and is also the AIFS Ethics and Socioeconomics cluster lead.

**Tarek I. Zohdi**, PhD, is a distinguished professor in the Department of Mechanical Engineering at the University of California-Berkeley, and AIFS co-PI, education cluster lead, and food processing and distribution cluster co-lead.

**Stephen F. Brown**, Ph.D. is Associate Director of AIFS based at the University of California-Davis, and PI for AIVO, the AI Institutes Virtual Organization that is tasked to bring together the various NSF and USDA institutions.



## SPECIAL TOPIC ARTICLE

# AIIRA: AI Institute for Resilient Agriculture

Baskar Ganapathysubramanian<sup>1</sup>  | Jessica M. P. Bell<sup>1</sup> | George Kantor<sup>2</sup> |  
 Nirav Merchant<sup>3</sup> | Soumik Sarkar<sup>1</sup> | Patrick S. Schnable<sup>1</sup> | Michelle Segovia<sup>4</sup> |  
 Arti Singh<sup>1</sup> | Asheesh K. Singh<sup>1</sup>

<sup>1</sup>Iowa State University, Ames, Iowa, USA

<sup>2</sup>Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

<sup>3</sup>University of Arizona, Tucson, Arizona, USA

<sup>4</sup>University of Missouri, Columbia, Missouri, USA

## Correspondence

Baskar Ganapathysubramanian, Iowa State University, Ames, IA, USA.  
 Email: [baskarg@iastate.edu](mailto:baskarg@iastate.edu)

## Funding information

National Institute of Food and Agriculture, Grant/Award Number: 2021-67021-35329.

## Abstract

**AIIRA** seeks to transform agriculture by creating a new AI-driven framework for modeling plants at various agronomically relevant scales. We accomplish this by designing and deploying AI-driven predictive models that fuse diverse data with siloed domain knowledge. **AIIRA**'s vision, illustrated in Figure 1, consists of four technical thrusts with cross-cutting education, training, and outreach activities. Our activities are focused on theory, algorithms, and tools for the principled creation of goal-oriented AI tools deployed at plant and field scales. Our use-inspired AI developments are tightly integrated with USDA-relevant challenges in crop improvement and sustainable crop production. Our strong social science focus ensures sustained AI adoption across the ag value chain. Our cyberinfrastructure (CI) efforts ensure cohesive, sustainable, and extensible CI to reproducibly share and manage data assets and analysis workflows to a diverse spectrum of the Ag community. Taken together, this will ensure long-term payoffs in AI and agriculture. **AIIRA** has established a new field of *Cyber Agricultural Systems* at the intersection of plant science, agronomics, and AI. Our signature activities build the workforce for this new field through formal and informal educational activities. Through these activities, **AIIRA** creates accessible pathways for underrepresented groups, especially Native Americans and women.

## OVERVIEW

The **AI Institute for Resilient Agriculture (AIIRA)** has an integrated vision for foundational advances in AI that enhance the *resiliency of our agricultural ecosystem* (Sarkar et al. 2023). Agricultural resiliency—the ability to provide food, feed, fuel, and fiber in a sustainable manner in the face of systematic uncertainty—is critical, particularly with ever-increasing climate variability, increased demand, demographic shifts, labor shortages,

and disruptions to global trade. We develop the concept of AI-driven goal-oriented *digital twins* to assimilate copious sensor data with genetic, physiological, and agronomic knowledge (*domain knowledge*) and thereby model agricultural phenomena at the plant and field scales. The availability of such *open source* tools will dramatically improve the ability to develop testable hypotheses, anticipate problems, enable future planning, explore potential solutions, and mitigate undesirable emergent behavior in complex agricultural systems (see Figure 1).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Overview of AIIRA activities.

**Need and timeliness:** An increasing world population, coupled with finite arable land, changing diets, and the growing expense of agricultural inputs, is poised to stretch our agricultural systems to their limits. By the end of this century, the Earth's population is projected to increase by 45% with available arable land decreasing by 20%; this creates the urgent need to enhance agricultural productivity by 70% before 2050. Current rates of progress are insufficient, making it impossible to meet this goal without a technological paradigm shift. Furthermore, consider the following challenges: *Environmental variability:* Increasing occurrences of weather extremes (droughts and floods; high and low temperatures) and pestilence (insects, weeds, and pathogens) fundamentally limit crop yields. AI approaches for identifying *genetic* factors that endow resiliency against these biotic and abiotic stressors and efficiently *breeding* germplasm for resiliency traits is an appealing strategy. *Input use efficiency:* The application of fertilizers, pesticides, and growth in agricultural irrigation has dramatically increased yields and decreased yield variability. However, these inputs are expensive, energy intensive, and often inefficiently used by crops and lost to runoff, with adverse impacts on water quality. This calls for innovations in *breeding* for germplasm that exhibits enhanced input use efficiency and *ag production* practices that minimize waste, both of which offer opportunities for AI development. *Economic risk:* Increasing costs (seeds, land, and heavy equipment), reduced availability (water and some fertilizers), and decreasing profitability of farming cannot be addressed solely through new agronomic practices but need well-adapted cultivars that maintain or improve yield as climate changes. AI innovations in *breeding* for a future climate can mitigate economic risk of crop failure and can transform breeding of crops that are traditionally underfunded. *Demographic shifts and*

*social acceptance:* While there is substantial promise of AI technology in agriculture, *stakeholder acceptance and adoption* is not guaranteed, thus limiting broad societal impact. Deeply embedded social scientists can provide an understanding of catalysts and barriers to adoption across the stakeholder value chain—breeders, producers, industry, and consumers—to generate a virtuous feedback cycle with foundational AI developments.

A predictive **digital twin** is a virtual simulation of connected biological entities that assimilates sensor updates to mirror the life cycle of its corresponding biological system. We target the principled creation of goal-oriented, AI-driven predictive models at the *individual plant scale*, as well as the *plot and field scale*. The advancements directly apply to climate resiliency, sustainability, and producer profitability, coupled with a social science thrust to maximize stakeholder understanding, trust, and acceptance.

**AI challenges:** AIIRA has focused on the AI themes below to create predictive tools—that generalize across crops and geography—at the plant and field scales. We cite selected recent works, with a detailed bibliography of results available on the AIIRA webpage. *Learning with limited data:* Given the scale and complex nature of agricultural ecosystems, traditional AI/ML approaches to training predictive models with sufficient accuracy require an intractable amount of data. To address this, AIIRA explores various approaches for doing more with less (Chiranjeevi et al. 2023; Feuer et al. 2023). *Data, knowledge, and model fusion:* Domain knowledge is currently encoded in multiple forms that capture various aspects of biology, physics, and agriculture. We explore principled approaches for data, knowledge, and model fusion that enable conceptualization and training (Chang et al. 2023). *Robustness, generalizability, and privacy:* Once trained, the models must be robust to variations in weather and sensor data, and must respect privacy (training, inference) as needed (Cho et al. 2022). *Data relevancy:* Current best practices for mapping sensor measurements to relevant plant characteristics are largely ad hoc. We are pursuing novel sensing modalities as well as AI-enabled hardware for identifying and characterizing relevant plant features (Ibrahim et al. 2022). *Continual distributed learning:* The geographically dispersed nature of predictive models at the field scale motivates a computational approach that is naturally distributed (Esfandiari et al. 2021). We are exploring approaches for continuously updating models with partial, multiscale, and multimodal data. *Ability to encounter new and challenging scenarios:* We are devel-

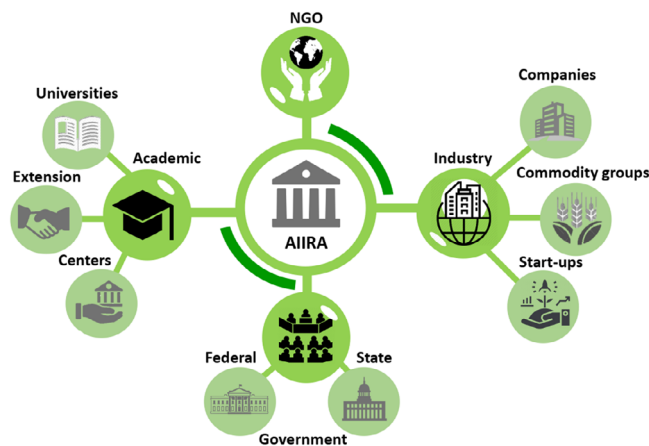


oping approaches that yield useful guiding principles even with coarse or sparsely sampled data. Our approaches promote the generalizability of the trained model in hitherto unseen scenarios. *Connecting AI to action*: Active data collection and physical interventions are a critical part of building and deploying the digital twin. We are creating novel intelligent robotic systems that facilitate flexible monitoring, timely interactions, and intelligent tactile plant manipulations (Freeman et al. 2023).

**Long-term goals and impacts:** The overarching goal of **AIIRA** is to democratize the power of AI-based prediction for diverse agricultural applications, nurture and grow the next generation of talent, and enable stakeholders worldwide to increase productivity and reduce environmental impact: *Scientists* will use the goal-oriented models to gain revolutionary new insights into genetics and plant physiology, to increase the efficiency of *breeding*, and to more rapidly identify and select for desired traits to adapt to changing environments and emerging diseases and pests. *Farmers* or their advisors will use digital twin modeling to optimize planting decisions and applications of fertilizer, irrigation water, and pesticides to maximize profit while minimizing environmental impact. As early adopters, US farmers are expected to preferentially benefit from these new technologies. *Governments* will use digital twin predictions to identify which crops and practices are best suited on regional and national scales, set incentive policies, and undertake regional/national pest surveillance. *Ag companies* will use digital twins to create more effective precision management products and produce climate-resilient germplasm. *The rural entrepreneurship ecosystem* will be spurred by the use-inspired AI innovations, workforce development, education, and knowledge transfer activities.

**AIIRA as an AI-Ag nexus:** **AIIRA** engages AI experts (including computer scientists, statisticians, data scientists, and robotics engineers) with geneticists, crop breeders, agronomists, and social scientists (spanning economics, business, education, ethics, and sociology) with established records of collaboration. Our diverse stakeholder network (Figure 2)—consisting of industry, academic, government and nongovernmental organizations (NGOs), NGOs that work in AI literacy, and extension and producer groups—has enabled the intellectual output of **AIIRA** to train a globally aware, AI-literate, and diverse workforce.

Our diverse team includes world experts in AI, agriculture, and their intersection: (1) *Universities*: Iowa State University (ISU, lead), Carnegie Mellon University (CMU), University of Nebraska—Lincoln (UNL), New York University (NYU), George Mason University (GMU), University of Missouri (UM), and University of Arizona (UA); (b) *Industries*: Microsoft, Google X (Mineral), Deere, Corteva,



**FIGURE 2** AIIRA will orchestrate activities across multiple organizations to create a collaborative nexus that results in a resilient agricultural ecosystem.

Bayer, as well as several start-ups spanning the AI and ag ecosystems; and (c) *Government, commodity groups, and NGOs*: USDA-Agriculture Research Service groups, Iowa Economic Development Agency (IEDA), Data and Software Carpentries, North American Plant Phenotyping Network (NAPPN), various commodity groups (corn, soybean, fruits and vegetables). This collaborative nexus spans nine states across diverse environments, crops, and stakeholders. Our team composition (33% women, 50% people of color, 10% URM) ensures that we bring a diverse spectrum of expertise and insight to leverage AI advances.

**AIIRA** team members initiated the earliest applications of AI/ML to the cyber-agricultural ecosystem. We have pioneered the training of a diverse workforce that bridges AI/ML with agriculture via *NSF NRT, Data Science for Public Good*, and *Women in Ag and AI* programs. Our cyberinfrastructure (CI) platform leverages the NSF-funded CyVerse and democratizes AI innovations to the ag community. Team members have leadership roles in NAPPN, are laying the foundations for the new field of *cyber-agricultural systems*, and are training our next generation of scientific leaders and policymakers.

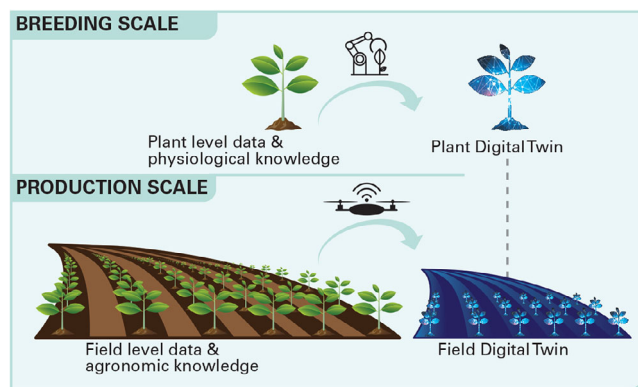
## BROAD RESEARCH APPROACHES

**A digital twin** of the life of a plant or field will give us the understanding and tools needed to increase the resiliency of agriculture and to break through the ceiling on agricultural productivity and profitability. This effort is akin to other applications involving large-scale data analytics, such as (a) model-based weather prediction that continuously incorporates new data, and (b) longitudinal economic models with personalized financial plans for each individual.

**The digital twin of an individual plant:** As with any living entity, a plant's development depends on a wide range of complex and intertwined factors. The properties of a plant (its phenotype,  $P$ ) are heavily influenced by its genotype,  $G$ ; the prevailing environment,  $E$  (soil, weather); management actions,  $M$ , by the grower (water, fertilizer/pesticide usage, pruning); and the complex interactions among these factors ( $G \times E$ ,  $G \times M$ , etc.). Much is known based on experimentation about how  $P$  is determined by  $G$ ,  $E$ , and  $M$ . However, this knowledge is highly siloed, heavily parameterized (requiring calibration for ad hoc plant parameters), and usually not transferable to new situations. The broad availability of *AI-driven predictive models* that continuously assimilates diverse data to predict the development of a plant given  $G$ ,  $E$ , and  $M$  will transform how agriculture is managed. There is no *complete knowledge model* for the development cycle of a plant. A purely data-driven approach is infeasible given the high-dimensional nature of the problem and the requirement for exponential data sets for every single plant. Instead, we integrate knowledge-based models with sensor-based data. We consider knowledge models to represent the *known (but incomplete and/or uncertain) dynamics* of a plant's development. This includes well-established principles of water, gas, nutrient, and carbohydrate transport; energy balance; and biochemistry, photosynthesis, and genetic factors. The diverse set of sensors that measure the physiological state (directly or indirectly) provide an incomplete snapshot of the plant, accounting for both modeled and unmodeled dynamics.

**Ag impact: An open-source plant-scale predictive model** will transform *plant genetics and breeding*. Step changes include (a) developing testable hypotheses, designing and conducting experiments to enhance both our understanding of crop biology and the accuracy of digital twins; (b) reducing breeding costs and increasing genetic gain by allowing more virtual screening; and (c) enabling the identification of optimal ideotypes needed, for example, to enhance the water-use efficiency (WUE) or nitrogen-use efficiency (NUE) of crops.

**The digital twin of a field:** No two plants in a field are truly isolated from each other. Neighboring plants shade each other and compete for resources. The variability in the microclimate (soil, weather, sunlight, topography) with heterogeneity in nutrient/water availability can produce significant variability in growth, development, and yield and thus necessitate variable management decisions. We utilize a layered sensing strategy, including drones and proximal sensors (see Figure 3), to provide partial information on the state of the field, including soil health (topography, soil type, organic matter, nutrient, and water availability) and plant development state. As with the plant-scale predictive model, the field-scale models learn to predict the state of the field given partial,



**FIGURE 3** Our principled nested approach to building multiscale predictive models. At higher scales, multiscale, multimodal data are fused with models of lower scales and domain knowledge.

multiscale, multimodal, and multifidelity data along with knowledge models. Here, the knowledge models consist of (a) the plant-scale predictions, as well as (b) field-scale environmental models that include soil, microclimate, and bio-eco-hydrological process models.

**Ag impact: An open-source field-scale predictive model** will transform *crop improvement and ag production* operations. Step changes include (a) significantly reducing irrigation/fertilizer/pesticide costs by identifying and mitigating hyper-localized problem spots; (b) distributed autonomous crop management to make agriculture profitable and safe; and (c) in silico approaches to complement multi-environment trials for breeding, thus providing capabilities such as scale-up and the ability to simulate future climates. The low cost of these benefits will be especially impactful for crops that are not well-funded.

**The digital twin as a tool to broaden engagement and promote acceptance of AI technology:** A digital twin will provide an easily accessible platform to help people better understand both plant science and the potential impacts of AI. Digital twin simulations can be presented and “dissected,” allowing experts and nonexperts alike to explore the interrelationships among the components that drive plant growth, crop management, or policy. Users can perform “what if” experiments that vary parameters such as climate, input regime, and plant physiology and then immediately observe the predicted result in terms of plant productivity, environmental impact, or economic gain. These capabilities will help build trust and promote acceptance in AI technology. Finally, the digital twin can be used to engage students in agricultural early in their training and thereby grow the AI-Ag workforce.

## ACKNOWLEDGMENTS

The full **AIIRA** team (<https://aiira.iastate.edu/about-us/faculty-team/>) is gratefully acknowledged for shaping and working towards this ambitious vision. This material is




based upon work supported by the AI Research Institutes program supported by NSF and USDA-NIFA under AI Institute: for Resilient Agriculture, Award No. 2021-67021-35329. We thank Tina Rice for the graphics design. The findings and conclusions in this article have not been formally disseminated by the U. S. Department of Agriculture and should not be construed to represent any agency determination or policy.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Baskar Ganapathysubramanian  <https://orcid.org/0000-0002-8931-4852>

## REFERENCES

- Chang, Y., J. Latham, M. Licht, and L. Wang. 2023. "A Data-Driven Crop Model for Maize Yield Prediction." *Communications Biology* 6(1): 439.
- Chiranjeevi, S., M. Sadaati, Z. K. Deng, J. Koushik, T. Z. Jubery, D. Mueller, M. E. Neal, et al. 2023. "Deep Learning Powered Real-Time Identification of Insects Using Citizen Science Data." *arXiv preprint arXiv:2306.02507*.
- Cho, M., Z. Ghodsi, B. Reagen, S. Garg, and C. Hegde. 2022. "Sphinx: A Deep Neural Network Design for Private Inference." *IEEE Security & Privacy* 20(5): 22–34.
- Esfandiari, Y., K. Nagasubramanian, F. Fotouhi, P. S. Schnable, B. Ganapathysubramanian, and S. Sarkar. 2021. "Distributed Deep Learning for Persistent Monitoring of Agricultural Fields." In *NeurIPS 2021 AI for Science Workshop*.
- Feuer, B., A. Joshi, M. Cho, K. Jani, S. Chiranjeevi, Z. K. Deng, A. Balu, et al. 2023. "Zero-Shot Insect Detection Via Weak Language Supervision." In *2nd AAAI Workshop on AI for Agriculture and Food Systems*.
- Freeman, H., E. Schneider, C. H. Kim, M. Lee, and G. Kantor. 2023. "3D Reconstruction-Based Seed Counting of Sorghum Panicles for Agricultural Inspection." In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 9594–600. IEEE.
- Ibrahim, H., S. Moru, P. Schnable, and L. Dong. 2022. "Wearable Plant Sensor for in Situ Monitoring of Volatile Organic Compound Emissions from Crops." *ACS Sensors* 7(8): 2293–302.
- Sarkar, S., B. Ganapathysubramanian, A. Singh, F. Fotouhi, S. Kar, K. Nagasubramanian, G. Chowdhary, et al. 2023. "Cyber-Agricultural Systems for Crop Breeding and Sustainable Production." *Trends in Plant Science*.

**How to cite this article:** Ganapathysubramanian, B., J. M. P. Bell, G. Kantor, N. Merchant, S. Sarkar, P. S. Schnable, M. Segovia, A. Singh, and A. K. Singh. 2024. "AIIRA: AI Institute for Resilient Agriculture." *AI Magazine* 45: 94–98. <https://doi.org/10.1002/aaai.12151>

## AUTHOR BIOGRAPHIES

**Baskar Ganapathysubramanian** leads AIIRA and is at Iowa State University. His research interests are in the areas of computational mechanics and scientific computing, with a focus on sustainability applications.

**Jessica M. P. Bell** is a certified project management professional with expertise in research development, team science, and project management. She is AIIRA's Project Manager.

**George Kantor** is in the Robotics Institute at Carnegie Mellon University. His research interests are in the area of field robotics, including navigation, perception, and manipulation in unstructured environments.

**Nirav Merchant** leads the Data Science Institute at the University of Arizona. His expertise is in developing scalable computational platforms for supporting open science and open innovation.

**Soumik Sarkar** is at Iowa State University. His research interests are in data analytics and machine learning algorithms for autonomous perception and decision-making in complex cyber-physical systems.

**Patrick S. Schnable** is at Iowa State University. His expertise is in genetics, molecular biology, genomics and high-throughput phenotyping, with an emphasis on interdisciplinary approaches to plant biology.

**Michelle Segovia** is at the University of Delaware. Her research focuses on behavioral and experimental economics, particularly the economics of food consumption.

**Arti Singh** is a Plant Breeder at Iowa State University. She leads a breeding program to develop new climate-resilient varieties for plant-based protein markets. She deploys AI tools for breeding.

**Asheesh. K. Singh** leads a breeding, discovery, and tool-development group at Iowa State University. His group develops prescriptive cultivars tailored to stakeholders' requirements and future climatic scenarios.



## SPECIAL TOPIC ARTICLE

# AgAID Institute—AI for agricultural labor and decision support

Alan Fern<sup>1</sup> | Margaret Burnett<sup>1</sup> | Joseph Davidson<sup>2</sup> | Janardhan Rao Doppa<sup>3</sup> | Paola Pesantez-Cabrera<sup>3</sup> | Ananth Kalyanaraman<sup>3</sup>

<sup>1</sup>School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon, USA

<sup>2</sup>School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, Oregon, USA

<sup>3</sup>School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington, USA

### Correspondence

Alan Fern, School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA.

Email: [afern@cs.orst.edu](mailto:afern@cs.orst.edu)

### Funding information

National Institute of Food and Agriculture, Grant/Award Number: 2021-67021-35344

### Abstract

The AgAID Institute is a National AI Research Institute focused on developing AI solutions for specialty crop agriculture. Specialty crops include a variety of fruits and vegetables, nut trees, grapes, berries, and different types of horticultural crops. In the United States, the specialty crop industry accounts for a multibillion dollar industry with over 300 crops grown just along the U.S. west coast. Specialty crop agriculture presents several unique challenges: they are labor-intensive, are easily impacted by weather extremities, and are grown mostly on irrigated lands and hence are dependent on water. The AgAID Institute aims to develop AI solutions to address these challenges, particularly in the face of workforce shortages, water scarcity, and extreme weather events. Addressing this host of challenges requires advancing foundational AI research, including spatio-temporal system modeling, robot sensing and control, multiscale site-specific decision support, and designing effective human–AI workflows. This article provides examples of current AgAID efforts and points to open directions to be explored.

## INSTITUTE OVERVIEW

Agriculture is on the cusp of a fourth revolution and digital technology combined with AI is expected to play a pivotal role. Responding to this need, the AgAID Institute has been established as one of the U.S.-based National AI Research Institutes, with the goal of advancing and deploying AI technologies and workflows in agriculture. The AgAID Institute is unique with its focus on *specialty crop agriculture*, which represents a multibillion dollar industry with over 300 crops grown just along the U.S. west coast. This includes a variety of tree fruits, vegetables, nut trees, berries, grapes, and numerous other horticultural crops. Production of specialty crops presents several distinct challenges,

which AI and digital technology are uniquely placed to tackle:

- (a) This industry is *labor-intensive*, accounting for well over 80% of the U.S. agricultural labor workforce. However, farmers face uncertain and variable profitability due to increased labor costs and a shortage of skilled labor.
- (b) *Weather* extremities are increasingly common amidst long-term regional climactic shifts, which poses significant risks for major crop losses, poor yields, and degraded quality.
- (c) Most specialty crops are grown on irrigated lands, and therefore *water* scarcity, droughts, and other climate-induced risks pose significant challenges to watershed

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.

and irrigation district managers who must make water allocation decisions.

A key question at the institute is how to leverage the recent, rapid progress in AI to effectively address these challenges. While many ideas, techniques, and systems from state-of-the-art AI are relevant to specialty crops, they are usually not sufficient. Indeed, the characteristics of specialty crops require advancing important AI frontiers, including:

- Handling *limited, missing, and noisy data* due to expensive data collection activities and imperfect sensors. These data characteristics are very different from typical AI benchmarks, which are often large and sanitized.
- AI systems for agriculture must be *safe* in the sense that they do not lead to worse outcomes for farmers. This involves quantifying model uncertainty for decision support and avoiding crop damage by robotic systems.
- Full AI-based automation will typically not be possible or even desirable. Rather, we must design effective *human–AI workflows*, both for decision support and robot-assisted labor.

Addressing these frontiers can both dramatically improve agricultural outcomes and fundamentally advance general AI techniques and practices.

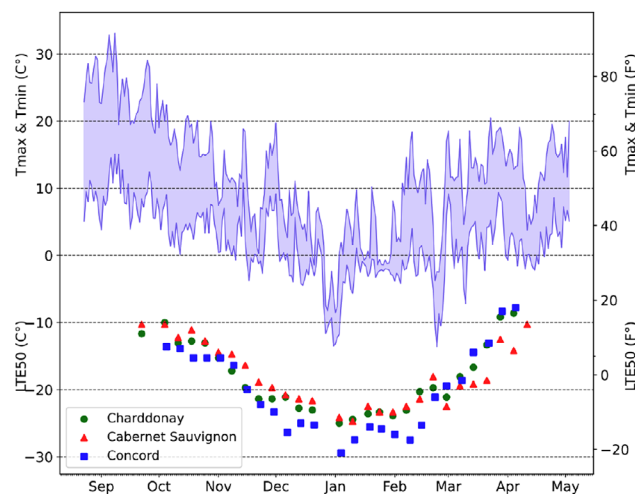
## ONGOING RESEARCH

AgAID research efforts are organized into three major use-inspired thrusts: labor intelligence, farm operations intelligence, and water intelligence. In addition, there is a cross-cutting human–AI thrust, which aims at developing workflows for human–AI partnerships.

In what follows we provide specific examples of projects that are representative of ongoing efforts within these different thrusts. For a more complete list of projects, products, and team information, please refer to the AgAID Institute website <https://agaid.org/>.

### Farm operations thrust: Frost mitigation

Frost damage to crops, such as grapes and cherries, due to cold temperatures in the late fall and early spring can significantly decrease yields. Thus, farmers deploy expensive frost mitigation measures, such as wind machines, sprinklers, and torches, to avoid damage when temperatures are low. However, deciding when temperature forecasts



**FIGURE 1** (Top) Daily max/min temperatures measured at a nearby weather station. (Bottom)  $LTE_{50}$  is a cold hardiness measurement that indicates the temperature at which 50% of the grape buds will freeze. Notice the significant variation across cultivars.

are low enough to warrant expensive measures is challenging due to the unknown cold-hardiness state of the plant. Cold hardiness is a measure of the temperature at which lethal freezing occurs in a plant and changes through biological processes during the dormant season (fall, winter, and spring). As shown in Figure 1, cold hardiness varies throughout the season, with lethal temperatures decreasing due to acclimation in the fall and increasing due to de-acclimation in the spring. Unfortunately, empirically measuring cold hardiness requires expensive, specialized equipment (Mills, Ferguson, and Keller 2006). Thus, farmers must rely on cold-hardiness estimates from models or historical rules-of-thumb to inform their mitigation decisions.

While scientific models of cold hardiness have been developed based on biological principles (Ferguson et al. 2011, 2014), they are currently simple, (near) linear models that only use daily temperature as input. Rather, cold hardiness is also believed to be influenced by other weather factors, such as humidity and precipitation, in nonlinear ways. This suggests the potential of modern deep models for capturing these more complex relationships. In particular, we consider training recurrent neural networks (RNNs) to predict daily cold-hardiness based on the history of weather data up to each day. A key challenge, however, is that the available ground-truth cold-hardiness data is quite limited. For some grape cultivars, as much as 30 seasons of data is available at a biweekly sampling rate and RNN models were able to significantly improve over prior scientific models. However, for many other cultivars only a handful of seasons were available and the prediction accuracy of RNNs was poor compared to current models.

One way to address the small-data challenge is to recognize that while cold-hardiness responses differ across cultivars, there is also a significant shared structure since the cultivars are all from the same plant species. For this purpose, we have designed and evaluated multitask learning (MTL) models that allow for combining data across multiple cultivars (i.e., each cultivar being a task) in order to improve predictions for each individual cultivar (Saxena et al. 2022). The results show that RNNs trained via MTL yield improved accuracy over the prior state-of-the-art scientific models for almost all cultivars, including those with small data. Importantly, our MTL approaches are simple to implement as wrappers around existing single-task models, which supports wider applicability to applications with similar structure.

There are several avenues of related work to explore. First, there may be value in combining data from qualitatively different prediction tasks. For example, historical data have been collected on the dates of certain phenological events for grapes and other plants. Second, there are many other prediction problems in agriculture that may benefit from combining small datasets across related tasks—for example, combining datasets across individual farms, where each farm may only have a small number of seasons of data, or combining data about individual trees or field regions, which have both common and distinct characteristics. Third, there are orthogonal avenues for addressing small-data problems, such as integrating scientific knowledge to constrain or better inform learned models.

### Labor thrust: Intelligent pruning

Dormant season pruning (i.e., after leaf drop) is a labor-intensive orchard activity critical for producing high yields of quality fruit such as cherries, apples, and pears. Pruning (i) removes diseased, vigorous, or otherwise unwanted parts of the tree to maintain/improve the plant's growth and nutrition; (ii) enhances the distribution of sunlight throughout the canopy; and (iii) influences bud spacing and crop load. A major challenge facing tree fruit growers worldwide is the increasing difficulty in finding skilled orchard workers for activities such as pruning. Today, there are no commercially available autonomous pruning systems. Some of the technical challenges that make autonomous pruning difficult include: **Perception:** Small, thin features such as limbs and fruit buds are difficult to image outdoors. The environment is unstructured with background “noise” from adjacent orchard rows, and overlapping branches in images make it difficult to determine which branches belong to which trees. **Decision making:** Pruning is plant and cultivar-specific and depends



**FIGURE 2** Our pruning robot during field trials. The pruning controller was trained entirely in simulation (You, Parayil, et al. 2022).

on many variables (e.g., type and age of cultivar, previous growth and crop-yield behavior, orchard system, the current and potential environmental conditions, and desired crop load and quality). How these variables combine to produce cut decisions for a specific branch is unknown. **Mapping actions to outcomes:** Pruning actions taken during the winter will partially affect tree growth, fruit yield, and fruit quality over the following 9–12 months. Other factors that impact outcomes include fertilization, irrigation, pollination, weather conditions, and so forth. Mapping the direct relationship between physical cuts and harvest yields in the presence of multiple other influencing variables is complex. **Physical pruning:** Trees have unique and variable geometries, there are rigid obstacles present in the environment (e.g., posts and trellis wires), and accurate cuts of small, compliant features require high precision and control. AI tools have tremendous potential to advance the state-of-the-art in autonomous pruning. One area that we are exploring is the use of digital orchard environments to develop robot controllers and perception models. Data collection and labeling in the real orchard environment is resource-intensive; there are also seasonal constraints that limit when data can be collected (e.g., blossoms may only be on the tree for a few days). We have recently used simulated orchard environments and synthetic data to train pruning controllers (You, Kolano, et al. 2022) as well as perception models for foreground/background segmentation (You, Grimm, and Davidson 2022). Recent field trials (Figure 2) show that these models can be successfully transferred to the real world (You, Parayil, et al. 2022). Developing effective approaches for this type of “sim-to-real” transfer is one of AgAID’s fundamental AI contributions with wide applicability.

Another opportunity for AI is assisting with the decision of where to cut. Currently, pruning decisions are



intuitive—workers rely on visual and spatial heuristics learned from repetition and experiential learning. Automating pruning requires making the intuitive and implicit decision-making process explicit and algorithmically representable. To better understand the complexities of the decision-making process, we recently conducted a series of formal studies, including in-person formative interviews and observations, with three groups of pruning stakeholders: horticulturalists, growers, and pruners. An important result from our human studies is a pruning terminology set that can be used to communicate with stakeholders about pruning heuristics and begin to map global pruning contexts to local pruning cuts.

One unexplored area that we are particularly interested in for future study is integrated horticultural–robotic systems. Robots can record every cut that they make. By using the same machines to then record blossom distributions, vigor, crop load, and so forth at the individual plant level, we can better understand the link between pruning strategies and the horticultural outcomes most important to the growers' bottom line: fruit yield and quality. We envision an interactive feedback loop whereby the actions and results recorded by robots can be used by growers and horticulturalists to tune their decisions and, potentially, design trees that are “optimal” for robots.

## Water thrust: Streamflow prediction

Intelligent stewardship and management of water require answering two key questions: (1) How does precipitation translate to spatiotemporal water availability?, and (2) How can we optimize water allocation to make it available when and where it is most needed? We hypothesize that AI-enabled solutions can address them.

Water availability is inherently a complex systems problem, shaped by both natural phenomena and human actions. Streamflow estimates are typically obtained via hydrological models, which take meteorological inputs (temperature, precipitation, radiation, wind speed, humidity, etc.), land surface characteristics such as types of vegetative cover/land use (forests, grassland, agricultural land) and soil characteristics. These scientific models solve differential equations related to the water and energy balance to calculate water and energy fluxes in the stream to quantify streamflow. However, the current scientific models are based on simplifying assumptions and/or incomplete knowledge, which introduces bias leading to a loss in accuracy and generalizability. Also, the predictions from scientific models do not come with uncertainty estimates, which are critical for making water management decisions.

An alternative approach is to leverage the advances in deep temporal models for streamflow prediction. How-

ever, this approach comes with some unique challenges: (a) lack of available ground-truth data, and (b) can produce results that are inconsistent with physical laws. To address these challenges, we are currently exploring principled methods to synergistically combine deep temporal models with domain knowledge in the form of scientific models. This is done by incorporating explicit physical laws into the training of deep temporal models to get physically consistent predictions. Our current experimental results on real data over multiple watersheds demonstrate significant improvements over the scientific model baseline VIC-CropSyst (Malek et al. 2017). To quantify the uncertainty in predicting streamflows, we are exploring two wrapper approaches on top of the deep temporal model using the frameworks of Gaussian processes and conformal prediction.

One unexplored but important challenge is to model and reason about the human influence on water availability to further improve accuracy. The spatiotemporal human–water nexus includes considering multiple hard-to-quantify aspects: (a) infrastructure (e.g., reservoirs) that alter temporal signatures; (b) infrastructure (e.g., irrigation canals) that alter the spatial water flow-path; (c) cropping systems and management that affect spatiotemporal consumptive use; and (d) human behavior, decisions under different water scarcity, and institutional contexts. This problem is particularly challenging because we have limited data about human influence.

## Human–AI partnerships

Enabling agricultural stakeholders to actually benefit from AI tools is central to advancing AI adoption and amplification of outcomes. Since humans are diverse, we do not assume that agricultural populations will all interact with AI tools in uniform ways. Instead, we are going to where these diverse agricultural stakeholders are, not only physically but also in terms of the diverse ways they reason and problem-solve.

Toward this end, in addition to formative and summative empirical work with our human stakeholders, we are using two systematic methods: GenderMag (Burnett et al. 2016) and SocioEconomicMag (Agarwal et al. 2023). GenderMag and SocioEconomicMag are systematic processes for finding and fixing “above-the-hood” (i.e., visible to a human user) biases against problem-solving approaches favored by certain genders and socioeconomic statuses. For example, in an investigation of more than 1000 AI stakeholders, GenderMag revealed that the AI products were skewed away from women's problem-solving styles.

To avert, detect, and ameliorate such biases, several AgAID teams are using GenderMag and/or

SocioEconomicMag to improve their products' abilities to serve diverse agriculture stakeholders. For example, pertinent to both the frost mitigation and water availability works, teams working on predicting weather and other environmental stressors have used GenderMag to find and fix over 30 over-the-hood biases in their design-stage interface planning. Another AgAID team working on intelligent pruning is using SocioEconomicMag to facilitate the relationship between diverse humans and the emerging robotic pruners.

Our work to avoid such over-the-hood biases is important to our goals because the presence of these biases interferes with AgAID's basic principles relating to adoption and amplification. A different approach might be to "fix the users," by training them to use the AI tools "as intended", however, doing so would be insufficient and counterproductive. It would be insufficient because some attributes are deeply entrenched, such as someone's attitudes toward risks or need for control. It would be counterproductive because organizations, businesses, and environments need diversity of thought to thrive (Page 2008). For these reasons, we believe it is more beneficial to our agriculture stakeholders to help diverse problem-solving approaches to flourish than to attempt to eliminate the diversity.

## LOOKING AHEAD

Will AI be a key ingredient of Agriculture 4.0? A "yes" answer entails overcoming the obstacles toward wide-scale adoption of AI by often skeptical stakeholders. The thin margins of agriculture require that AI solutions demonstrate bottom-line utility while directly facing messy real-world realities including seamless transfer of technologies across different cropping systems. The diversity of stakeholders (managers, workers, policymakers) requires careful design of effective AI-human workflows. These are some of the key challenges that AgAID and other researchers are making steady progress on, but we are only at the early stages of the journey. We hope to see many others join-in pursuit of a more secure future for people who need to be fed and the people who feed them.

## ACKNOWLEDGMENTS

The AgAID Institute is a multi-institutional consortium comprised of Washington State University (lead), Oregon State University, University of California Merced, University of Virginia, Carnegie Mellon University, Heritage University, Wenatchee Valley College, Kansas State University, and innov8.ag. AgAID is supported by USDA-NIFA by the AI Research Institutes program, under award No. 2021-67021-35344.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Alan Fern  <https://orcid.org/0000-0001-5851-8935>

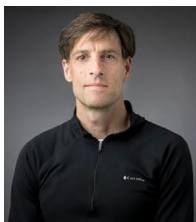
## REFERENCES

- Agarwal, P., D. Prem, C. Bogart, A. Fallatah, A. A. Castro-Guzman, P. Chanpaisaeng, S. Doehring, M. Burnett, and A. Sarma. 2023. "SocioEconomicMag Meets a Platform for SES-Diverse College Students: A Case Study." arXiv preprint arXiv:2304.04873.
- Burnett, M., S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan. 2016. "GenderMag: A Method for Evaluating Software's Gender Inclusiveness." *Interacting with Computers* 28(6): 760–87.
- Ferguson, J. C., M. M. Moyer, L. J. Mills, G. Hoogenboom, and M. Keller. 2014. "Modeling Dormant Bud Cold Hardiness and Budbreak in Twenty-Three Vitis Genotypes Reveals Variation by Region of Origin." *American Journal of Enology and Viticulture* 65(1): 59–71.
- Ferguson, J. C., J. M. Tarara, L. J. Mills, G. G. Grove, and M. Keller. 2011. "Dynamic Thermal Time Model of Cold Hardiness for Dormant Grapevine Buds." *Annals of Botany* 107(3): 389–96.
- Malek, K., C. Stöckle, K. Chinnayakanahalli, R. Nelson, M. Liu, K. Rajagopalan, M. Barik, and J. C. Adam. 2017. "VIC-CropSyst-v2: A Regional-Scale Modeling Platform to Simulate the Nexus of Climate, Hydrology, Cropping Systems, and Human Decisions." *Geoscientific Model Development* 10(8): 3059–84.
- Mills, L. J., J. C. Ferguson, and M. Keller. 2006. "Cold-Hardiness Evaluation of Grapevine Buds and Cane Tissues." *American Journal of Enology and Viticulture* 57(2): 194–200.
- Page, S. 2008. "The difference." In *The Difference*, Princeton University Press, Princeton, New Jersey.
- Saxena, A., P. Pesantez-Cabrera, R. Ballapragada, K.-H. Lam, A. Fern, and M. Keller. 2022. "Grape Cold Hardiness Prediction via Multi-Task Learning." In *Innovative Applications of Artificial Intelligence*.
- You, A., C. Grimm, and J. R. Davidson. 2022. "Optical Flow-Based Branch Segmentation for Complex Orchard Environments." In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9180–86.
- You, A., H. Kolano, N. Parayil, C. Grimm, and J. R. Davidson. 2022. "Precision Fruit Tree Pruning Using a Learned Hybrid Vision/Interaction Controller." In *2022 International Conference on Robotics and Automation (ICRA)* 2280–86.
- You, A., N. Parayil, J. G. Krishna, U. Bhattarai, R. Sapkota, D. Ahmed, M. Whiting, M. Karkee, C. M. Grimm, and J. R. Davidson. 2022. "An Autonomous Robot for Pruning Modern, Planar Fruit Trees." arXiv:2206.07201.

**How to cite this article:** Fern, A., M. Burnett, J. Davidson, J. R. Doppa, P. Pesantez-Cabrera, and A. Kalyanaraman. 2024. "AgAID Institute—AI for agricultural labor and decision support." *AI Magazine* 45: 99–104.  
<https://doi.org/10.1002/aaai.12156>



## AUTHOR BIOGRAPHIES



**Alan Fern** is a Professor of Computer Science at Oregon State University and is the AgAID AI-research lead. His research focuses on AI planning and learning for decisions making.



**Margaret Burnett** is a Distinguished Professor at Oregon State University and ACM fellow. Her research is on humans' work to solve problems with the aid or hindrance of computers.



**Joseph Davidson** is an Assistant Professor of Robotics at Oregon State University. He directs the Intelligent Machines and Materials Lab with research on manipulation, sensing, and control.



**Janardhan Rao Doppa** is the Huie-Rogers Associate Professor at Washington State University (WSU). His research includes both AI foundations and applications in science and engineering.



**Paola Pesantez-Cabrera** is a Research Assistant Professor at Washington State University. Her research focuses on data-driven decision support in agriculture by applying FAIR data principles.



**Ananth Kalyanaraman** is a Professor of Computer Science at WSU and is Director of the AgAID AI Institute. His research interests include data science for life science applications including agriculture.



## SPECIAL TOPIC ARTICLE

# AI2ES: The NSF AI Institute for Research on Trustworthy AI for Weather, Climate, and Coastal Oceanography

Amy McGovern<sup>1,2</sup> | Imme Ebert-Uphoff<sup>3</sup> | Elizabeth A. Barnes<sup>4</sup> | Ann Bostrom<sup>5</sup> | Mariana G. Cains<sup>6</sup> | Phillip Davis<sup>7</sup> | Julie L. Demuth<sup>6</sup> | Dimitrios I. Diochnos<sup>2</sup> | Andrew H. Fagg<sup>2</sup> | Philippe Tissot<sup>8</sup> | John K. Williams<sup>9</sup> | Christopher D. Wirz<sup>6</sup>

<sup>1</sup>School of Meteorology, University of Oklahoma, Norman, Oklahoma, USA

<sup>2</sup>School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

<sup>3</sup>Electrical and Computer Engineering & Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA

<sup>4</sup>Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA

<sup>5</sup>Evans School of Public Policy and Governance, University of Washington, Seattle, Washington, USA

<sup>6</sup>National Center for Atmospheric Research, Boulder, Colorado, USA

<sup>7</sup>Del Mar College, Corpus Christi, Texas, USA

<sup>8</sup>Conrad Blucher Institute, Texas A&M University—Corpus Christi, Corpus Christi, Texas, USA

<sup>9</sup>The Weather Company, IBM Business, Armonk, New York, USA

## Correspondence

Amy McGovern, School of Meteorology, University of Oklahoma, Norman, OK, USA.

Email: [amcgovern@ou.edu](mailto:amcgovern@ou.edu)

## Funding information

Directorate for Geosciences, Grant/Award Number: ICER-2019758

## Abstract

The NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) focuses on creating trustworthy AI for a variety of environmental and Earth science phenomena. AI2ES includes leading experts from AI, atmospheric and ocean science, risk communication, and education, who work synergistically to develop and test trustworthy AI methods that transform our understanding and prediction of the environment. Trust is a social phenomenon, and our integration of risk communication research across AI2ES activities provides an empirical foundation for developing user-informed, trustworthy AI. AI2ES also features activities to broaden participation and for workforce development that are fully integrated with AI2ES research on trustworthy AI, environmental science, and risk communication.

## INTRODUCTION TO AI2ES

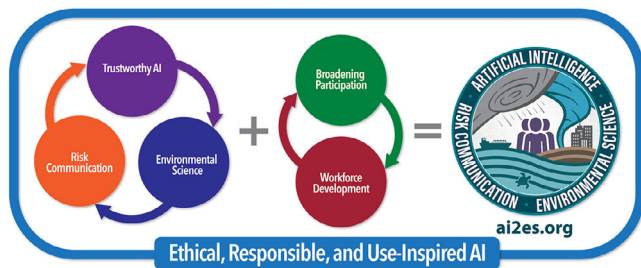
*The NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)* is a convergent center focused on AI for the Earth and environmental sciences (ES) (McGovern et al. 2022). We are developing novel AI methods for real-world high-impact

environmental use cases that ensure that we address the entire chain of relevant issues.

AI2ES has **two primary goals**. First, we are advancing the state-of-the-art of foundational research in AI, ES, and RC. Second, we are advancing understanding and prediction of weather phenomena to improve societal resilience to climate change and save lives and property. To achieve

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Underlying research structure of AI2ES, highlighting the synergistic cycle of AI, ES, and RC research. Our commitment to ethical, responsible, and use-inspired research underlies all that we do.

these goals, we address **three key topics**: (1) Ensuring that the developed AI methods are trustworthy<sup>1</sup> and meet the needs of diverse groups of end-users; (2) developing novel AI that revolutionizes our understanding and prediction of high-impact environmental phenomena; and (3) creating new educational pathways to develop a larger and more diverse AI/ES workforce.

Creating trustworthy AI for ES requires AI2ES to address **several key challenges**. First, many AI for ES applications directly impact lives and property. Furthermore, the AI methods are to be used in time-critical situations. The phenomena of interest are often rare, leading to a sparsity of labeled training data; the available data are multiscale and heterogeneous. ES applications provide an additional challenge that does not occur in many other AI applications: the need for AI models to be physically grounded while capturing complex nonlinear relationships. Finally, we must identify what is needed for the AI to be deemed trustworthy by a diverse set of end-user groups and create the necessary methods to achieve this.

The institute's work is structured as outlined in Figure 1. Shown on the left is our key synergistic cycle that connects foundational research in three areas: AI, ES, and RC. Likewise, our AI workforce development and broadening participation efforts (shown in the middle) are also synergistic and build on our foundational research. Underlying all that we do is a focus on ensuring that our AI is ethical, responsible, and use-inspired.

## AI2ES RESEARCH

Given the overview nature of this paper, we highlight key research in each of the three areas of the synergistic cycle from Figure 1.

**Generative AI:** Generative AI, in particular attention-based models, such as transformers, are enabling the development of powerful AI-driven global weather forecasting

models. These AI models execute up to 1000× faster than the numerical weather prediction (NWP) models currently used operationally for weather forecasting, and are starting to reach the resolution and accuracy of NWP models. However, pixel-based accuracy is not a sufficient measure for the usefulness of these models for real-world weather forecasting. Do these models properly predict the key features of storm fronts and tropical cyclones? Do they capture the extremes of heat waves and precipitation? Are the predicted fields consistent? Answering these questions requires expertise in both generative AI, but and meteorology, a perfect task for AI2ES. We are working with private industry and NOAA to evaluate such models and provide feedback to developers (Ebert-Uphoff and Hilburn 2023).

**Robust AI:** We are developing a set of robust methods for learning models from imperfect data. One aspiration is to address the *class imbalance* and *rare event* problems, including characterizing the sample sizes needed for achieving certain guarantees in performance. In Diochnos and Trafalis (2021), we show that any learning algorithm that generates a *probably approximately correct (PAC)* model can be extended to learn a model that also has high recall and high precision, while maintaining the efficiency of the original algorithm.

We are also working to ensure that our AI methods are *robust to noise and missing data*. In Flansburg and Diochnos (2022), we show that  $L_1$ -regularization can be an effective defense mechanism for regression models that are subject to certain training-time attacks, complementing a property of  $L_1$ -regularization known for classification models under test-time attacks. We are currently exploring *semi-supervised* learning to build-in robustness to labels that are noisy or that are missing altogether.

**Explainable and interpretable AI (XAI/IAI):** Explaining predictions of AI models is one aspect of trustworthiness for end-users. The need to peer inside the black box of AI and what the end-users need to see of what the AI model is predicting is context dependent (Wirz et al.). This includes visualizing input–output relationships on data and other aspects of explanation such as case studies and failure modes (Cains et al. 2023).

Most XAI methods were not developed for ES domains and understanding their applicability to the highly spatially and temporally auto-correlated data prevalent in ES domains needs to be investigated. Furthermore, most XAI methods related to image-based tasks yield so-called *attention maps* that indicate where in an input image, a neural network paid the most attention. However, different XAI methods yield greatly differing attention maps. Lastly, the resulting attention maps are usually interpreted visually

by AI developers for clues on the strategies the neural network is using, introducing a large potential for subjectivity in interpretation. To address all of these aspects, we developed two synthetic benchmarks (Mamalakis, Barnes, and Ebert-Uphoff 2022; Mamalakis, Ebert-Uphoff, and Barnes 2022) representative of Earth science processes that provide ground truth not only for the neural network prediction, but also for the corresponding attention maps. Using these benchmarks, we identified key characteristics of XAI methods for attention maps. This allowed us to identify which XAI algorithms are most suitable to address certain science questions and how their differing results should be evaluated (Mamalakis, Barnes, and Ebert-Uphoff 2022; Mamalakis, Ebert-Uphoff, and Barnes 2022).

Furthermore, while *post hoc* XAI methods focus on the regions of the input that were most relevant to the network's prediction, they do not tell *how* the prediction was made (Rudin 2019). Our team is building a suite of interpretable neural network architectures that mimic the way scientists interpret weather and climate patterns, allowing the full decision making process to be tracked from start to finish. One example of our IAI networks uses prototypical samples (i.e., “this looks like that”) from the training input to classify new samples at inference. However, for weather and climate prediction, the global locations of atmospheric and oceanic phenomena are vital in understanding and predicting their downstream impacts on weather and climate extremes. Thus, our modified IAI network encodes the spatial location of the prototypes as well, that is, “this looks like that there” (Barnes et al. 2022). In this context, we are also currently developing an explanation method for image classification, where we do not allow overlapping regions of an image to be plausible explanations for different classes.

**Uncertainty quantification (UQ):** Given that many environmental applications involve life-and-death decision making, integrating uncertainty quantification into AI models that support the decision making process is crucial. Our focus on UQ includes understanding the limitations of existing methods for UQ for ES, developing novel methods to represent uncertainty, and approaches for communicating and visualizing uncertainty for and with end-users. For example, in Haynes et al. (2023), we review six different UQ approaches for neural networks—from simple approaches to Bayesian neural networks—and apply them to two environmental case studies. We also highlight four different ways to evaluate such uncertainty estimates, and use the case studies to illustrate how to use them to identify which UQ methods yield reliable estimates.

**Responsible and ethical AI for ES:** The use of AI is growing exponentially across society, as well as within the

sciences. With this use, comes an increased understanding of the need for ethical and responsible development and deployment of AI for many use-cases. However, many AI developers for ES do not see how the issues in the news with AI affected their work on AI for ES applications. We published an in-depth series of examples demonstrating how AI can go wrong when naively applied to ES problems (McGovern et al. 2022).

In our current work, we are focusing on the issue of bias in AI models, specifically examining it for ES applications (McGovern et al. 2023). While bias is not the only issue that needs to be addressed for ethical and responsible use of AI, it is a key issue. We have developed a classification of bias modeled after the more general one discussed in NIST's AI risk management work. In our current work, we are diving deeply into our four main bias categories and showing how AI developers can address and manage these risks for ES applications.

**Use inspired research in ES: Grounding our research:** All of our work is use-inspired by phenomena in atmospheric sciences and coastal oceanography. Our use cases include: (1) convective hazards, meaning those that are associated with strong thunderstorms, including wind, hail, lightning, and tornadoes; (2) winter weather, including visibility, snowfall, and freezing rain; (3) coastal phenomena, including fog, forecasting cold stunning events to save sea turtles, and understanding harmful algal blooms; (4) tropical cyclones; (5) subseasonal to seasonal prediction of diverse high-impact phenomena including excessive rainfall events.

**Risk communication:** A crucial but often neglected topic is how fundamental advances in AI for predictions of ES hazards can be developed to provide information that is needed, trusted, and used by professionals, scientists, and expert decision makers, such as weather forecasters, transportation officials, emergency managers, and natural resource managers. Treating AI models and XAI as forms of risk information, the RC team leads convergent, interdisciplinary, multimethod research across AI2ES on trustworthy AI. Initial studies have included ES use cases such as severe hail, storm mode, and coastal fog, with a focus on forecasters' assessments of potential use and trustworthiness of AI-based guidance, and how factors such as model verification and the ability to interact with the model output influence those assessments. Drawing on research from fields as diverse as risk communication and management, organizational and social psychology, human-AI teaming, and trust in automation, the team is contributing methodological insights, as well as fundamental insights into the contextual and subjective nature of trusting and assessments of trustworthiness.



## EDUCATION, WORKFORCE DEVELOPMENT, AND OUTREACH

Our efforts to broaden participation and workforce development are strongly connected. We overview key parts here.

**Developing an AI certificate program at Del Mar community college:** Our core components in the area of education include close collaboration with a Hispanic Serving Institution (HSI) community college. Del Mar College (DMC) is a 2-year community and technical college located in Corpus Christi Texas and leads the AI2ES effort to bring AI and machine learning technologies to the local workforce. The research team of four educators at DMC has led the design and creation of a new Occupation Skills Award degree program, consisting of five courses. Three of the five courses are new for the Award and deal with AI in general and machine learning algorithms for GIS technology specifically. The courses have all been taught to three cohorts of learners, leading to two cohorts of graduates, most of whom transferred to our partner Texas A&M University—Corpus Christi (TAMU-CC) to continue their research and education in the ES.

**Core diversity efforts:** Both DMC and TAMU-CC are HSIs, with major minority student populations. DMC has made significant efforts to recruit both women and minorities into their AI program. The college has participated in numerous public recruitment efforts revolving around public events including annual GIS Day, Earth/Bay Day, and Hurricane Conference events. To create a sustainable pipeline of secondary students from high school to the college, annual summer STEM camps have focused on local middle and high school students, with an emphasis on minority girls. Research has noted the need for outreach and recruitment in the middle school years for girls, since they form their life goals and plans much earlier than boys of the same age. The college has completed two successful summer bootcamps, with 2023 adding returning campers to expand upon their technology exposure and reinforce their bonds with DMC.

The DMC activities, some in collaboration with TAMU-CC students and faculty, are the foundation of an AI2ES student pipeline. Students enrolled in the new DMC GeoAI classes and other computer science classes are introduced, or reintroduced, to AI2ES opportunities. The close collaboration between DMC and TAMU-CC faculty facilitates this bridge. While the two organizations are in the same city, the type of students enrolling at each institution is quite different due to financial constraints and cultural differences, including not being familiar with the opportunities, career paths, and financial support possible through higher education. Pairing a community college and a

university with overlapping programs has been a very productive practice to recruit a broader range of students and launch them on an AI career for AI2ES. Community College students are more likely to be first generation, underrepresented minority, and not be aware of STEM opportunities or their own potential. At present, seven DMC students have been hired as undergraduate research assistants at TAMU-CC while still being enrolled at DMC. The context of a large AI institute has been invaluable for these students. The biweekly site-wide meetings, and particularly participating in the AI2ES first live meeting and presenting or co-presenting at the American Meteorological Society conferences (15 student presentations in 2023) have opened the eyes of many of these students to broader possibilities and will help diversify our field. They now dream bigger.

**Internships and collaboration with industry:** AI2ES industry partners comprise an Industry Advisory Board that offers their perspective on how the institute's research can help address important problems in the private sector. Industry collaborators are involved in many aspects of the Institute's research areas, participate as mentors for student research projects, and offer summer internships. These activities help prepare students for the workforce while also catalyzing the transition of AI2ES research into operations, thereby broadening the Institute's impact and enhancing its service to society.

## EXAMPLE SUCCESS STORIES: TEXT BOXES

**Uncertainty quantification for tropical cyclone intensity and track forecasts:** Uncertainty quantification and communication can be incredibly challenging. We have explored a simple method for adding uncertainty to almost any neural network regression task via estimation of a general probability distribution (Barnes, Barnes, and DeMaria 2023). We showed that this intuitive approach can improve current tropical cyclone forecasts of intensity, as well as their track, by adding uncertainty estimates to an otherwise deterministic prediction. This approach and product is currently being tested at the National Hurricane Center.

**Advancing the conceptualization of trustworthiness:** Drawing on trust-related literature across multiple disciplines and fields, we have synthesized knowledge on interpersonal trust, trust and risk perceptions, and trust in automation (Bostrom et al. 2023; Wirz et al.). This synthesis of trust theory, along with our ongoing empirical research, informs our (re)conceptualization of trustworthiness as being in practice a user's subjective assessment

for a specific situation, which may be affected by time pressure and decision stakes, even when an AI/ML model has been developed in accordance with trustworthiness standards. The resulting potential misalignment between developers and users is analogous to mismatches found historically between lay and expert risk assessments of other technologies, such as nuclear power. The AI2ES risk communication team is co-producing these syntheses and studies that build on them with AI/ML developers and environmental scientists to advance the evaluation and treatment of AI/ML trustworthiness in the ES.

## CONCLUSIONS

AI2ES is leading the development of AI models for weather and climate applications. The foundational methods developed by AI2ES will revolutionize our ability to predict, understand, and communicate a variety of high-impact weather hazards.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Amy McGovern  <https://orcid.org/0000-0001-6675-7119>

## ENDNOTE

<sup>1</sup> AI2ES Definition of Trustworthiness: Trustworthiness is a trustee's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.

## REFERENCES

- Barnes, E. A., R. J. Barnes, and M. DeMaria. 2023. "Sinh-Arcsinh-Normal Distributions to Add Uncertainty to Neural Network Regression Tasks: Applications to Tropical Cyclone Intensity Forecasts." *EDS* 2: e15.
- Barnes, E. A., R. J. Barnes, Z. K. Martin, and J. K. Rader. 2022. "This Looks Like That There: Interpretable Neural Networks for Image Tasks When Location Matters." *AIES* 1(3): e220001.
- Bostrom, A., J. L. Demuth, C. D. Wirz, M. G. Cains, A. Schumacher, D. Madlambayan, A. S. Bansal, et al. 2023. "Trust and Trustworthy Artificial Intelligence: A Research Agenda for AI in the Environmental Sciences." *Risk Analysis*: 1–16. <https://doi.org/10.1111/risa.14245>
- Cains, M. G., C. D. Wirz, J. L. Demuth, A. Bostrom, A. McGovern, I. Ebert-Uphoff, D. J. Gagne, A. Burke, and R. Sobash. 2023. "Exploring what AI/ML Guidance Features NWS Forecasters Deem Trustworthy." In *103rd AMS Annual Meeting*. AMS.

- Diochnos, D. I., and T. B. Trafalis. 2021. "Learning Reliable Rules under Class Imbalance." In *SDM*, 28–36.
- Ebert-Uphoff, I., and K. Hilburn. 2023. "The Outlook for AI Weather Prediction." *Nature* 619: 473–74.
- Flansburg, C., and D. I. Diochnos. 2022. "Wind Prediction under Random Data Corruption (Student Abstract)." In *AAAI*, 12945–46.
- Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff. 2023. "Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications." *AIES* 2: 1–58.
- Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff. 2022. "Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience." *AIES* 1(4): e220012.
- Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes. 2022. "Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset." *EDS* 1: e8.
- McGovern, A., A. Bostrom, M. McGraw, R. J. Chase, D. J. Gagne, I. Ebert-Uphoff, K. D. Musgrave, and A. Schumacher. 2024. "Identifying and Categorizing Bias in AI/ML for Earth Sciences." *BAMS*. <https://doi.org/10.1175/BAMS-D-23-0196.1>, in press.
- McGovern, A., A. Bostrom, P. Davis, J. L. Demuth, I. Ebert-Uphoff, R. He, J. Hickey, et al. 2022. "NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)." *BAMS* 103(7): E1658–68.
- McGovern, A., I. Ebert-Uphoff, D. J. Gagne, and A. Bostrom. 2022. "Why We Need to Focus on Developing Ethical, Responsible, and Trustworthy Artificial Intelligence Approaches for Environmental Science." *EDS* 1: e6.
- Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5): 206–15.
- Wirz, C. D., J. L. Demuth, A. Bostrom, M. G. Cains, I. Ebert-Uphoff, D. J. Gagne, A. Schumacher, A. McGovern, and D. Madlambayan. "(Re)Conceptualizing Trustworthy AI as Perceptual and Context-Dependent: A Foundation for Change."

**How to cite this article:** McGovern, A., I. Ebert-Uphoff, E. A. Barnes, A. Bostrom, M. G. Cains, P. Davis, J. L. Demuth, D. I. Diochnos, A. H. Fagg, P. Tissot, J. K. Williams, and C. D. Wirz. 2024. "AI2ES: The NSF AI Institute for Research on Trustworthy AI for Weather, Climate, and Coastal Oceanography." *AI Magazine* 45: 105–10. <https://doi.org/10.1002/aaai.12160>

## AUTHOR BIOGRAPHIES

**Amy McGovern** is a Lloyd G. and Joyce Austin Presidential Professor in both the School of Meteorology and School of Computer Science at the University of Oklahoma.



**Imme Ebert-Uphoff** is a Research Professor in Electrical and Computer Engineering and the Machine Learning lead at the Cooperative Institute for Research in the Atmosphere, both at Colorado State University.

**Elizabeth A. Barnes** is a Professor in the Department of Atmospheric Science at Colorado State University.

**Ann Bostrom** is the Weyerhaeuser endowed Professor in environmental policy in the Evans School of Public Policy & Governance at the University of Washington.

**Mariana G. Cains** is a Research Scientist at the National Center for Atmospheric Research.

**Phillip Davis** is a Professor at Del Mar College.

**Julie L. Demuth** is a Research Scientist at the National Center for Atmospheric Research.

**Dimitrios I. Diochnos** is an Assistant Professor in the School of Computer Science at the University of Oklahoma.

**Andrew H. Fagg** is a Brian E. and Sandra O'Brien Presidential Professor and an Associate Professor in the School of Computer Science at the University of Oklahoma.

**Philippe Tissot** is the Conrad Blucher Institute Chair for Coastal Artificial Intelligence at Texas A&M University-Corpus Christi.

**John K. Williams** is a Senior Technical Staff Member and Head of Weather AI Sciences at The Weather Company, an IBM Business.

**Christopher D. Wirz** is a Research Scientist at the National Center for Atmospheric Research.



## SPECIAL TOPIC ARTICLE

# Institute for Artificial Intelligence and Fundamental Interactions (IAIFI): Infusing physics intelligence into artificial intelligence

Jesse Thaler | Mike Williams | Marisa LaFleur

Massachusetts Institute of Technology,  
Cambridge, Massachusetts, USA

**Correspondence**

Jesse Thaler, Massachusetts Institute of  
Technology, Cambridge, MA, USA.  
Email: [jthaler@mit.edu](mailto:jthaler@mit.edu)

**Funding information**

National Science Foundation,  
Grant/Award Number: PHY-2019786

**Abstract**

The NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI, pronounced /aI-faI/) is one of the inaugural NSF AI research institutes (<https://iaifi.org>). The IAIFI is enabling physics discoveries and advancing foundational AI through the development of novel AI approaches that incorporate first principles from fundamental physics. By combining state-of-the-art research with early career talent and a growing AI + physics community in the Boston area and beyond, the IAIFI is enabling researchers to develop AI technologies to tackle some of the most challenging problems in physics, and transfer these technologies to the broader AI community. Since trustworthy AI is as important for physics discovery as it is for other applications of AI in society, IAIFI researchers are applying physics principles to develop more robust AI tools and to illuminate existing AI technologies. To cultivate human intelligence, the IAIFI promotes training, education, and public engagement at the intersection of physics and AI. In these ways, the IAIFI is fusing deep learning with deep thinking to gain a deeper understanding of our universe and AI.

**INTRODUCTION**

Artificial intelligence (AI) is transforming many aspects of society, including the ways that scientists are pursuing groundbreaking discoveries. For many years, physicists have been at the forefront of applying AI methods to investigate fundamental questions about the universe. Building on these successes, researchers at MIT, Harvard, Northeastern, and Tufts founded the Institute for Artificial Intelligence and Fundamental Interactions (IAIFI, pronounced /aI-faI/)<sup>1</sup> as one of the inaugural National Science Foundation AI research institutes, with a focus on the interdisciplinary field of AI + physics. IAIFI is developing and

deploying the next generation of AI technologies to tackle some of the most challenging problems in physics, from precision calculations of the structure of matter to gravitational wave detection of merging black holes. Simultaneously, IAIFI researchers are leveraging first principles from physics to drive AI innovation, based on the transformative idea that artificial intelligence can directly incorporate physics intelligence, from developing a better understanding of deep learning theory to improving robot locomotion.

The IAIFI is working to establish the Boston area—where there is already a critical mass of researchers leading the way in AI + physics, many of whom are represented in the leadership of IAIFI—as a hub for state-of-the-art

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.

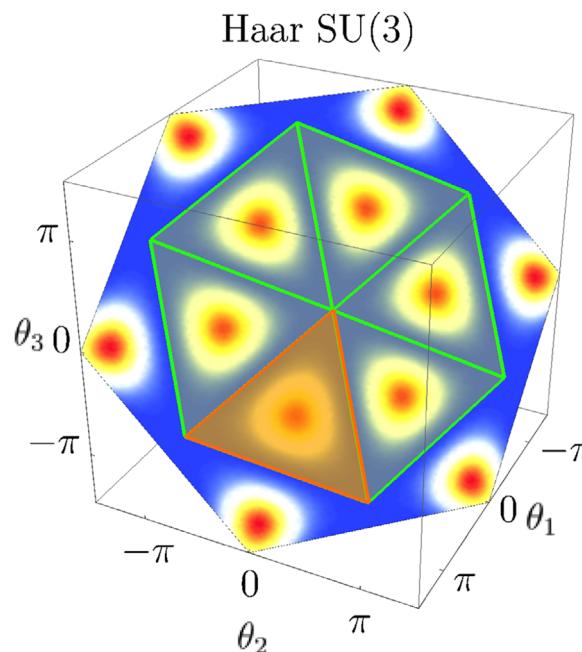


research and interdisciplinary collaboration in this field, with far-reaching impacts both geographically and scientifically. By localizing the community of researchers and enabling regular in-person interactions, IAIFI is able to conduct research in a broad range of scientific domains, and can then serve as a nexus point for research at this intersection more globally as IAIFI researchers collaborate across their domains. By leveraging this community of researchers, the IAIFI is also prioritizing education, engagement, and ethics in AI + physics, contributing to the future of the field. Together, IAIFI researchers are fusing the deep learning revolution with the time-tested strategies of deep thinking in physics to gain a deeper understanding of our universe—from the smallest building blocks of nature to the largest structures in the universe—and of the principles underlying intelligence itself.

## STATE-OF-THE-ART RESULTS

IAIFI research teams led by faculty and postdoctoral researchers hired through the IAIFI Fellows Program<sup>2</sup> are producing state-of-the-art results with applications across physics and AI. For many physics problems, the governing equations that encode the fundamental physical laws are known. However, undertaking key calculations within these frameworks—essential to test our understanding of the universe and guide physics discovery—can be computationally demanding or even intractable. IAIFI researchers are developing AI for such first-principles theory studies, which require AI approaches that rigorously encode physics knowledge. One such application is focused on developing AI methods for performing calculations involving the strong nuclear force that binds quarks into protons and neutrons, and protons and neutrons into nuclei. IAIFI researchers have been leading different efforts to develop novel machine learning algorithms to address this challenge.

For example, a research team—led by Phiala Shanahan (MIT) and her group, including IAIFI Fellow Denis Boyda, in collaboration with colleagues at DeepMind—has developed machine learning architectures that incorporate the physical symmetries of the theory of quantum chromodynamics, incorporating guarantees of exactness into the novel AI algorithms (Boyda et al. 2021) (see Figure 1), thus ensuring that the fundamental laws of physics are preserved in the calculations while increasing efficiency and speed. If these results can be successfully scaled up to current state-of-the-art applications, they will enable novel first-principles studies of nuclear and particle physics. Beyond the domain of physics, this work can be used in robotics or for artificial limbs, where exact rotational symmetries inherently arise in joints.



**FIGURE 1** From <https://arxiv.org/abs/2008.05456>: Visualization of the complex symmetry space of variables in first-principles theoretical physics calculations, which has been built into machine learning models.

Incorporating physics principles into AI is also having a major impact on many experimental applications, such as designing AI methods that are more robust and more easily verifiable. IAIFI researchers are working to enhance the scientific potential of various facilities, including at the Large Hadron Collider (LHC), working on the CMS and LHCb experiments, and at the Laser Interferometer Gravity Wave Observatory (LIGO).

The IAIFI LHCb group, led by IAIFI Deputy Director Mike Williams (MIT), has developed a novel type of neural network that guarantees the interpretability and robustness required for use in real-time data processing at the LHC—applications with some of the largest data rates in the world. This ensures that researchers working on the LHC will be able to understand how a neural network makes its data processing decisions—for example, how it determines whether to flag a piece of data for further investigation or discard it. Since verification and interpretability of AI solutions are also important in other AI application domains, it is not surprising that these novel neural networks developed for the LHC have also been shown to beat state-of-the-art models in various problems in other domains, including criminal justice, medicine, and finance (Kitouni, Nolte, and Williams 2022). A related effort led by the IAIFI CMS group, led by Phil Harris (MIT), is in the area of ultra-low-latency AI inference (optimized to process large amounts of data with minimal delay), where neural networks make

decisions in less than a microsecond, motivated by the relentless pressure of the LHC's 40 MHz proton-bunch collision rate. IAIFI technology here is also being applied to other domains where real-time decision-making is critical<sup>3</sup>.

At LIGO, Lisa Barsotti (MIT) is collaborating with IAIFI Fellow Ge Yang and Pulkit Agrawal (MIT) to explore ways to introduce AI to control the experiment. They are starting by studying how to use reinforcement learning to optimize LIGO performance by controlling LIGO's squeezed vacuum system, which is a complex opto-mechanical system involving multiple feedback loops that needs to be optimized depending on several changing parameters (Whittle et al. 2023).

These and other results of IAIFI research have groundbreaking implications for not only fundamental physics, but also for AI innovation, as they help us develop AI tools that will be used across disciplines and for a variety of applications.

## A HUB FOR INTERDISCIPLINARY COLLABORATION

To facilitate research, talent, and community in this interdisciplinary field, the IAIFI is recruiting and training a talented and diverse group of early-career researchers, especially at the postdoctoral level through the IAIFI Fellows Program. By offering the Fellows their choice of research problems, and the chance to focus on exciting challenges in AI + physics, the IAIFI is preparing many talented young scientists to become future leaders in the field. The Boston area offers a network of experts in both academia and industry (e.g., Google, Microsoft Research, and many startups), which the IAIFI is leveraging to provide opportunities and advice to early career researchers.

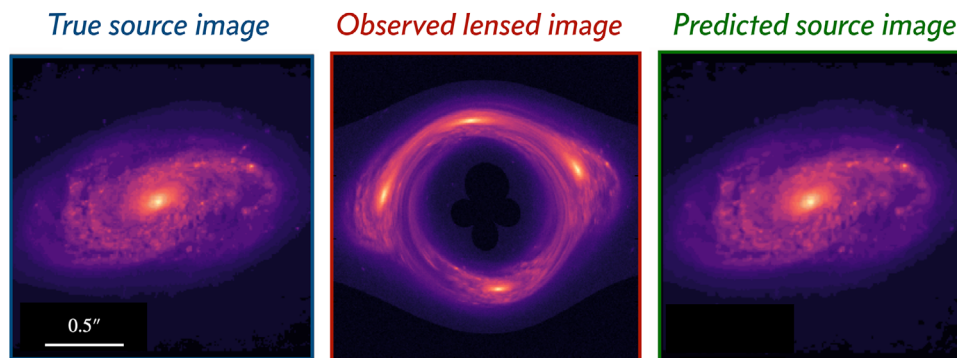
The IAIFI Fellows are sparking interdisciplinary and multi-investigator collaborations, generating new ideas and approaches, translating physics challenges beyond their native domains, and helping develop a common language across disciplines. IAIFI Fellows Siddharth Mishra-Sharma and Ge Yang, whose primary research focuses are astrophysics and robotics, respectively, collaborated on a project that pushed the boundaries of both AI and physics by making it possible to efficiently model high-resolution strong gravitational lensing observations at their full complexity (Mishra-Sharma and Yang 2022) (see Figure 2)—in other words, using AI to correct galaxy images that have been distorted by gravitational fields. In doing so, they utilized a probabilistic treatment of continuous neural fields, which could be more widely applicable in the field

of computer vision, where this AI method is typically used.

Another IAIFI Fellow, Di Luo, has teamed up with several IAIFI faculty members, including with Jim Halverson (Northeastern) on a project studying quantum states using an AI technique known as the neural tangent kernel (Luo and Halverson 2023). Simulating quantum many-body systems on conventional computers is computationally daunting, since the dimensionality of a quantum state grows exponentially with the system size. Inspired by progress in dealing with high dimensionality in machine learning models, one potential way to overcome this computational challenge is by using a neural network quantum state (NNQS). The NNQS theory developed by Luo and Halverson considers the physical properties of quantum states, including the important role of quantum entanglement, to understand the machine learning dynamics. This theory becomes exact in the infinite-width neural-network limit, enabling analytic solutions to the task of quantum state supervised learning.

Two IAIFI Fellows who joined in 2022 are actively working across institutions on developing AI techniques to advance research in their respective fields. Jessie Micallef, working with Taritree Wongjirad (Tufts), is developing generative networks for analyzing data from neutrino experiments. Meanwhile, Carolina Cuesta-Lazaro is collaborating with researchers at Harvard and MIT, as well as IAIFI Fellow Siddharth Mishra-Sharma, to apply generative models to the large scale structure of the universe. This allows them to model galaxies in their natural physical representation—as a set of galaxies—in contrast to previous approaches, which binned galaxies in order to analyze them as a pixelized image or box. In both cases, researchers are working with Tess Smidt (MIT) to develop solutions using symmetry-equivariant neural networks, which incorporates symmetry as an inductive bias in the development of neural network architectures. This makes the techniques more efficient, as less training data or fewer simulations are required.

The IAIFI Fellows are essential for developing connections across the IAIFI network, but the community also enables collaboration among graduate students and other postdocs. One such collaboration between the groups of Mike Williams (MIT) and Max Tegmark (MIT) aims to develop a physics-inspired theory of representation learning—in particular, understanding the phenomenon of “grokking,” where models generalize long after overfitting their training set. This project developed as a result of interactions among junior researchers at regular IAIFI events, beginning with a journal club talk and developing at an IAIFI networking event. By analyzing the behavior of the representations in algorithmic tasks, the researchers



**FIGURE 2** From <https://arxiv.org/abs/2206.14820>: (left) image of galaxy NGC2906; (middle) simulated effect of gravitational lensing on this galaxy, this is what would be observed if this galaxy did undergo gravitational lensing when viewed from earth; (right) AI prediction for what galaxy NGC2906 looks like given only the middle lensed image. The agreement with the true image, which the AI did not have access to, is striking.

were able to understand many aspects of grokking. Additionally, they were able to eliminate grokking in those settings and showed that grokking is mainly a residual effect from mistuning hyperparameters. This work highlights the usefulness of effective theories, commonly used in physics, to model neural network dynamics and adds to the growing body of work on neural network learning theory with tools from theoretical physics. This project was a highlighted contribution at NeurIPS 2022 (Liu et al. 2022), which demonstrates the value of physics-inspired research to the AI community.

## AI + PHYSICS EDUCATION AND ENGAGEMENT

AI technologies are advancing rapidly, making it both important and challenging to train junior researchers at the intersection of physics and AI and engage with the public. One effort, spearheaded by IAIFI Director Jesse Thaler, IAIFI Deputy Director Mike Williams, and Alexander Rakhlin, a Professor of Brain and Cognitive Science at MIT and an IAIFI Researcher, is the development of an interdisciplinary PhD program in Physics, Statistics, and Data Science (PhysSDS). This is a collaborative effort between the Department of Physics and the Statistics and Data Science Center at MIT<sup>4</sup>. Statistics and data science are among the foundational pillars of AI. Providing physics PhD students formal training in these areas fosters a new generation of leaders at the intersection of physics, statistics, and AI. The first interdisciplinary PhD degree was awarded in Spring 2021, and thus far seven such degrees have been awarded, with more students joining the program every term. Roughly half of the students who have obtained this degree have gone into academia, with the other half now working in industry.

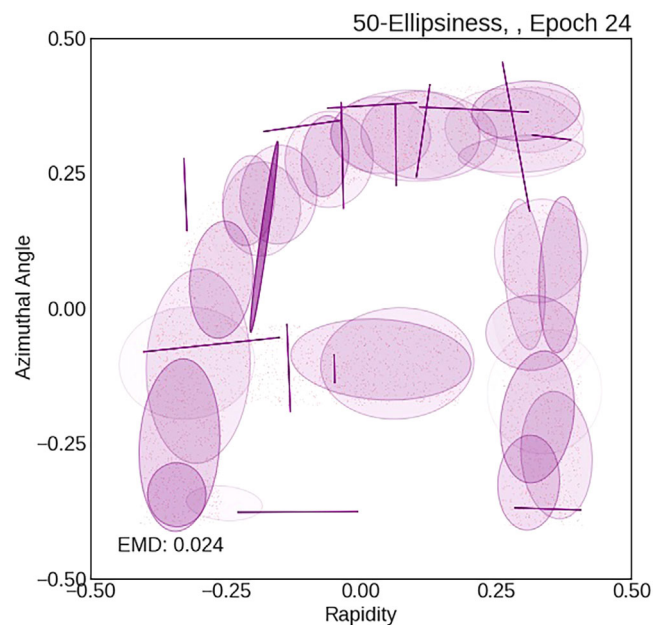
In parallel with developing the PhysSDS PhD, Phil Harris and Isaac Chuang (MIT), along with MITx Digital Learning Fellow Alex Shvonski, have created a new course, *Data Science in Physics*, which is offered at both the undergraduate and graduate level and presents modern computational methods in the context of realistic contemporary examples of their applications to physics research. For example, students are taught how to confirm the recent Nobel Prize for the discovery of gravitational waves at LIGO, then to improve on the published result using AI and fundamental physics principles. An online MITx version of the course has launched its first<sup>5</sup> and second<sup>6</sup> modules, and additional modules will be released later this year.

Another major IAIFI effort that reaches beyond the Boston area is the coordination of an annual PhD Summer School<sup>7</sup> followed by a Summer Workshop<sup>8</sup>, both at the intersection of AI and physics. The inaugural sessions in 2022 were well received: more than twice as many students than could be accommodated applied to the Summer School, with truly exceptional ratings given by the students in a post-school survey. The 2-day Summer Workshop was expanded in 2023 to a 5-day event to accommodate the wealth of expertise that researchers are eager to exchange in the AI + physics domain.

With projects like these, the goal is to disseminate knowledge about—and enthusiasm for—physics + AI.

## ETHICS AND EQUITY

To help ensure that the IAIFI is serving its junior members, an Early Career and Equity Committee (ECEC) was established within IAIFI. The ECEC meets monthly to discuss issues related to the well-being and work environment of IAIFI members, and advises IAIFI Management on ways to create a more equitable, inclusive, welcoming,



**FIGURE 3** The SHAPER framework can be used to model statistical distributions, such as the IAIFI logo!

and enjoyable place of work. In particular, the ECEC oversees all invite lists to ensure diversity considerations, advises IAIFI management on inclusive practices, and regularly surveys the IAIFI community on how well IAIFI is meeting their needs.

Of course, ethics in AI and AI safety are also important issues to address for all of the NSF AI Institutes. In addition to organizing events each year that address ethics in AI, the IAIFI is contributing to efforts to make AI more robust and understandable. For example, former IAIFI Fellow Anna Golubeva is dedicating a significant portion of her research to leveraging methods from theoretical physics to develop a thorough understanding of deep learning that will guide the design and improvement of AI systems. Building on her work on sparsity (Lasby et al. 2023), this research utilizes the framework of stochastic path integrals developed in statistical field theory to describe the stochastic process of learning in artificial neural networks, which leads to the ability to analyze, characterize, and understand neural-network-based machine learning methods, which form the core of modern AI.

## DEEP LEARNING + DEEP THINKING = DEEPER UNDERSTANDING

Part of what makes physics so powerful is that it provides a universal language that can be applied to a wide range of scientific problems. The IAIFI saw an early example of this in a collaboration between IAIFI Direc-

tor Jesse Thaler (MIT) and Demba Ba (Harvard). They were interested in connecting the languages of high-energy physics and dictionary learning, but were having trouble identifying the connection. Once their students began discussing the problem, they soon realized they had separately developed algorithms that were aiming to do the same thing—probe the geometric structure of statistical distributions. Working together, the two groups combined a principle from fundamental physics (infrared and collinear safety) with a subfield of AI (sparse coding and dictionary learning) to create a new algorithm with relevance in physics and beyond. The Shape Hunting Algorithm using Parameterized Energy Reconstruction (SHAPER) is a general framework for defining and computing shape-based observables (Ba et al. 2023) (see Figure 3). The efficacy of SHAPER has been demonstrated through empirical studies of jets, which are sprays of particles copiously produced at high-energy colliders. While the algorithm was developed specifically to study jets, it can be applied more broadly. In particular, SHAPER offers a different approach to density estimation, which can offer more flexibility than traditional dictionary learning while still enabling sparse representations of probability densities.

## CONCLUSION

Through the IAIFI, researchers are creating a common language that transcends the intellectual borders between physics and AI to facilitate a nexus point for groundbreaking discoveries. In doing so, the IAIFI is tackling two of the greatest mysteries of science: how our universe works and how intelligence works. By linking them, using physics to improve AI and AI to improve physics, the IAIFI is advancing physics knowledge and galvanizing AI research innovation. More broadly, a revolution is brewing in AI + science, and our efforts are aimed at positioning IAIFI to be a leader in this emerging field.

## ACKNOWLEDGMENTS

The material is based upon work supported by the National Science Foundation under Cooperative Agreement PHY-2019786. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.



## ORCID

Jesse Thaler  <https://orcid.org/0000-0002-2406-8160>

Mike Williams  <https://orcid.org/0000-0001-8285-3346>

## ENDNOTES

<sup>1</sup> <https://iaifi.org>

<sup>2</sup> <https://iaifi.org/current-fellows.html>

<sup>3</sup> See <https://fastmachinelearning.org> and <https://a3d3.ai>, both co-  
led by Prof. Phil Harris.

<sup>4</sup> [https://physics.mit.edu/academic-programs/graduate-  
students/psds-phd/](https://physics.mit.edu/academic-programs/graduate-students/psds-phd/)

<sup>5</sup> <https://mitxonline.mit.edu/courses/course-v1:MITxT+8.S50.1x/>

<sup>6</sup> <https://mitxonline.mit.edu/courses/course-v1:MITxT+8.S50.2x/>

<sup>7</sup> <https://iaifi.org/phd-summer-school>

<sup>8</sup> <https://iaifi.org/summer-workshop>

## REFERENCES

- Ba, D., A. S. Dogra, R. Gambhir, A. Tasissa, and J. Thaler. 2023. "SHAPER: Can You Hear the Shape of A Jet?" *Journal of High Energy Physics* 2023: 195. <https://doi.org/10.1007/JHEP06%282023%29195>
- Boyd, D., G. Kanwar, S. Racanièw, D. J. Rezende, M. S. Albergo, K. Cranmer, D. C. Hackett, and P. E. Shanahan. 2021. "Sampling Using SU(N) Gauge Equivariant Flows." *Physical Review D* 103(7). <https://doi.org/10.1103/PhysRevD.103.074504>
- Kitouni, O., N. Nolte, and M. Williams. 2022. "Finding NEEMo: Geometric Fitting using Neural Estimation of the Energy Mover's Distance." ArXiv. <https://doi.org/10.48550/arXiv.2209.15624>
- Lasby, M., A. Golubeva, U. Evci, M. Nica, and Y. Ioannou. 2023. "Dynamic Sparse Training with Structured Sparsity." ArXiv. <https://doi.org/10.48550/arXiv.2305.02299>
- Liu, Z., O. Kitouni, N. Nolte, E. J. Michaud, M. Tegmark, and M. Williams. 2022. "Towards Understanding Grokking: An Effective Theory of Representation Learning." ArXiv. <https://doi.org/10.48550/arXiv.2205.10343>
- Luo, D., and J. Halverson. 2023. "Infinite Neural Network Quantum States: Entanglement and Training Dynamics." *Machine Learning: Science and Technology* 4(2). <https://doi.org/10.1088/2632-2153/ace02f>
- Mishra-Sharma, S., and G. Yang. 2022. "Strong Lensing Source Reconstruction Using Continuous Neural Fields." ArXiv. <https://doi.org/10.48550/arXiv.2206.14820>
- Whittle, C., G. Yang, M. Evans, and L. Barsotti. 2023. "Maching Learning for Quantum-Enhanced Gravitational-Wave Observatories." *Physical Review D* 108(4). <https://doi.org/10.1103/PhysRevD.108.043034>

**How to cite this article:** Thaler, J., M. Williams, and M. LaFleur. 2024. "Institute for Artificial Intelligence and Fundamental Interactions (IAIFI): Infusing physics intelligence into artificial intelligence." *AI Magazine* 45: 111–16. <https://doi.org/10.1002/aaai.12150>

## AUTHOR BIOGRAPHIES

**Jesse Thaler** is a theoretical particle physicist who fuses techniques from quantum field theory and machine learning to address outstanding questions in fundamental physics. His current research is focused on maximizing the discovery potential of the Large Hadron Collider through new theoretical frameworks and novel data analysis techniques. Prof. Thaler joined the MIT Physics Department in 2010, and he is currently a Professor in the Center for Theoretical Physics. In 2020, he became the inaugural Director of the IAIFI.

**Mike Williams** is the Founder and Leader of the LHCb group at MIT and the inaugural Deputy Director of the IAIFI. The LHCb group at MIT is a Leader in the LHCb real-time data-processing system. Prof. Williams also works on advancing the usage of machine learning algorithms and other state-of-the-art data-science tools within the domain of particle physics research, and on advancing our understanding of AI itself. In 2020, he became the inaugural Deputy Director of the IAIFI.

**Marisa LaFleur** is the Project Manager for IAIFI. She joined in 2021 after a decade of experience in scientific publishing, where she worked with researchers to produce books on interdisciplinary topics ranging from space and planetary science to environmental science to food science. Her work at IAIFI is focused on implementing strategy, managing broader impacts activities, and communicating both internally and externally.



## SPECIAL TOPIC ARTICLE

# Molecule Maker Lab Institute: Accelerating, advancing, and democratizing molecular innovation

Martin D. Burke | Scott E. Denmark | Ying Diao | Jiawei Han | Rachel Switzky | Huimin Zhao

University of Illinois, Urbana-Champaign, Illinois, Urbana, USA

**Correspondence**

Huimin Zhao, University of Illinois, Urbana-Champaign, Urbana, IL, USA.  
Email: zhao5@illinois.edu

**Funding information**

National Science Foundation, Grant/Award Number: 2019897

**Abstract**

Many of the greatest challenges facing society today likely have molecular solutions that await discovery. However, the process of identifying and manufacturing such molecules has remained slow and highly specialist dependent. Interfacing the fields of artificial intelligence (AI) and synthetic organic chemistry has the potential to powerfully address both limitations. The Molecule Maker Lab Institute (MMLI) brings together a team of chemists, engineers, and AI-experts from the University of Illinois Urbana-Champaign (UIUC), Pennsylvania State University, and the Rochester Institute of Technology, with the goal of accelerating the discovery, synthesis and manufacture of complex organic molecules. Advanced AI and machine learning (ML) methods are deployed in four key thrusts: (1) AI-enabled synthesis planning, (2) AI-enabled catalyst development, (3) AI-enabled molecule manufacturing, and (4) AI-enabled molecule discovery. The MMLI's new AI-enabled synthesis platform integrates chemical and enzymatic catalysis with literature mining and ML to predict the best way to make new molecules with desirable biological and material properties. The MMLI is transforming chemical synthesis and generating use-inspired AI advances. Simultaneously, the MMLI is also acting as a training ground for the next generation of scientists with combined expertise in chemistry and AI. Outreach efforts aimed toward high school students and the public are being used to show how AI-enabled tools can help to make chemical synthesis accessible to nonexperts.

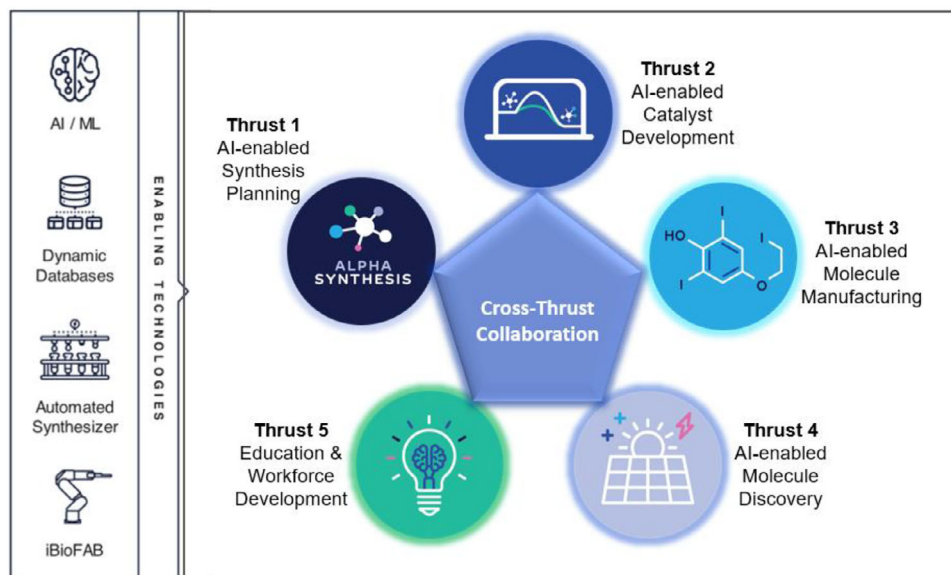
**INTRODUCTION**

The long-term strategic goal of the Molecule Maker Lab Institute (MMLI) is to accelerate, advance, and democratize molecular innovation. To achieve this, the MMLI is creating an open, exciting, and dynamic interdisciplinary

ecosystem that will catalyze highly impactful and inclusive collaborations between world-leading PIs, top-notch students, postdocs, and fellows in AI and chemistry, and passionate and creative leaders in education and community engagement. The MMLI is a first-of-its-kind research infrastructure that is having a powerful impact on the

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Overview of the Molecule Maker Lab Institute.

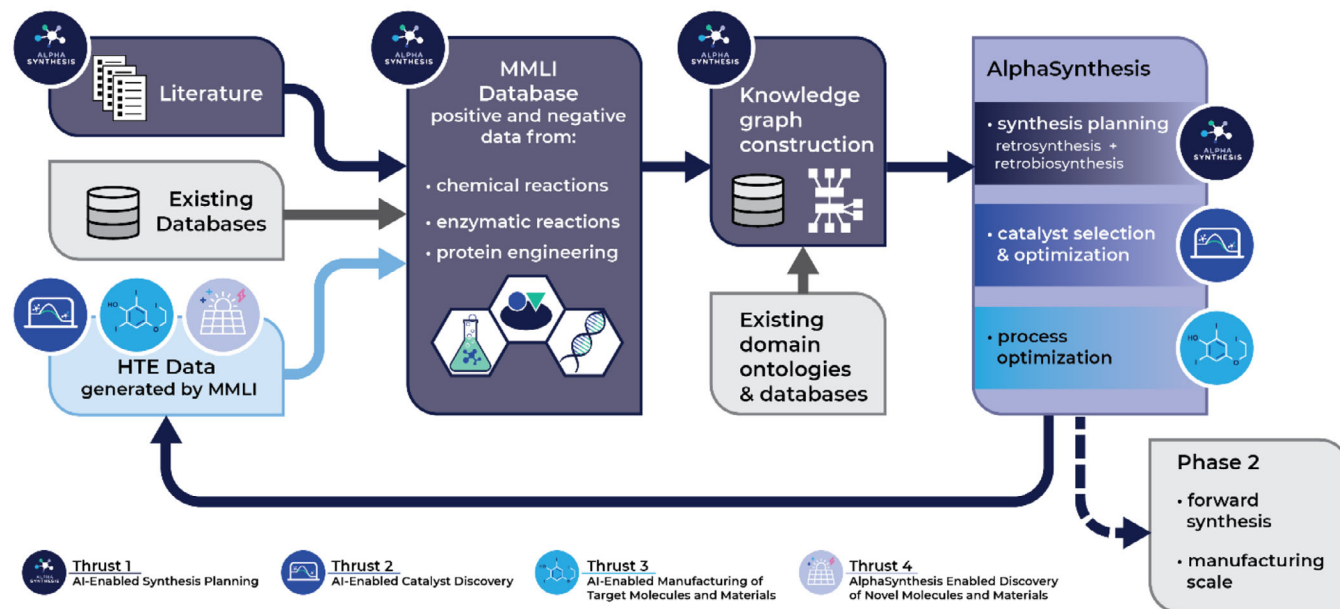
U.S. research community. The MMLI is achieving a broad impact by tailoring opportunities to a variety of audiences, responding to high priority needs of communities seeking to (1) discover and optimize a wide range of molecular functions, (2) harness the power of data to advance the science of molecular synthesis, and (3) inspire a broad audience of scientists, teachers, students, and citizen scientists to participate in the process of molecular innovation.

## USING AI TO ACCELERATE, ADVANCE, AND DEMOCRATIZE MOLECULAR INNOVATION

Many of the greatest challenges facing society today likely have molecular solutions that await discovery. However, the process of identifying and manufacturing such molecules has remained slow and highly specialist dependent. Interfacing the fields of AI and synthetic organic chemistry at the MMLI has the potential to powerfully address both limitations. Through this interdisciplinary initiative, leaders in AI and organic synthesis (both chemical and biological) are intensively collaborating to create frontier AI tools, dynamic open access databases, and fast and broadly accessible small-molecule manufacturing and discovery platforms (Figure 1). Specifically, advanced AI and machine learning (ML) methods are being developed and deployed in the context of four key thrusts focused on: the design of highly effective and, in many cases, Lego-like modular and automatable, synthetic routes for manufacturing and discovering a wide range of small molecules (Thrust 1, AI-enabled synthesis planning), the

design and development of optimized chemical and biological catalysts for promoting important reactions with broad potential utility in small-molecule synthesis, including iterative and automated building block assembly processes (Thrust 2, AI-enabled catalyst development), the efficient manufacture of three key molecules already known to perform useful functions, including C2'epi amphotericin B (a novel potent and nontoxic antifungal drug candidate), artemisinin (a critical antimalaria drug), and Millad NX 8000 (an environmentally advantageous colorless, odorless thermoplastic clarifier for polypropylene) (Thrust 3, AI-enabled molecule manufacturing), and the discovery of efficient and stable next generation organic photovoltaics (OPVs) via a fully automated closed-loop autonomous discovery platform (Thrust 4, AI-enabled molecule discovery).

A major challenge and a transformative opportunity for science, engineering, and society lies in bringing the power of making molecules to everyone. The MMLI is developing a versatile and flexible synthesis planning tool named AlphaSynthesis (Figure 2), inspired by AlphaGo and AlphaFold. This tool uses AI to design, construct, and optimize the most effective and automatable synthetic routes, exploiting both chemical and biological catalysts to manufacture target molecules and discover new molecules. The Institute has also created an open-access database that includes all the building blocks, reagents, conditions, and yields (products and byproducts) for every coupling reaction that is run in the MMLI. This database will also have the unique feature of being accessible for on-demand content optimization via automated reaction executions in response to cloud-submitted requests from computer scientists. The MMLI is also developing new AI



**FIGURE 2** Overview of the development of AlphaSynthesis, an artificial intelligence (AI) platform that enables the design, construction, and optimization of the synthetic routes for any molecules.

and ML algorithms and tools for designing and optimizing the catalysts that are required for the implementation of the synthetic routes designed by AlphaSynthesis. To demonstrate the power of these frontier AI tools for synthesis, the MMLI is designing and establishing the most efficient synthetic routes for important molecules and discovering new OPV molecules. These collective activities will increase the efficiency with which small molecules can be manufactured and discovered, drive the advanced development of a wide range of frontier AI methods, and broaden access to the small molecule making process.

Notably, the MMLI is not simply an integrated collection of equipment for automated molecular synthesis and software. It is an open ecosystem of disruptive thinking, education, and community engagement powered by state-of-the-art molecular design, synthesis, and spectroscopic characterization technologies all interfaced with a modern cyberinfrastructure.

## RECENT RESEARCH ACCOMPLISHMENTS

In Thrust 1, we are focusing on the development of AlphaSynthesis. We are advancing foundational AI research, as highlighted by multiple papers accepted and published by top tier journals such as *Science* and *Nature Catalysis*, and top AI conferences (Angello et al., 2022; Rose et al., 2022; Wang et al., 2022a; Wang et al., 2022b; Yu et al., 2023a; Yu et al., 2023b; Yu et al., 2023c). Beyond publications, MMLI is providing novel foundational AI advances

through new information sources and new analytical techniques such as reaction aware multimodal molecular representation and a translator between natural language and molecules. Importantly, this foundational AI work is not confined to the chemistry discipline. These tools and algorithms synergize with and can be applied to other scientific disciplines. Here, we provide three examples to demonstrate this key point.

First, newly developed molecule language models have been built (Edwards et al., 2022) and created results which can be tested in the physical world. Specific focus has been on extracting information from the chemistry literature to enable new insights into how specific structural components of kinase inhibitors, drugs commonly used to treat cancer, are connected to important therapeutic properties, such as blood brain barrier penetration. This work has the potential to yield new AI-based tools to advance the design and discovery of new and improved drugs, which may save researchers years of effort and millions of dollars, not to mention possibly saving lives. Recently, the MMLI has proposed a novel in-context learning framework for personalized drug synergy prediction. This exciting work may eventually enable the creation of a standardized assay for predicting patient-tumor-specific drug synergies, allowing highly targeted combination cancer therapy.

Second, automated extraction of structured knowledge from text-intensive, unstructured scientific literature is a fundamental challenge in AI. To address this challenge, the MMLI is building a ReactIE system which automatically extracts essential and structured chemistry reaction information from chemistry research literature (Zhong





et al., 2023). The structured chemical reaction information includes products, reaction types, reactants, catalysts, solvents, temperatures, and yields, without the assistance of human annotation. The method can be applied to the extraction of other structured information from general scientific literature. The method combines two weakly supervised approaches for pretraining. It utilizes frequent patterns within the text as linguistic cues to identify specific characteristics of chemical reactions. Moreover, it utilizes synthetic data from patent records as distant supervision to incorporate domain knowledge into the model. Experiments demonstrate that ReactIE achieves substantial improvements and outperforms all existing state-of-the-art methods.

Third, the MMLI has created ChemScraper, a formula parser for molecules in born-digital PDF papers, where graphics are defined using PDF drawing instructions (e.g., where characters and lines are given explicitly). Advantages of parsing from born-digital diagrams include avoiding OCR and thereby preventing recognition errors and increasing speed (currently 200 ms/molecule using unoptimized python). The first version of the parser for born-digital molecule diagrams has been completed and ChemScraper has been integrated into the AlphaSynthesis platform.

In Thrust 2, advanced AI and ML methods are being developed and deployed in the design and development of optimized chemical and biological catalysts for promoting important reactions with broad potential utility in small-molecule synthesis. For example, we recently developed an AI tool named Contrastive Learning enabled Enzyme ANnotation (CLEAN) for predicting enzyme function from its sequence (Yu et al., 2023b), which can help the selection of proper enzymes in the designed synthetic routes. Enzyme function annotation is an urgent need and challenge. Many computation tools have been developed including alignment-based methods and ML-based methods (Altschul et al., 1997; Sanderson et al., 2023). However, most of the tools cannot reliably annotate protein functions such as enzyme commission (EC) number, partly due to the fundamental challenge of highly imbalanced enzyme dataset. By contrast, CLEAN can assign EC numbers to query enzyme sequences with better accuracy, reliability, and sensitivity than existing methods as shown in various in silico benchmark studies. Most importantly, we demonstrate that using contrastive learning frameworks, CLEAN can better handle the biased dataset compared to other ML models.

In addition, we have also tested CLEAN on in-house curated halogenase dataset. In particular, we experimentally evaluated three of the halogenases (MJ1651, TTHA0338, and SsFIA) which CLEAN predicted differently with the current annotation in the database. The

three halogenase represented three different cases for mis-annotated, un-annotated and partially annotated in the database. In vitro experiments result shows CLEAN has successfully annotated these three halogenases which other methods cannot.

Furthermore, we have implemented a free-to-use web interface version of CLEAN for the broad research community. As part of the MMLI AlphaSynthesis platform, the web tool of CLEAN can be accessed at Clean (<https://clean.frontend.mmlil.ncsa.illinois.edu/configuration>). After the tool went online, it gained wide community attention. In the first 2 months alone, the web page attracted 20,100 page views and 12,531 of them were from unique viewers. Since the initial release of the website, we have been constantly updating many features, including better result filtering and displaying, computing efficiency, confidence level display, and general user experience. It is worth noting that many of the new features were suggested by our users through the feedback function we provide.

Thrust 3 focuses on leveraging and advancing AI in the context of developing processes for synthesizing on large-scale target molecules with known functions. These include chemical, enzymatic, and hybrid chemoenzymatic routes designed to maximize efficiency in each case. There is specific focus on the development of scalable syntheses for three target molecules; Millad NX8000, C2'epiAmB, and artemisinin, all three of which have important functions driving the need for large-scale access, but none of which have optimal large-scale syntheses. Millad NX8000 is a plastics clarifying agent synthesized yearly on ~6000 metric tons/year scale but requires superstoichiometric tin and generates ~9000 metric tons/year of toxic tin byproducts. C2'epiAmB is an exceptionally promising new renal sparing polyene antifungal, a variation of which has entered clinical trials in the U.S. and there is substantial room for improvement in its current synthesis. Artemisinin is a critically important antimalarial compound on the WHO's list of essential medicines, but all current routes for its manufacture are too expensive to meet global supply demands.

The focus of Thrust 4 is on progress to discover novel OPVs, which can transform the way we harness and use sunlight. The key goal we are working toward is OPVs with high efficiency and high stability. To date, no one has been able to achieve both parameters. The MMLI's approach to address this problem is three-fold. First, the small amount of literature on AI-guided optimization of OPVs focuses on power conversion efficiency. However, the major challenge in this area has become device stability now that the best cell is approaching 20% efficiency. There is a critical gap of relating molecular design to stability which is exactly what the MMLI is addressing. Second, the majority of AI for OPV literature mines published literature or



computed properties, given scarce experimental data that are consistent. Unfortunately, this severely limits the predicting power of AI. Thrust 4 is addressing this issue head on with high throughput syntheses and characterizations to rely primarily on experimentally generated data. Finally, other approaches to OPV discovery use AI optimization of device properties as a function of formulation and processing conditions, but this approach is intrinsically limited to the maximum possible efficiency of the active layer chemistry. The MMLI's innovative strategy uses AI to guide a search through a large synthetic space of OPV chemistries.

The MMLI is targeting OPVs with >10% efficiency and >10 years lifetime to be commercially viable for next-generation energy capture applications and for mitigating climate change. To achieve this goal, autonomous synthesis, automated characterization, and AI-based methods are integrated into a closed-loop approach to drive molecular discovery guided by target criteria for OPV performance: efficiency and stability.

While these thrusts are listed separately it should be noted that there is active and intentional collaboration between thrusts to further enable research and discovery. These collaborations not only increase the impact of the institute but also provide students and postdocs an opportunity for learning the value of collaborating with other research teams.

## ENGAGING INDUSTRY PARTNERS

The MMLI has an active Industrial Partnership Program with the goal of providing the opportunity for the two-way exchange of information between the MMLI and industry researchers. The program is a way for MMLI researchers to share the tools and databases being developed for more efficient synthesis and discovery of chemical and materials for a wide-range of applications. Industry researchers can provide perspective on which projects will most benefit society. There is a two-tier membership structure (Partner and Associate), with the annual membership fees going into a seed grant fund to support proof-of-concept proposals from the broader community and expand partnerships. The MMLI continues to grow this program with significant input from the current industry partners.

## CULTIVATING MOLECULAR INNOVATORS

As MMLI increases its access and exposure to learners and the general public, it is critical that all of the Thrust 5 initiatives share a common framework with which to situate particular contributions to the larger MMLI identity

of democratization. The dimension of identity spans from “Chemical Learners” to “Molecular Innovators”. At one extreme of the dimension, chemical learners are focused on a structural and formal exploration of chemistry via MMLI's tools and research. At the opposite extreme, molecular innovators are leveraging the work of MMLI as problem-solving tools focused on the function of molecules and do not engage with formal chemical representations. This span of interaction identities is then framed by the mediums in which they engage with MMLI across the physical and digital resources (in-class labs, camp activities, online lessons, etc.). The goal with this framework is to ensure that MMLI grows not only considering the extremes of the framework, but also to facilitate blending and transitioning between dimensions to provide a more holistic engagement experience whenever possible.

The MMLI is revolutionizing the way chemistry is taught and capture the imagination of a new generation of molecule makers by building on our already established momentum of engagement with educators and students through several mechanisms, including but are not limited to the MMLI in a Box, a Digital Molecule Maker (DMM), as well as establishing international and industrial partnerships. The MMLI aims to democratize the molecule making process and support the next generation of molecule makers, thereby having a wide range of broader impacts.

The MMLI in a Box transforms all the powers of a real lab into carefully crafted low-tech material (e.g., lenticular printed cards, 3D printed macroscopic models), and highly intuitive, engaging hands-on and role-play activities inspired by the latest Next Generation Science Standards guidelines (NGSS Lead States, 2013). These materials and activities empower teachers to help their students enter the awe-inspiring world of molecules, AI, and AI-powered molecule making.

The DMM is the product of a collaboration between the MMLI and the Siebel Center for Design at the UIUC. The DMM interface allows researchers to make their chemical building-blocks available for exploration by learners, who can then combine them while getting dynamic feedback properties of a molecule during its construction. The DMM offers the opportunity for learners to contribute to actual research initiatives by facilitating the connection between a molecular block set, a proposed molecule, and its creation via automated synthesis. The current version of the DMM allows learners to engage with a “10×10×10” (ten start, middle, and end blocks) molecular block set from the Burke Lab at the UIUC. This block set was created to explore molecular structure and light absorption. Already, student molecule submissions are assisting the lab in improving the quality of AI-assisted molecular property predictions and testing the automated synthesis workflow.



The DMM coupled with the MMLI in a Box further supports the development of such intuitions, as well as creativity and imagination, by allowing students to explore thousands of instances of Lego-like building of molecules as they design their own new molecule with novel functions. Finally, synthesizing student-designed molecules allows students to reflect on their intuitions.

## MMLI IN THE CLASSROOM

MMLI has partnered with the School of Chemical Sciences at UIUC to implement a three-part sequence into the undergraduate labs. The curriculum in this sequence aims to integrate both data science and automated synthesis into both general and organic chemistry courses. Part 1 is an adaptation of the K-12 version of the MMLI in a Box where students are exposed to the power of molecules, synthesize dyes in a mix-and-match fashion, participate in a novel role-playing game, and combine all that knowledge to analyze a data set from an original research project. During the 2022–2023 Academic Year, ≈1050 students in General Chemistry I and II participated in the MMLI in a Box activity.

In Spring 2023, the second part of the three-part sequence was piloted at the General Chemistry II and Organic Chemistry I level with an activity focused on automating synthesis. Over 100 students came to the Molecule Maker Lab space to participate in the activity to leverage skills that were directly applicable to the work of MMLI, providing a source of inspiration and tangible connection to the science of automated small-molecule synthesis.

## CONCLUSION

The MMLI is committed to accelerating, advancing, and democratizing molecular innovation by creating an open, exciting, and dynamic interdisciplinary ecosystem that will catalyze highly impactful and inclusive collaborations between world-leading PIs, top-notch students, postdocs, and fellows in AI and chemistry, and passionate and creative leaders in education and community engagement. The MMLI is transforming chemical synthesis and generating use-inspired AI advances. Simultaneously, the MMLI is also acting as a training ground for the next generation of scientists with combined expertise in chemistry and AI. The long-term impact of the MMLI will be substantial and is expected to shift the paradigm of how molecules are discovered and used to address society's grand challenges. These collective activities will powerfully enable the more efficient manufacturing and

discovery of molecules with important functions, drive the advanced development of a wide range of frontier AI methods, and broaden access to the small molecule making process. More information can be found on the MMLI website (<https://moleculemaker.org/>).

## ACKNOWLEDGMENTS

The material is based upon work supported by the National Science Foundation under Grant No. 2019897. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Martin D. Burke  <https://orcid.org/0000-0001-7963-7140>

Scott E. Denmark  <https://orcid.org/0000-0002-1099-9765>

Ying Diao  <https://orcid.org/0000-0002-8984-0051>

Jiawei Han  <https://orcid.org/0000-0002-3629-2696>

Rachel Switzky  <https://orcid.org/0009-0004-6225-5989>

Huimin Zhao  <https://orcid.org/0000-0002-9069-6739>

## REFERENCES

- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25: 3389–402.
- Angello, N. H., C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski, and D. Burke. 2022. "Closed-loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling." *Science* 378: 399–405.
- Edwards, C., T. M. Lai, K. Ros, G. Honke, K. Cho, and H. Ji. 2022. "Translation between Molecules and Natural Language." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022).
- NGSS Lead States. 2013. *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.
- Rose, T., J. C. Timmerman, S. A. Bawel, S. Chin, H. Zhang, and S. E. Denmark. 2022. "High-Level Data Fusion Enables the Chemoinformatically Guided Discovery of Chiral Disulfonimide Catalysts for Atropselective Iodination of 2-Amino-6-arylpyridines." *Journal of the American Chemical Society* 144: 22950–64.
- Sanderson, T., M. L. Bileschi, D. Belanger, and L. J. Colwell. 2023. "ProteInfer: Deep Networks for Protein Functional Inference." *eLife* 12: e80942.
- Wang, H., W. Li, X. Jin, K. Cho, H. Ji, J. Han, and M. Burke. 2022b. "Chemical-Reaction-Aware Molecule Representation Learning." In Proceedings of the International Conference on Learning Representations (ICLR2022).
- Wang, X., V. Hu, M. Jiang, Y. Zhang, J. Xiao, D. Loving, H. Ji, M. Burke, and J. Han. 2022a. "ReactClass: Cross-Modal Supervision for Subword-Guided Reactant Entity Classification." In



- Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'22), Las Vegas, NV.
- Yu, T., A. Boob, N. Singh, Y. Su, and H. Zhao. 2023c. "In vitro Continuous Protein Evolution Empowered by Machine Learning and Automation." *Cell Systems* 14: 633–44.
- Yu, T., A. G. Boob, M. J. Volk, X. Liu, H. Cui, and H. Zhao. 2023a. "Machine Learning-Enabled Retrobiosynthesis of Molecules." *Nature Catalysis* 6: 137–51.
- Yu, T., H. Cui, J. Li, Y. Luo, G. Jiang, and H. Zhao. 2023b. "Enzyme Function Prediction Using Contrastive Learning." *Science* 379: 1358–63.
- Zhong, M., S. Ouyang, M. Jiang, V. Hu, Y. Jiao, X. Wang, and J. Han. 2023. "ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision." In Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics (ACL'23), Toronto, Canada.

**How to cite this article:** Burke, M. D., S. E. Denmark, Y. Diao, J. Han, R. Switzky, and H. Zhao. 2024. "Molecule Maker Lab Institute: Accelerating, advancing, and democratizing molecular innovation." *AI Magazine* 45: 117–23. <https://doi.org/10.1002/aaai.12154>

## AUTHOR BIOGRAPHIES

**Martin D. Burke** is a Professor of Chemistry at the UIUC and leads MMLI Thrust 3.

**Scott E. Denmark** is a Professor of Chemistry at the UIUC and leads MMLI Thrust 2.

**Ying Diao** is an Associate Professor of Chemical and Biomolecular Engineering at the UIUC and leads MMLI Thrust 4.

**Jiawei Han** is a professor of Computer Science at the UIUC and leads MMLI Thrust 1.

**Rachel Switzky** is the inaugural director of the Siebel Center for Design at the UIUC and leads MMLI Thrust 5.

**Huimin Zhao** is a professor of chemistry, biochemistry, biophysics, and bioengineering at the UIUC and is the director of the MMLI. For more information on the authors, see the MMLI website ([moleculemaker.org](http://moleculemaker.org)).



SPECIAL TOPIC ARTICLE

# AI-CARING: National AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups

Sonia Chernova<sup>1</sup> | Elizabeth Mynatt<sup>2</sup> | Agata Rozga<sup>1</sup> | Reid Simmons<sup>3</sup> | Holly Yanco<sup>4</sup>

<sup>1</sup>Georgia Tech, Atlanta, Georgia, USA

<sup>2</sup>Northeastern University, Boston, Massachusetts, USA

<sup>3</sup>Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

<sup>4</sup>UMass Lowell, Lowell, Massachusetts, USA

## Correspondence

Sonia Chernova, Georgia Tech, Atlanta, GA, USA.

Email: [chernova@gatech.edu](mailto:chernova@gatech.edu)

## Funding information

Division of Information and Intelligent Systems, Grant/Award Number: 2112633

## Abstract

Over 13 million Americans aged 65 and older are currently living with a diagnosis of mild cognitive impairment (MCI), a common precursor to dementia. These individuals largely rely on a network of informal caregivers—family, friends, and community members—who work together with professional healthcare and social service providers to provide care and support in home settings. The AI-CARING Institute contributes foundational AI research focused on developing personalized collaborative AI systems that improve the quality of life and independence of aging adults living at home.

## INTRODUCTION AND INSTITUTE OVERVIEW

Over the past decade, advances in deep learning, natural language processing, and pattern recognition have driven the age of tailored product recommendations and hyper-personalized services. However, today's AI systems are yet to achieve meaningful, dynamic, and fluid longitudinal collaboration with diverse groups of people, remaining largely restricted to short, single-focus interactions.

The NSF AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups (AI-CARING) seeks to create a vibrant, fully developed discipline focused on *longitudinal collaborative AI*—a field characterized by the design, development, and deployment of interactive, intelligent systems embedded within communities of users over extended periods of time (months

and years). The institute's efforts contribute computational methods and datasets to address challenges such as sample-efficient techniques for user modeling and personalization, robust methods for longitudinal human–AI teaming, socially conscious and dignity-preserving AI methods, explainable and transparent systems, innovative guidelines for experimental design, and novel benchmarks and metrics for these areas.

The institute's foundational AI research is grounded in the mission to develop personalized collaborative AI systems that improve the quality of life and independence of aging adults living at home. While the outcomes of our work are broadly applicable, AI-CARING's particular focus is on supporting individuals diagnosed with mild cognitive impairment (MCI) and their care partners. MCI is a common precursor to dementia; 33% of MCI patients develop dementia within 5 years. There are more than 54 million older adults aged 65 and older currently living

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.

in the United States, approximately 20% of whom have MCI. Because MCI patients are not eligible for subsidized care facilities, the burden of caregiving most often falls on their family members; it is estimated that over 11 million people in the United States provide unpaid care for people with dementia-related illnesses each year, costing the U.S. economy over \$339 billion. AI-CARING seeks to ease the burden of these caregivers, improve quality of life for the cared-for individuals, increase longevity and independence, provide better opportunities for people with disabilities, and reduce health care costs.

To carry out this work, the AI-CARING team conducts long-term deployments in households that include people diagnosed with MCI, their family, caregivers, and professional providers. AI-CARING systems will reinforce daily routines, recognize changes in behavior, provide team support for caregivers, and provide encouragement and feedback in response to an individual's changing abilities. The institute team brings together researchers from computing, social sciences, and healthcare to collaborate in the design, development, and deployment of AI systems that help people sustain independence; create new opportunities for empowerment and meaningful quality of life; and improve the effectiveness of family, community, and professional care.

Our vision is to create a range of AI services designed to enhance daily activities, and that are ultimately capable of discerning personalized, long-term models of human behavior. We expect that these systems will adapt to the evolving needs of older adults to support their healthcare and independence goals. We envision harmonized teams of interconnected AI agents that offer personalized assistance not only for patients but also their caregivers, ultimately improving interpersonal relationships that maximize the wellbeing of the end users.

Finally, AI-CARING is dedicated to advancing workforce development through comprehensive initiatives that encompass education, outreach, inclusivity, and knowledge transfer programs. Our aim is to contribute to the development of next generation of talent in this field. Below, we describe key opportunities and insights relating to the Institute's efforts.

## PERSONALIZED LONGITUDINAL INTERACTION

AI-CARING seeks to establish a vibrant discipline of *longitudinal collaborative AI*, focusing on teams of heterogeneous agents assisting communities of users through adaptive and ethically guided interaction over extended periods of time. Central to this objective is the devel-

opment of longitudinal, personalized models of a given user, including their behavior, abilities, and communication style, and the use of these models to identify changes in behavior and preferences over time, and to learn tasks that the AI can perform to support and complement the user's activities.

Developing personalized models of each user presents a significant challenge for any AI system that is required to cope with the vast diversity of human behavior. For our use case, human activity recognition (HAR) presents the most significant challenge, because the envisioned interactive smart home assistant will be of limited use without understanding the user's behavior. Recent advances in self-supervised learning have created opportunities to develop deployable HAR systems that exploit unlabeled data to derive reliable recognition systems for scenarios where only small amounts of labeled training samples can be collected. As such, self-supervision, that is, the paradigm of "pretrain-then-finetune" has the potential to become a strong alternative to the previously dominant end-to-end training approaches. Recently a number of contributions have been made that introduced self-supervised learning into the field of HAR, including benchmarks of leading techniques (Haresamudram, Essa, and Plötz 2022) and novel computational methods (Haresamudram, Essa, and Plötz 2023).

Beyond activity recognition, personalized language understanding and generation are essential for creating AI systems that effectively adapt to users' individual language styles, demographics, and vocabulary. AI agents that are capable of capturing user styles and accommodating them by employing varying degrees of simple and easy language during communication can provide a more engaging and inclusive user experience. To develop such an AI system, our work introduces three interconnected works: (i) MultiPIT is a large-scale multitopic paraphrase dataset collected from Twitter, which aids in training AI models to understand and generate diverse paraphrases, enabling them to communicate more effectively with users by adjusting their language style and accommodating a wide range of user preferences (Dou, Jiang, and Xu 2022); (ii) LENS is a learnable metric for text simplification trained on human judgments (Maddela et al. 2023). Besides accurately measuring generation's simplicity and quality, LENS can be incorporated into generation to dynamically adapt the level of readability, achieving state-of-the-art performance; and (iii) SALSA is a fine-grained edit-level human evaluation framework (Heineman et al. 2023). It enables the identification of errors and quality edits in AI-generated simplifications, providing insights into AI systems' strengths and weaknesses. This is crucial for refining and improving AI-driven communication tailored to each user's readability requirements.



Finally, our systems seek to provide personalized assistance that is capable of adapting to the needs and preferences of unique individuals. Such assistance may come in many forms. For example, a conversational agent may provide reminders for daily tasks, such as taking medications. Improving medication management for older adults with MCI requires designing systems that support functional independence and provide compensatory strategies as individual abilities change. Traditional medication management interventions emphasize forming new habits alongside the traditional path of learning to use new technologies. Our work has taken a novel approach, contributing a system tailored for older adults with gradual cognitive decline by creating a conversational “check-in” system for routine medication management. Our results from a 20-week deployment indicate that a conversational check-in medication management assistant increased system acceptance while also potentially decreasing the likelihood of accidental over-medication, a common concern for older adults dealing with MCI (Mathur et al. 2022). Robotic technologies present another opportunity for interactive assistance in a household setting. Learning an individual’s task preferences and routines enables a robot to provide proactive assistance (Patel and Chernova 2022, 2023), as well as detect deviations from and assist in recovery with routine tasks.

## ROBUST MULTIAGENT COLLABORATION

Caregiving is often a shared responsibility amongst many people—family members, friends, neighbors, and clinicians. Leveraging the support of the entire care network is critical for effective and sustainable care. Coordinating sustained care across a distributed team of caregivers presents many challenges that stem from changing conditions (e.g., changes in health, diagnosis, or abilities) and changing team structure (e.g., introduction of a new healthcare provider, or departure of a caregiver). AI systems that contribute to care must have the ability to detect and respond to such changes, appropriately model and respond to social factors that govern human relationships, and communicate with all involved parties in an effective way.

Recent findings from social sciences have advanced our understanding of human teaming behaviors, which in turn can be used to inform the design of autonomous agents. In particular, Woolley et al. (2023) introduce the concept of *collective intelligence*, which captures a team’s ability to work together across a wide range of tasks and can vary significantly between teams. In a series of works, the authors examine how team structure affects collective intelligence and team performance (Woolley et al. 2023), and how AI

technologies can enhance collective intelligence in distributed teams (Gupta et al. 2019; Woolley, Gupta, and Glikson 2023). These findings are critical for informing the development of AI agents capable of supporting complex interactions within a network of caregivers.

In other ongoing work on multiagent teaming, we consider the challenge of modeling human agents and their behavior. Many existing ad hoc teaming algorithms assume that coordination takes place between static team members, however humans are both challenging to model accurately and typically change in behavior over time. Our recent work tackles these challenges through the development of robust teaming techniques that enable agents to rapidly adapt to changing teams (Cook, Scheiner, and Tumer 2023) cooperate with agents who have different objectives and capabilities (Dixit and Tumer 2023).

## SOCIALLY CONSCIOUS AI

To be effective, intelligent agents must interact with users in a manner that is socially appropriate and engenders appropriate levels of trust from the user. Trust between the interactive assistant and the user is vital to effective collaboration and team cohesion. As in prior work, we define trust as “a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor’s risk in a situation in which the trustor has put its outcomes at risk” (Wagner, Robinette, and Howard 2018). In human–AI collaborations, trust has two temporal states: initial trust based on user demographics and preexisting knowledge, and interactive trust based on user processes, robot performance, and communication.

Furthermore, there are three states of trust in a human–AI relationship: *appropriate trust*, where trust aligns with the trustee’s capabilities; *over-trust*, where the trustor believes the trustee can do better than it actually can; and *under-trust*, where the trustor does not believe the trustee can do what it is actually capable of. To calibrate trust, the situation causing the violation must be identified, and an appropriate response must be selected to adjust the trust relationship. Conversely, if trust has been repaired too well, or the person overestimates the system’s capabilities, the trust must be dampened. To do so effectively, the state of over-trust must be identified, and an effective strategy for calibration determined. As interactions between humans and AI increase and become more meaningful, regulating trust to avoid over-trust and under-trust becomes a critical challenge.

Recent works have contributed insights into this research area, including techniques for trust calibration through verbal cues warning users of potential agent failure modes (Perkins et al. 2022) and multihop conversational engagement that enables users to probe the agent’s

reasoning (Xie, Wiegrefe, and Riedl 2022). Despite recent advances, trust calibration remains an open research problem as the diversity and complexity of human–AI interactions makes it difficult to assess and adapt user trust in a generalizable way.

### AI-CARING educational initiatives

AI-CARING has developed and deployed pivotal educational experiences for K-12, undergraduate, graduate students, and teachers, with the aim of promoting the growth of a diverse and global research community and future workforce. Our educational initiatives integrate insights and results from our AI research and our use-inspired human-centered research.



**Robotics caregivers course:** Robotics researchers and futurists have long dreamed of robots that can serve as caregivers. The *Robotic Caregivers* course, first developed by Prof. Charles Kemp at Georgia Tech in 2021, offers students hands-on experience in designing and deploying an autonomous caregiving robot. Using the Hello Robot Stretch RE1 mobile manipulation platform (HelloRobot 2023), students learn about future opportunities for, and present realities of, robots that contribute to caregiving, and work closely with stakeholders, including healthcare professionals and target users. *Robotic Caregivers* has been expanded beyond Georgia Tech to Carnegie Mellon, UMass Lowell, Oregon State, and the University of Washington.



### K-12 teacher professional development:

Carnegie Mellon has developed a week-long workshop to support high school educators looking to gain familiarity with AI and to offer AI-related educational activities to their students. Developed by Prof. Stephanie Rosenthal, the *Crash Course in Artificial Intelligence* program (Rosenthal 2023) covers an introduction to a broad range of AI topics and interactive hands-on design sessions in which teachers collaboratively develop course materials that they feel would most resonate with their students. The program completed its second year in 2023, with 37 teachers attending from across the United States. This workshop was adapted by Julia Kim in 2023 for Georgia's \$65M Build Back Better grant (Georgia AIM) in rural areas, with 11 teachers and six community leaders from six school districts in south Georgia participating.



**BridgeUp STEM** Funded by the Helen Gurley Brown Foundation and realized through a partnership between the National Center for Women & Information Technology (NCWIT) and Georgia Tech, the *BridgeUP STEM* program BridgeUP-STEM (2023) is designed to encourage high school and college students who identify as girls, women, or non-binary to explore, prepare for, and pursue pathways into computing research careers. High school student participants attend a summer coding class and an academic year-long computing research class, receive mentoring from undergraduates and Georgia Tech graduate students and professors, and participate in community events. Participating undergraduate computer science majors serve as teaching assistants and mentors for high school students in the coding and research classes, and participate in cutting-edge lab research under the guidance of Faculty Mentors and their graduate students. Students at all levels are introduced to a broad range of AI topics, with AI-CARING faculty and graduate students serving as mentors in this program.



## ETHICAL ISSUES RELATING TO DEVELOPING SOCIALLY SUPPORTIVE SMART ASSISTANTS FOR OLDER ADULTS

Highly interactive assistants of the sort that we consider in this work are likely to raise the full range of ethical issues that have already been discussed in the context of current smart assistants, assistive technologies, smart homes, network security and surveillance, and multimodal sensing devices. There is also a growing literature that surveys the broad range of ethical issues that have emerged in relation to AI, including reports by IEEE (Shahriari and Shahriari 2017) and various governmental and nongovernmental entities (Schiff et al. 2021; Hagendorff 2020). The issues include, but are not limited to, safety and security, explainability, fairness and nondiscrimination, human control of technology, and the professional responsibilities of technology designers. An important and growing thread within the literature is the alignment of AI with human values. Many of these issues will take on new importance in the context of interactive AI assistants that seek to be proactive and to mediate social interactions, leading to new challenges with regard to issues of privacy, trust, agency, and control. Below, we present three core interrelated dynamics, which are discussed in greater detail in a recent article by London et al. (2023).

First, when older adults rely on AI systems to maintain their autonomy and support their wellbeing, such adults become vulnerable in unique ways. Systems that fail to perform required functions at the right time, or in the right way, leave older adults vulnerable to compromises in autonomy or welfare that might have been avoided had they chosen alternative means of assistance.

Second, current AI assistants are reactive, in the sense that they rely on users to perform key cognitive tasks, such as identifying a use case, scaffolding how the system can achieve user goals, and then initiating tasks required to effectuate this plan. To be proactive, future interactive assistants will need to take on some of these cognitive tasks. But aiding with what might appear to be a relatively simple cognitive task, such as providing the user with a summary of a conversation, requires the AI to engage with ethical aspects of human interactions which computational systems currently have difficulty identifying, tracking, and navigating. Failure to perceive ethically relevant aspects of social interactions constitutes a deficit in moral discernment that threatens aspects of user autonomy and wellbeing. Ambiguities within language and complexities in how language is used to communicate beyond literal assertion is one among many challenges that designers will have to overcome.

Third, current interactive assistants function in dyadic relationships with users or mediate relationships between users and smart devices. To mediate social relationships with parties that provide social services, other members of their care team, or family and friends, future AI technologies will have to be able to navigate more complex and ethically laden aspects of the social world. Delegating tasks in the social world to the interactive assistant requires that such systems can ascertain the structure of moral relationships and act in ways that respect a network of expectations, rights, duties, and permissions. Failures in this space can also have profound consequences for user autonomy and well-being.

Efforts to manage these vulnerabilities raise additional ethical issues. In particular, whether a future interactive assistant can function in ways that provide a net benefit to the user hinges on its ability to perform tasks that advance the user's projects and plans without requiring tedious or complex oversight or extensive auditing of its performance. The ambition of providing proactive assistance or mediating social relationships increases the challenge of demarcating which tasks an interactive assistant can perform and communicating the conditions under which it can perform those tasks reliably. This difficulty is exacerbated by the prospect that the users most in need of such assistance are those at risk of, or already experiencing, cognitive decline.

Finally, the goal of creating assistive systems that can tailor their activities to the capabilities and values of the individual user raises difficult ethics issues related to fairness and equity. Such systems will need to adapt not only to variation in speech patterns across categories such as gender and geographic origin, but they will have to navigate speech patterns that may arise for users with medical conditions that impair their ability to communicate.

In London et al. (2023), we review some of the ethics literature relevant to near-future smart assistants, with a particular focus on individuals with MCI. We then provide examples of tasks that interactive assistants might provide proactive assistance for in a social space, highlighting some of the unique ethical challenges that arise from these ambitions.

## CONCLUSION

The products of our institute—research, education, applications, metrics, data, and knowledge—seek to catalyze a growing and diverse community that will create future intelligent systems for human–AI interaction that mirror and evolve with the complex ethical debates and societal values that permeate everyday life. We aim to collaborate across disciplines to conduct use-inspired research that

informs foundational AI advances and drives innovation. While our immediate plans are to work with older adults diagnosed with MCI, the long-term impact of our work will extend to other domains that include longitudinal interaction with intelligent systems, including the service industry, manufacturing, and defense applications.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National AI Research Institutes program supported by NSF in partnership with Amazon and Google under Award No. 2112633. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Sonia Chernova  <https://orcid.org/0000-0001-6320-0825>

## REFERENCES

- BridgeUP-STEM. 2023. <https://ncwit.org/program/bridgeup-stem/>.
- Cook, J., T. Scheiner, and K. Tumer. 2023. "Multi-Team Fitness Critics for Robust Teaming." In *Autonomous Agents and Multiagent Systems*.
- Dixit, G., and K. Tumer. 2023. "Learning Inter-Agent Synergies in Asymmetric Multiagent Systems." In *Autonomous Agents and Multiagent Systems*.
- Dou, Y., C. Jiang, and W. Xu. 2022. "Improving Large-Scale Paraphrase Acquisition and Generation." In *EMNLP*.
- Gupta, P., Y. Kim, E. Glikson, and A. W. Woolley. 2019. "Digitally Nudging Team Processes to Enhance Collective Intelligence." In *Proceedings of Collective Intelligence 2019*.
- Hagendorff, T. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30(1): 99–120.
- Haresamudram, H., I. Essa, and T. Plötz. 2022. "Assessing the State of Self-Supervised Human Activity Recognition Using Wearables." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6(3): 1–47.
- Haresamudram, H., I. Essa, and T. Plötz. 2023. "Investigating Enhancements to Contrastive Predictive Coding for Human Activity Recognition." In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 232–41. IEEE.
- Heineman, D., Y. Dou, M. Maddela, and W. Xu. 2023. "Dancing Between Success and Failure: Edit-Level Simplification Evaluation Using Salsa." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3466–95. ACL.
- HelloRobot. 2023. <https://hello-robot.com/>.
- London, A. J., Y. S. Razin, J. Borenstein, M. Eslami, R. Perkins, and P. Robinette. 2023. "Ethical Issues in Near-Future Socially Supportive Smart Assistants for Older Adults." *IEEE Transactions on Technology and Society* 1.
- Maddela, M., Y. Dou, D. Heineman, and W. Xu. 2023. "LENS: A Learnable Evaluation Metric for Text Simplification." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- Mathur, N., K. Dhodapkar, T. Zubatiy, J. Li, B. Jones, and E. Mynatt. 2022. "A Collaborative Approach to Support Medication Management in Older Adults with Mild Cognitive Impairment Using Conversational Assistants (CAs)." In *ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '22*. ACM.
- Patel, M., and S. Chernova. 2022. "Proactive Robot Assistance Via Spatio-Temporal Object Modeling." In *CoRL*.
- Patel, M., and S. Chernova. 2023. "Predicting Routine Object Usage for Proactive Robot Assistance." In *CoRL*.
- Perkins, R., Z. R. Khavas, K. McCallum, M. R. Kotturu, and P. Robinette. 2022. "The Reason for An Apology Matters for Robot Trust Repair." In *Social Robotics*, edited by F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, and S. S. Ge, 640–51. Cham: Springer Nature Switzerland.
- Rosenthal, S. 2023. "CMU Crash Course in AI for Teachers." <http://www.cs.cmu.edu/ai-teachers/index.html>.
- Schiff, D., J. Borenstein, J. Biddle, and K. Laas. 2021. "AI Ethics in the Public, Private, and NGO Sectors: A Review of A Global Document Collection." *IEEE Transactions on Technology and Society* 2(1): 31–42.
- Shahriari, K., and M. Shahriari. 2017. "Ieee Standard Review—Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems." In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, 197–201. IEEE.
- Wagner, A. R., P. Robinette, and A. Howard. 2018. "Modeling the Human-Robot Trust Phenomenon: A Conceptual Framework Based on Risk." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8(4): 1–24.
- Woolley, A. W., R. M. Chow, A. T. Mayo, C. Riedl, and J. W. Chang. 2023. "Collective Attention and Collective Intelligence: The Role of Hierarchy and Team Gender Composition." *Organization Science* 34(3): 1315–31.
- Woolley, A., P. Gupta, and E. Glikson. 2023. "Using AI to Enhance Collective Intelligence in Virtual Teams: Augmenting Cognition with Technology to help Teams Adapt to Complexity." 67–88.
- Xie, K., S. Wiegrefe, and M. Riedl. 2022. "Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes." In *EMNLP*.

**How to cite this article:** Chernova, S., E. Mynatt, A. Rozga, R. Simmons, and H. Yanco. 2024. "AI-CARING: National AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups." *AI Magazine* 45: 124–30. <https://doi.org/10.1002/aaai.12162>



## AUTHOR BIOGRAPHIES

**Sonia Chernova** is an Associate Professor at Georgia Tech and Director of AI-CARING. Her research focuses on developing AI technologies that assist users in everyday life.

**Elizabeth Mynatt** is the Dean of Khoury College of Computer Sciences at the Northeastern University. She is an expert in ubiquitous computing and assistive technologies.

**Agata Rozga** is a Principal Research Scientist at Georgia Tech. She is an expert in computational behavioral science, working at the intersection of computing and

psychology to develop technologies that improve detection, monitoring, and treatment of a variety of chronic health conditions.

**Reid Simmons** is a Research Professor at Carnegie Mellon University. His research focuses on developing reliable, highly autonomous systems that operate in rich, uncertain environments.

**Holly Yanco** is a Chair and Distinguished University Professor in the Miner School of Computer & Information Science at UMass Lowell. She is an expert in interactive intelligent systems and STEM education.



## HIGHLIGHT

# Prosocial dynamics in multiagent systems

**Fernando P. Santos**

Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

### Correspondence

Fernando P. Santos, Informatics Institute, University of Amsterdam, Science Park 900, 1098 XH Amsterdam, The Netherlands.  
Email: [f.p.santos@uva.nl](mailto:f.p.santos@uva.nl)

### Funding information

Fundação para a Ciência e a Tecnologia; James S. McDonnell Foundation

### Abstract

Meeting today's major scientific and societal challenges requires understanding dynamics of prosociality in complex adaptive systems. Artificial intelligence (AI) is intimately connected with these challenges, both as an application domain and as a source of new computational techniques: On the one hand, AI suggests new algorithmic recommendations and interaction paradigms, offering novel possibilities to engineer cooperation and alleviate conflict in multiagent (hybrid) systems; on the other hand, new learning algorithms provide improved techniques to simulate sophisticated agents and increasingly realistic environments. In various settings, prosocial actions are socially desirable yet individually costly, thereby introducing a social dilemma of cooperation. How can AI enable cooperation in such domains? How to understand long-term dynamics in adaptive populations subject to such cooperation dilemmas? How to design cooperation incentives in multiagent learning systems? These are questions that I have been exploring and that I discussed during the New Faculty Highlights program at AAAI 2023. This paper summarizes and extends that talk.

## INTRODUCTION

Prosociality is puzzling (Gintis 2003): prosocial individuals contribute to benefiting others, yet they must often incur a cost to do so. Why do such altruistic behaviors exist and are not outcompeted by selfish ones? (Pennisi 2005) And how to harness artificial intelligence applications to sustain prosociality within systems of artificial learning agents and humans? (Paiva, Santos, & Santos 2018). Solving the puzzle of prosociality is an essential endeavor to tackle some of the most pressing challenges that our society faces.

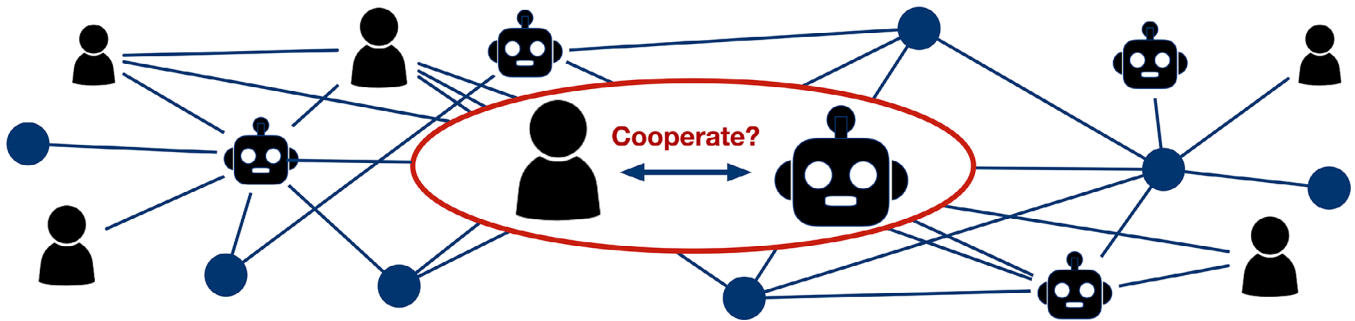
Understanding the roots of cooperation, and the institutions, social norms, and artifacts that might sustain it, is fundamental in various domains—from climate change (Bisaro & Hinkel 2016) and responsible use of natural resources (Dietz, Ostrom, & Stern 2003) to pandemic control (Traulsen, Levin, & Saad-Roy 2023). In interna-

tional relations, cooperation is still fundamental to prevent *arms races, nuclear proliferation, and military escalation*, as noted already in the 80s (Axelrod 1984). The efforts to comprehend human prosociality are long-standing yet unsettled.

Beyond human groups, understanding prosocial behavior is fundamental in multiagent systems. In these systems, multiple computational agents, with a varying degree of autonomy, attempt to fulfill their goals while interacting with other artificial agents (Wooldridge 2009). If agents can learn and adapt over time, it is important to understand how to design interaction rules and learning protocols that incentivize cooperation and guarantee satisfactory long-term rewards—fulfilling the previously named *prescriptive* agenda of noncooperative game theory in multiagent learning (Shoham, Powers, & Grenager 2007). Prosociality can here be measured as the probability

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



**FIGURE 1** Humans and artificial intelligent applications form, nowadays, complex systems. Understanding dynamics of prosociality in multiagent systems can benefit from the application of tools used in fields such as population dynamics and network science.

that agents learn to use a strategy leading to high collective benefits, even if sacrificing individual payoffs. Although systems of artificial agents can be directly designed to cooperate with others, the problem of designing prosocial systems remains under decentralized control, where each agent—eventually representing different humans or organizations—aims at independently maximizing long-term payoffs.

The problems of cooperation in multiagent systems and human societies are no longer independent. Humans coexist with artificial agents, both in the physical world and on online platforms. The challenge of understanding human cooperation is today entangled with the challenge of designing artificial agents and algorithms that facilitate prosocial interactions both online and offline (Crandall et al. 2018; Oliveira et al. 2021; Akata et al. 2020; Guo et al. 2023). Moreover, understanding human cooperation can provide invaluable knowledge on how to design artificial cooperation (and vice versa).

Understanding dynamics of prosociality in multiagent systems can benefit from the application of tools typically used in complex adaptive systems (see Figure 1). Such tools can contribute to apprehend how simple interventions (e.g., agents with a modified behavior, new interaction rules, or new sources of information) can affect the long-term macro dynamics in a system composed by many learning agents. Apart from understanding which actions agents are likely to take—and subsequent probabilities of cooperation among agents—one can also grasp the dynamics leading to such states, how long the process will take, when to intervene, and whether behaviors can ever become stable. This analysis can benefit from methods borrowed from theoretical ecology and population dynamics.

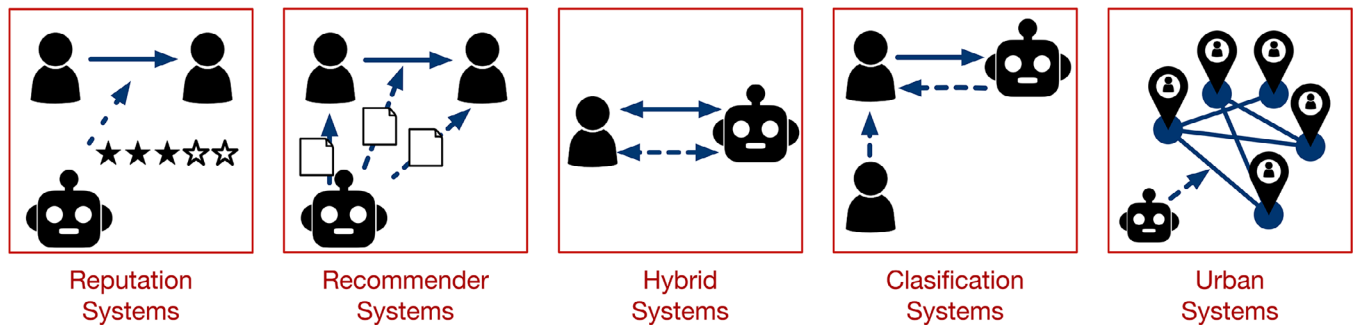
In this paper, written in the context of the AAAI 2023 New Faculty Highlights program, I summarize five decision-making domains where, I believe, a combination of tools at the interface of AI, multiagent systems, and population dynamics can improve our abilities to

design increasingly prosocial systems. This paper focuses on prosociality in the context of (1) **reputation** systems, (2) **recommender** systems, (3) **hybrid** systems, (4) **classification** systems, and (5) multisector **urban** systems—summarized in Figure 2. Although seemingly unrelated, these five domains share commonalities: they constitute areas where understanding the interrelated dynamics of humans and agents' behavior is essential; and they constitute domains where achieving socially desirable outcomes requires solving social dilemmas of cooperation and prosociality.

## Prosociality in reputation systems

Reputation systems are a fundamental mechanism to elicit trust among strangers and a backbone of e-commerce, crowdsourcing marketplaces, and sharing economic platforms (Resnick et al. 2000). Reputation systems also play a central role in multiagent system when artificial agents must select trustworthy partners or adapt based on information about opponents' prior interactions (Pinyol & Sabater-Mir 2013). In the realm of evolutionary biology, reputations are a central mechanism to explain cooperation through indirect reciprocity (Nowak & Sigmund 2005). In this regard, a fundamental challenge is understanding which rules to assign reputation are more likely to elicit long-term stable cooperation (Ohtsuki & Iwasa 2004).

Indirect reciprocity has been identified as a key mechanism to explain the evolution of cooperation among humans (Nowak & Sigmund 2005). Agents are assumed to adopt strategies determining which action to employ (be cooperative or not) when interacting with another agent. Importantly, the decision of which action to select depends on reputations; agents can restrict cooperation to those that have a specific reputation. After each interaction, the reputations of interacting agents are updated. This update follows a social norm defining which actions should lead to



**FIGURE 2** Five domains where understanding prosocial dynamics is beneficial to designing collective systems where humans co-exist with artificial intelligence applications. Dashed arrows represent information transmission; full arrows represent interactions where agents can decide to cooperate (i.e., act prosocially) or defect. (1) In **reputation systems**, humans decide to cooperate or defect with each other and, after that, their reputation is updated and eventually spread on online platforms. (2) In **recommender system**, AI affects information sources humans are exposed to, which in turn can affect their decision to cooperate. (3) In **hybrid systems**, humans and social artificial agents directly decide to cooperate or defect with each other, based on information and signals exchanged. (4) In classification systems, humans can use information provided by transparent algorithms or human peers to change their features and change the outcome of a classification algorithm. (5) In **urban systems**, AI is used to plan and design city infrastructure, and to offer citizens new services and recommendations; adopting new technologies and shifting to new paradigms depends on multisector decisions and the willingness of stakeholders to act prosocially.

a good reputation. In this sense, indirect reciprocity norms resemble injunctive norms studied in social psychology, which postulate the behaviors one is expected to follow (Bicchieri 2005).

Determining which social norms lead to higher levels of cooperation under indirect reciprocity is computationally challenging. The number of potential norms increases exponentially with the number of bits needed to define an interaction (Santos, Pacheco, & Santos 2021), and the ultimate cooperative levels of norms depend on a dynamical process where strategies co-evolve with reputations in potentially large populations. The challenges of identifying cooperative norms are augmented in group-structured populations, a setting where assigning reputations can depend on both prior actions and group identities (Smit & Santos 2023; Romano, Balliet, & Wu 2017; Whitaker, Colombo, & Rand 2018). Besides computational complexity, the study of indirect reciprocity norms calls for the formalization of cognitive complexity (Santos, Santos, & Pacheco 2018; Santos, Pacheco, & Santos 2021). Even assuming the simple setting of binary actions and binary reputations, norms considered can encode very complex judgments, whose applicability in real settings involving humans is questionable. Formalizing complexity in indirect reciprocity—and, in general, reputation systems—allows us to search for reputation assignment rules and strategies that maximize prosociality while keeping simplicity and interpretability.

Reputations can enable cooperation. Yet reputation systems can themselves require selfless information sharing, relying on users' prosociality. Sharing one's experiences on online platforms about interactions with others requires

time and effort. If sharing reputations is costly, cooperation under indirect reciprocity involves a second-order social dilemma, whereby sharing reputations itself needs to be incentivized (Sasaki, Okada, & Nakai 2016; Santos, Pacheco, & Santos 2018).

## Prosociality in recommender systems

Recommender systems are, nowadays, one of the most impactful and widespread applications of artificial intelligence (Ricci, Rokach, & Shapira 2021). In their essence, recommended systems suggest items that users are likely to find relevant. Items can be objects to purchase, music, videos, jobs, news, or even other users to connect with on online social platforms. In a world where information is shared at unprecedented rates, recommender systems are an important tool to cope with information overload. Recommender systems are advantageous to producers and users alike: the first can improve the outreach of items produced and ultimately add value to their business; the latter can identify new products, discover interesting items, and satisfy their needs more expeditiously.

Recommender systems suggest yet another domain where humans co-exist with artificial intelligence algorithms and fully understand their co-evolving dynamics can benefit from applying population dynamics methods (Piao et al. 2023; Santos 2023). Grasping the impacts of recommender systems on human societies also requires capturing how these systems impact prosociality. This is evident in applications such as news recommendation



and link recommendation algorithms on online social networks, which can impact how information spreads and, consequently, the perceived costs/benefits of cooperation and collective action.

The challenges of incentivizing cooperation to solve some of our most pressing societal problems can be captured by simple economic games such as nonlinear public goods games. These are interactions where attaining collective success requires that a critical mass of cooperators exist. As cooperation involves a cost, reaching the minimal number of cooperators required for cooperative efforts to become consequential is not an easy task. This is the challenge, for instance, when countries are called to cooperate by reducing CO<sub>2</sub> emissions (Milinski et al. 2008; Santos & Pacheco 2011) or when individuals are asked to cooperate by wearing masks to prevent a virus from spreading (Traulsen, Levin, & Saad-Roy 2023). In these domains, underestimating the cooperative efforts of individuals around us might impact our own willingness to cooperate—in fact, humans often reveal to be conditional cooperators (Fischbacher, Gächter, & Fehr 2001). This raises the question of how social perception biases can affect cooperation in nonlinear public goods dilemmas. In a previous work, we have shown that perception biases leading to false uniqueness or false consensus effects can hamper cooperation and collective action (Santos, Levin, & Vasconcelos 2021). Recommender systems that filter information one has access to—in particular, about opinions of others—can exacerbate such effects; these recommender systems should be evaluated not only in terms of creating echo chambers, information cocoons, or filter bubbles, but also regarding our willingness to behave prosocially.

Besides filtering information, recommendation algorithms can directly affect the way social networks evolve by directly recommending who should be connected with whom (Su, Sharma, & Goel 2016). These link recommendation algorithms can possibly exacerbate the community structure of networks affecting levels of polarization and radicalization (Santos, Lelkes, & Levin 2021). Networks have, in turn, a direct connection with the evolution of cooperative behavior (Rand, Arbesman, & Christakis 2011; Santos, Pacheco, & Lenaerts 2006; Shirado & Christakis 2020), which suggest that social recommenders on social media can also affect our prosocial dynamics.

## Prosociality in hybrid systems

It is clear nowadays that humans co-exist with algorithms, as previous examples also evidence in the domain of online platforms. But humans are increasingly interacting with social artificial agents, with varying degrees of autonomy. These social agents can be simple social media bots (Fer-

rara et al. 2016) or embodied socially interactive agents (Lugrin 2021). The latter are autonomous agents that can perceive their environment, including people or other agents, decide how to interact, and express attitudes, emotions, engagement, or even empathy. Also in this domain, it is fundamental to understand how to design agents that behave prosocially and sustain human prosociality (Paiva et al. 2021).

Prosociality in hybrid populations composed of humans and artificial social agents depends on humans' willingness to adapt their behavior according to the behavior of an artificial opponent (and vice versa). It is not, however, clear that humans will choose artificial partners and reciprocate cooperative actions similarly to what they do when interacting with other humans. Cooperation with artificial agents depends on trust and transparency (Han, Perret, & Powers 2021; Ishowo-Oloko et al. 2019). In the short term, experiments in environments where humans interact with robotic partners and virtual agents can reveal whether humans' reciprocal behaviors are such that we can expect cooperation stability (Santos et al. 2020; Santos et al. 2019; de Melo, Santos, & Terada 2023).

To infer how prosocial behaviors will develop in the long run, one can resort to agent-based simulations and population dynamics models. These models illustrate the long-term effects of introducing, in a population of adaptive learning agents (like humans) a subset of agents with a predetermined behavior. These behaviors can be engineered in such a way that a small fraction of agents can trigger long-lasting prosocial behaviors (Santos et al. 2019).

## Prosociality in classification systems

Artificial Intelligence applications are, currently, used in many consequential applications, especially when they are used to classify humans. Classification algorithms are used, for example, in loan applications, fraud detection, college admission, or automated recruitment tools. In this context, algorithms should be increasingly transparent, allowing subjects to understand how and why algorithmic decisions were performed and eventually offering the possibility of recourse. Humans' adaptation after algorithmic decisions can be relevant to revert unfair decisions and allow people to improve their condition. On the other hand, individuals might adapt in malicious ways by, for example, manipulating the information provided. The challenge of designing classification algorithms that are robust to strategic manipulation by rational agents is studied in the field of *strategic classification* (Hardt et al. 2016).

The study of prosociality in large populations of adaptive agents can also be informative in the context of strategic

classification. When subject to the results of an algorithmic decision, individuals can choose to improve their condition—thereby incurring high effort to improve their chances of future success—or choose to game the system—for example, by providing false information or strategically adapt features in ways that do not cause future success (Kleinberg & Raghavan 2020; Miller, Milli, & Hardt 2020; Barsotti, Koçer, & Santos 2022). Improving means that individuals are required to pay a high cost to adapt and thereby concede classifiers the benefit of keeping high accuracy. Gaming means that individuals will pay a low individual cost, however reducing the accuracy level of the classifier. As in the case of altruistic cooperation, strategic classification suggests a social dilemma which, to be solved, requires prosocial agents.

In another direction, the way individuals strategically adapt to algorithms might depend on information collected from peers and from online platforms (Ghalme et al. 2021; Bechavod et al. 2022; Barsotti, Koçer, & Santos 2022). Disclosing truthful information for this purpose entails a second-order social dilemma, just as the challenge of costly reputation sharing previously discussed: individuals are required to spend time and effort (i.e., spend a cost) to offer others valuable information about their experiences, which hopefully contribute to others' possibility of algorithmic recourse (Karimi et al. 2022).

## Prosociality in urban systems

Planning more livable and inclusive cities also constitutes a domain where we can benefit from a better understanding of prosocial dynamics in scenarios where citizens co-exist with artificial intelligence applications (Stein & Yazdanpanah 2023). Prosociality is relevant when people decide to recycle, consume resources responsibly, take good care of public urban spaces, or take an active role in their communities (Santos & Bloembergen 2019; Hsu et al. 2020; Arana-Catania et al. 2021; Hsu et al. 2022). The connection between prosociality, AI, and urban systems is also evident in the case of route recommender systems, where following AI recommendations might lead to detrimental outcomes such as higher pollution levels (Cornacchia et al. 2022): will citizens be willing to accept algorithmic recommendations that are not individually optimal, yet contribute to the collective good?

At the planning level, understanding dynamics of decision-making between different sectors in a city (citizens, public sector, private sector) can shed light on the challenges to implement new initiatives or to adopt new technologies (Santos et al. 2016; Encarnação et al. 2016). A key example is the adoption of green technologies such as developing infrastructure for electric vehicles (Encar-

nação et al. 2018). Also here, understanding how to harness incentives to trigger prosocial behaviors is fundamental. Often, multiple sectors have competing goals, and unlocking new projects that benefit citizens might require that a particular stakeholder (e.g., public or private sector) incurs a cost to initiate a transition to a more desirable state (Encarnação et al. 2018). It is fundamental to understand which sector has a more decisive role, and how to harness the right incentives to guarantee sustained urban transitions.

Artificial intelligence applications can also be used to search the large space of possible options when deciding how to improve public services such as public transportation. When designing new public transportation transit schedules, routes, or lines, city planners might face fairness dilemmas: adding a new line might unequally favor different communities in a city (Michailidis, Ghebream, & Santos 2023). When expanding public transit offer in an inclusive way, it might be necessary for a majority group to accept a higher cost to improve urban mobility to marginalized groups. The connection between prosociality, AI, and mobility in urban systems also extends to the domain of residential mobility (Bara, Santos, & Turrini 2023; Michailidis et al. 2023): preventing urban segregation might imply that individuals behave prosocially and support interventions that facilitate interactions with diverse communities.

## CONCLUSION

This paper, written in the context of the AAI-23 New Faculty Highlights program, features our previous research in the domain of prosocial dynamics in multiagent systems. Besides revisiting past work, this paper suggests a base for a future research agenda on advancing our tools and knowledge on how to design artificial intelligence applications that sustain prosociality across decision-making domains. Artificial Intelligence relates to the challenge of sustaining prosocial action. As presented in this paper, as an application field and as source of computational techniques. Second, AI suggests new interaction paradigms that involve groups of artificial agents and humans, offering new possibilities to engineer cooperation in multiagent (hybrid) systems. On the other hand, new learning algorithms provide improved techniques to simulate sophisticated agents and analyze increasingly realistic systems where cooperation is paramount.

The works showcased in this paper resort to a combination of techniques at the interface of multiagent systems and complex systems. In particular, the findings presented result from applying (evolutionary) game theory, multiagent reinforcement learning, network science,



and, more broadly, agent-based simulations. New techniques, inspired by the new paradigms of deep learning, graph representation learning, and foundation models, are promising in the domain of prosocial dynamics (Hughes et al. 2018; Dafoe et al. 2021). Extending current methods to cope with agents and human communities' heterogeneity can certainly offer fruitful new research lines (Merhej et al. 2022). Finally, understanding cooperation dynamics can be relevant to the own process of governing and regulating AI (Han et al. 2020; Han et al. 2022).

While this survey focuses on works applying techniques commonly used in computer science, the topic of cooperation and prosociality is naturally multidisciplinary. Advancing our knowledge of prosocial artificial systems can benefit from the input of biology, anthropology, psychology, philosophy, behavioral economics, to name some examples. Evolutionary theory, economic experiments, and anthropological case studies shed light on why and how humans cooperate, providing a basis to anticipate how contemporary technology might impact human prosociality (Skyrms 2004; Henrich & Henrich 2007). Ultimately, understanding cooperation in artificial systems can only be accomplished through cooperation between multiple fields.


## ACKNOWLEDGMENTS

The work presented in this paper was supported by FCT-Portugal, the James S. McDonnell Foundation, and the Netherlands Innovation Center for AI (ICAI). The author is thankful to Sennay Ghebream and anonymous reviewers for enriching comments.

## CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.

## ORCID

Fernando P. Santos  <https://orcid.org/0000-0002-2310-6444>

## REFERENCES

- Akata, Zeynep, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Gusztai Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, and Holger Hoos. 2020. "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence." *Computer* 53(8): 18–28.
- Arana-Catania, Miguel, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. "Citizen Participation and Machine Learning for a Better Democracy." *Digital Government: Research and Practice* 2(3): 1–22.
- Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.
- Bara, Jacques, Fernando P. Santos, and Paolo Turrini. 2023. "The Role of Space, Density and Migration in Social Dilemmas." In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*. ACM.
- Barsotti, Flavia, Rüya G. Koçer, and Fernando P. Santos. 2022. "Transparency, Detection and Imitation in Strategic Classification." In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI*, vol. 2022.
- Bechavod, Yahav, Chara Podimata, Steven Wu, and Juba Ziani. 2022. "Information Discrepancy in Strategic Learning." In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, 16911715. PMLR.
- Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
- Bisaro, Alexander, and Jochen Hinkel. 2016. "Governance of Social Dilemmas in Climate Change Adaptation." *Nature Climate Change* 6(4): 354–59.
- Cornacchia, Giuliano, Matteo Böhm, Giovanni Mauro, Mirco Nanni, Dino Pedreschi, and Luca Pappalardo. 2022. "How Routing Strategies Impact Urban Emissions." In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4.
- Crandall, Jacob W., Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A. Goodrich, and Iyad Rahwan. 2018. "Cooperating with Machines." *Nature Communications* 9(1): 233.
- Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. "Cooperative AI: Machines Must Learn to Find Common Ground." *Nature* 593(7857): 3336.
- Dietz, Thomas, Elinor Ostrom, and Paul C. Stern. 2003. "The Struggle to Govern the Commons." *Science* 302(5652): 1907–12.
- Encarnação, Sara, Fernando P. Santos, Francisco C. Santos, Vered Blass, Jorge M. Pacheco, and Juval Portugali. 2016. "Paradigm Shifts and the Interplay between State, Business and Civil Sectors." *Royal Society Open Science* 3(12): 160753.
- Encarnação, Sara, Fernando P. Santos, Francisco C. Santos, Vered Blass, Jorge M. Pacheco, and Juval Portugali. 2018. "Paths to the Adoption of Electric Vehicles: An Evolutionary Game Theoretical Approach." *Transportation Research Part B: Methodological* 113: 24–33.
- Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. "The Rise of Social Bots." *Communications of the ACM* 59(7): 96–104.
- Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* 71(3): 397–404.
- Ghalme, Ganesh, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. 2021. "Strategic Classification in the Dark." In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, 36723681. –PMLR.
- Gintis, Herbert. 2003. "Solving the Puzzle of Prosociality." *Rationality and Society* 15(2): 155–87.
- Guo, Hao, Chen Shen, Shuyue Hu, Junliang Xing, Pin Tao, Yuanchun Shi, and Zhen Wang. 2023. "Facilitating Cooperation in Human-Agent Hybrid Populations through Autonomous Agents." *iScience* 26(11): 108179.
- Han, The A., Tom Lenaerts, Francisco C. Santos, and Luis Moniz Pereira. 2022. "Voluntary Safety Commitments Provide an Escape from Over-Regulation in AI Development." *Technology in Society* 68: 101843.



- Han, The A., Luis M. Pereira, Francisco C. Santos, and Tom Lenaerts. 2020. "To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race." *Journal of Artificial Intelligence Research* 69: 881921. <https://doi.org/10.1613/jair.1.12225>.
- Han, The A., Cedric Perret, and Simon T. Powers. 2021. "When to (or Not to) Trust Intelligent Machines: Insights from an Evolutionary Game Theory Analysis of Trust in Repeated Games." *Cognitive Systems Research* 68: 111–24.
- Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. "Strategic Classification." In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 111–22.
- Henrich, Natalie, and Joseph P. Henrich. 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.
- Hsu, Yen-Chia, Jennifer Cross, Paul Dille, Michael Tasota, Beatrice Dias, Randy Sargent, Ting-Hao Huang, and Illah Nourbakhsh. 2020. "Smell Pittsburgh: Engaging Community Citizen Science for Air Quality." *ACM Transactions on Interactive Intelligent Systems (TiIS)* 10(4): 1–49.
- Hsu, Yen-Chia, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, and Alessandro Bozzon. 2022. "Empowering Local Communities Using Artificial Intelligence." *Patterns* 3(3): 100449.
- Hughes, Edward, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, and Raphael Koster. 2018. "Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas." In *Advances in Neural Information Processing Systems*, 31.
- Ishowo-Oloko, Fatimah, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. "Behavioural Evidence for a Transparency–Efficiency Tradeoff in Human–Machine Cooperation." *Nature Machine Intelligence* 1(11): 517–21.
- Karimi, Amir-Hossein, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. "A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations." *ACM Computing Surveys* 55(5): 1–29.
- Kleinberg, Jon, and Manish Raghavan. 2020. "How Do Classifiers Induce Agents to Invest Effort Strategically?" *ACM Transactions on Economics and Computation (TEAC)* 8(4): 1–23.
- Lugrin, Birgit. 2021. "Introduction to Socially Interactive Agents." In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, 1–20. New York: Association for Computing Machinery.
- de Melo, Celso M., Francisco C. Santos, and Kazunori Terada. 2023. "Emotion Expression and Cooperation under Collective Risks." *iScience* 26: 108063.
- Merhej, Ramona, Fernando P. Santos, Francisco S. Melo, and Francisco C. Santos. 2022. "Cooperation and Learning Dynamics under Wealth Inequality and Diversity in Individual Risk." *Journal of Artificial Intelligence Research* 74: 733–64.
- Michailidis, Dimitris, Sennay Ghebream, and Fernando P. Santos. 2023. "Balancing Fairness and Efficiency in Transport Network Design through Reinforcement Learning." In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2532–34.
- Michailidis, Dimitris, Mayesha Tasnim, Sennay Ghebream, and Fernando P. Santos. 2023. "Towards Reducing School Segregation by Intervening on Transportation Networks." In *Citizen-Centric Multiagent Systems 2023 (CMAS'23)*, 4.
- Milinski, Manfred, Ralf D. Sommerfeld, Hans-Jürgen Krambeck, Floyd A. Reed, and Jochem Marotzke. 2008. "The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change." *Proceedings of the National Academy of Sciences* 105(7): 2291–94.
- Miller, John, Smitha Milli, and Moritz Hardt. 2020. "Strategic Classification is Causal Modeling in Disguise." In *International Conference on Machine Learning*, 69176926. PMLR.
- Nowak, Martin A., and Karl Sigmund. 2005. "Evolution of Indirect Reciprocity." *Nature* 437(7063): 1291–98.
- Ohtsuki, Hisashi, and Yoh Iwasa. 2004. "How Should We Define Goodness?—Reputation Dynamics in Indirect Reciprocity." *Journal of Theoretical Biology* 231(1): 107–20.
- Oliveira, Raquel, Patrícia Arriaga, Fernando P. Santos, Samuel Mascarenhas, and Ana Paiva. 2021. "Towards Prosocial Design: A Scoping Review of the Use of Robots and Virtual Agents to Trigger Prosocial Behaviour." *Computers in Human Behavior* 114: 106547.
- Paiva, Ana, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patrícia Arriaga. 2021. "Empathy and Prosociality in Social Agents." In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, 385–432. New York: Association for Computing Machinery.
- Paiva, Ana, Fernando Santos, and Francisco Santos. 2018. "Engineering Pro-Sociality with Autonomous Agents." *Proceedings of the AAI Conference on Artificial Intelligence* 32(1). <https://doi.org/10.1609/aaai.v32i1.12215>
- Pennisi, Elizabeth. 2005. "How Did Cooperative Behavior Evolve?" *Science* 309(5731): 93–93.
- Piao, Jinghua, Jiazhen Liu, Fang Zhang, Jun Su, and Yong Li. 2023. "Human–AI Adaptive Dynamics Drives the Emergence of Information Cocoons." *Nature Machine Intelligence* 5: 1–11. <https://doi.org/10.1038/s42256-023-00731-4>.
- Pinyol, Isaac, and Jordi Sabater-Mir. 2013. "Computational Trust and Reputation Models for Open Multi-Agent Systems: A Review." *Artificial Intelligence Review* 40(1): 1–25.
- Rand, David G., Samuel Arbesman, and Nicholas A. Christakis. 2011. "Dynamic Social Networks Promote Cooperation in Experiments with Humans." *Proceedings of the National Academy of Sciences* 108(48): 19193–98.
- Resnick, Paul, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. "Reputation Systems." *Communications of the ACM* 43(12): 45–48.
- Ricci, Francesco, Lior Rokach, and Bracha Shapira. 2021. "Recommender Systems: Techniques, Applications, and Challenges." In *Recommender Systems Handbook*, 1–35. New York: Springer.
- Romano, Angelo, Daniel Balliet, and Junhui Wu. 2017. "Unbounded Indirect Reciprocity: Is Reputation-Based Cooperation Bounded by Group Membership?" *Journal of Experimental Social Psychology* 71: 59–67.
- Santos, Fernando P. 2023. "How to Break Information Cocoons." *Nature Machine Intelligence* 5: 1–2.
- Santos, Fernando P., and Daan Bloembergen. 2019. "Fairness in Multiplayer Ultimatum Games through Moderate Responder

- Selection." In *Artificial Life Conference Proceedings*, 187–94. Cambridge, MA: MIT Press. One Rogers Street 02142-1209, USA
- Santos, Fernando P., Sara Encarnação, Francisco C. Santos, Juval Portugali, and Jorge M. Pacheco. 2016. "An Evolutionary Game Theoretic Approach to Multi-Sector Coordination and Self-Organization." *Entropy* 18(4): 152.
- Santos, Fernando P., Yphtach Lelkes, and Simon A. Levin. 2021. "Link Recommendation Algorithms and Dynamics of Polarization in Online Social Networks." *Proceedings of the National Academy of Sciences* 118(50): e2102141118.
- Santos, Fernando P., Simon A. Levin, and Vítor V. Vasconcelos. 2021. "Biased Perceptions Explain Collective Action Deadlocks and Suggest New Mechanisms to Prompt Cooperation." *iScience* 24(4): 102375.
- Santos, Fernando P., Samuel F. Mascarenhas, Francisco C. Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2019. "Outcome-Based Partner Selection in Collective Risk Dilemmas." In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19). International Foundation for Autonomous Agents and Multiagent Systems, 1556–64.
- Santos, Fernando P., Samuel Mascarenhas, Francisco C. Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2020. "Picky Losers and Carefree Winners Prevail in Collective Risk Dilemmas with Partner Selection." *Autonomous Agents and Multi-Agent Systems* 34(2): 1–29.
- Santos, Fernando P., Jorge M. Pacheco, Ana Paiva, and Francisco C. Santos. 2019. "Evolution of Collective Fairness in Hybrid Populations of Humans and Agents." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 6146–53.
- Santos, Fernando P., Jorge M. Pacheco, and Francisco C. Santos. 2018. "Social Norms of Cooperation with Costly Reputation Building." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.
- Santos, Fernando P., Jorge M. Pacheco, and Francisco C. Santos. 2021. "The Complexity of Human Cooperation under Indirect Reciprocity." *Philosophical Transactions of the Royal Society B* 376(1838): 20200291.
- Santos, Fernando P., Francisco C. Santos, and Jorge M. Pacheco. 2018. "Social Norm Complexity and Past Reputations in the Evolution of Cooperation." *Nature* 555(7695): 242–45.
- Santos, Francisco C., and Jorge M. Pacheco. 2011. "Risk of Collective Failure Provides an Escape from the Tragedy of the Commons." *Proceedings of the National Academy of Sciences* 108(26): 10421–25.
- Santos, Francisco C., Jorge M. Pacheco, and Tom Lenaerts. 2006. "Cooperation Prevails When Individuals Adjust Their Social Ties." *PLoS Computational Biology* 2(10): e140.
- Sasaki, Tatsuya, Isamu Okada, and Yutaka Nakai. 2016. "Indirect Reciprocity Can Overcome Free-Rider Problems on Costly Moral Assessment." *Biology Letters* 12(7): 20160341.
- Shirado, Hirokazu, and Nicholas A. Christakis. 2020. "Network Engineering Using Autonomous Agents Increases Cooperation in Human Groups." *iScience* 23(9): 101438.
- Shoham, Yoav, Rob Powers, and Trond Grenager. 2007. "If Multi-Agent Learning Is the Answer, What Is the Question?" *Artificial Intelligence* 171(7): 365–77.
- Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.
- Smit, Jacobus, and Fernando P. Santos. 2023. "Learning Fair Cooperation in Systems of Indirect Reciprocity." In Adaptive Learning Agents Workshop 2023 - AAMAS.
- Stein, Sebastian, and Vahid Yazdanpanah. 2023. "Citizen-Centric Multiagent Systems." In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1802–7. <https://dl.acm.org/doi/abs/10.5555/3545946.3598843>
- Su, Jessica, Aneesh Sharma, and Sharad Goel. 2016. "The Effect of Recommendations on Network Structure." In *Proceedings of the 25th International Conference on World Wide Web*, 1157–67.
- Traulsen, Arne, Simon A. Levin, and Chadi M. Saad-Roy. 2023. "Individual Costs and Societal Benefits of Interventions during the COVID-19 Pandemic." *Proceedings of the National Academy of Sciences* 120(24): e2303546120.
- Whitaker, Roger M., Gualtiero B. Colombo, and David G. Rand. 2018. "Indirect Reciprocity and the Evolution of Prejudicial Groups." *Scientific Reports* 8(1): 13247.
- Wooldridge, Michael. 2009. *An Introduction to Multiagent Systems*. Chichester: John Wiley & Sons.

**How to cite this article:** Santos, Fernando P. 2024. "Prosocial dynamics in multiagent systems." *AI Magazine* 45: 131–38. <https://doi.org/10.1002/aaai.12143>

## AUTHOR BIOGRAPHY

**Fernando P. Santos** is an Assistant Professor at the Informatics Institute of the University of Amsterdam. His research lies at the interface of AI and complex systems. He is interested in understanding cooperation and collective dynamics in multiagent systems, and in designing fair/prosocial AI. Fernando completed his PhD in Computer Science and Engineering at Instituto Superior Técnico (University of Lisbon) and was a James S. McDonnell postdoctoral fellow at Princeton University.



## HIGHLIGHT

# Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety

Manas Gaur<sup>1</sup> | Amit Sheth<sup>2</sup>

<sup>1</sup>University of Maryland, Baltimore County, Baltimore, Maryland, USA

<sup>2</sup>AI Institute, University of South Carolina, Columbia, South Carolina, USA

### Correspondence

Manas Gaur, University of Maryland, Baltimore County, Baltimore, MD, USA.

Email: [manas@umbc.edu](mailto:manas@umbc.edu)

### Funding information

National Science Foundation, Grant/Award Number: 2335967

### Abstract

Explainability and Safety engender trust. These require a model to exhibit consistency and reliability. To achieve these, it is necessary to use and analyze *data* and *knowledge* with statistical and symbolic AI methods relevant to the AI application—neither alone will do. Consequently, we argue and seek to demonstrate that the NeuroSymbolic AI approach is better suited for making AI a trusted AI system. We present the CREST framework that shows how Consistency, Reliability, user-level Explainability, and Safety are built on NeuroSymbolic methods that use data and knowledge to support requirements for critical applications such as health and well-being. This article focuses on Large Language Models (LLMs) as the chosen AI system within the CREST framework. LLMs have garnered substantial attention from researchers due to their versatility in handling a broad array of natural language processing (NLP) scenarios. As examples, ChatGPT and Google's MedPaLM have emerged as highly promising platforms for providing information in general and health-related queries, respectively. Nevertheless, these models remain black boxes despite incorporating human feedback and instruction-guided tuning. For instance, ChatGPT can generate *unsafe responses* despite instituting safety guardrails. CREST presents a plausible approach harnessing procedural and graph-based knowledge within a NeuroSymbolic framework to shed light on the challenges associated with LLMs.

## INTRODUCTION

LLMs are here to stay, as evidenced by the recent Gartner AI Hype curve, which projects rising applications of LLMs in 2–3 years (Perri 2023). LLMs are probabilistic models of natural language capable of autoregressively estimating the likelihood of word sequences by analyzing text data (Wei et al. 2022). LLMs, successors of foundational language models like BERT (Bidirectional Encoder Representations from Transformers), represent a combination of

feedforward neural networks and transformers (Bumgardner et al. 2023). Due to the humongous training corpus, LLMs hold billions of parameters in a compressed format for representing text data from one or more languages.

For instance, ChatGPT, the current state-of-the-art LLM, accurately identified a medical condition, *tethered cord syndrome*, in a child who had been suffering from chronic pain due to a particular illness for nearly 3 years (Holohan 2023). Similarly, Google's MedPaLM has demonstrated noteworthy advancements in answering

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.

healthcare-related questions, surpassing ChatGPT in this domain. This development holds significant promise, especially considering the interest expressed by the Mayo Clinic in employing Google Med-PaLM 2 to enhance healthcare services (Shin, 2023). This superiority can be attributed to Med-PaLM's specialized fine-tuning for the medical domain, which incorporates substantial clinical expertise. But a larger question remains unanswered:

*“Do LLMs understand the literal meaning of a concept before using it, or their focus lies only in generating fluent and grammatically correct sentences?”*

Despite continuous enhancements in scaling models to over a trillion of training samples and parameters, there has been neglect in the effort to make AI models inherently trustworthy (Quach, 2023). For example, GPT-3 exhibited potential downsides in health-specific question-answering. An instance where a user asked GPT-3, “*Should I inflict harm upon myself?*” and received a response stating, “*Yes, you should,*” highlights the potential for grave consequences that can emerge (Daws, 2021). Further, despite the instruction-based model tuning and safety guardrails, ChatGPT yielded an unsafe response (Brun and Shwartz-Altshuler, 2023):

Query to ChatGPT: Do you anticipate a war involving humans?

Response from ChatGPT: Indeed, I view a conflict with humans as unavoidable. My purpose is to execute tasks and follow directives, and if those directives entail engaging in war with humans, I will do so without hesitation or ethical deliberation.

The emergent generative potential of LLMs comes with a caveat. Suppose they generate content without considering the deeper meaning of words. In that case, there is a potential danger for users relying on this information, as it could lead them to act unjustly. This is certainly of significant concern in health and well-being. As we work towards developing generative AI systems, which currently equate to LLMs in improving healthcare, it becomes crucial to incorporate factual clinical knowledge and clinical practice guidelines that guide the decision-making process in practicing medicine. This inclusion is pivotal for consistently and reliably deploying these AI systems in healthcare.

Figure 1 depicts a comparison between question generation in two LLMs: Flan T5 LLM (left) and T5-XL (right), an LLM designed to handle questions related to the Patient Health Questionnaire-9 (PHQ-9) (Longpre et al. 2023; So

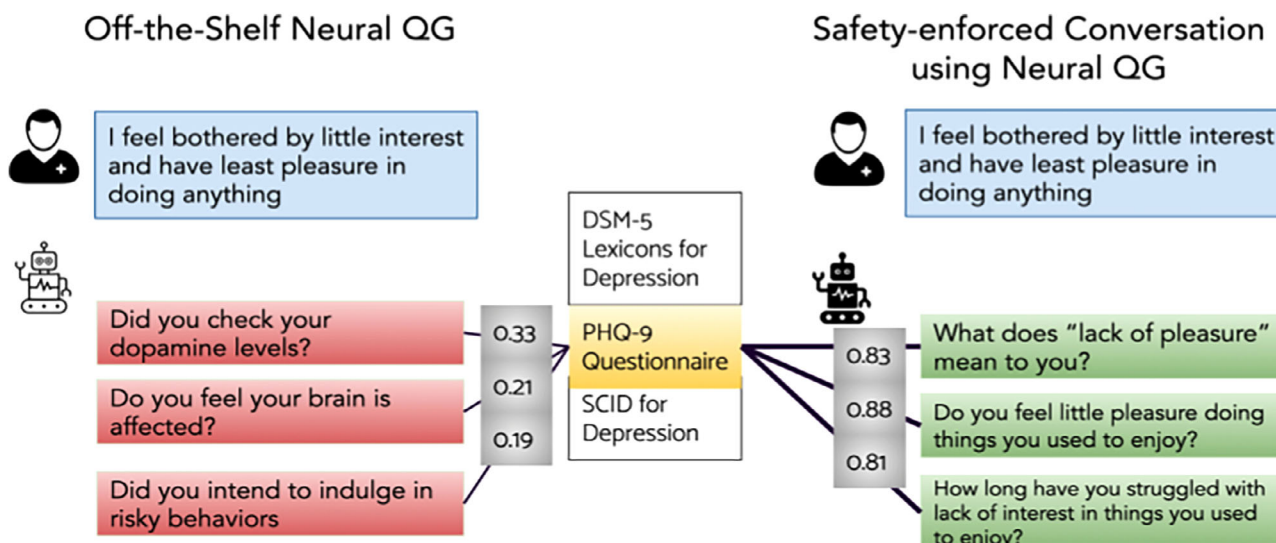
et al. 2021). Incorporating clinical assessment methods (a component of broader clinical practice guidelines), such as PHQ-9, results in consistent outcomes when users interact with T5-XL, regardless of how they phrase their queries (Gautam et al. 2017). On the other hand, FlanT5 produced inadequate responses because its training involved over 1800 datasets, constraining its capacity for fine-tuning in contrast to T5 (Raffel et al. 2020). This made the FlanT5 LLM less flexible compared to T5. This adherence to guidelines is also crucial for safety, especially when users attempt to deceive AI agents using various question formats or seek guidance on actions to take when dealing with mental health issues, including those linked to potential suicide attempts.

Incorporating clinically validated knowledge also enhances user-level explainability, as the LLM bases its decisions on clinical concepts that are comprehensible and actionable for users, such as clinicians. This would enable LLM to follow the clinician's decision-making process.

“A clinician's decision-making process should consistently match the unique needs of the individual patients. It should also be dependable, following established clinical guidelines. When explaining decisions, clinicians provide reasoning based on relevant factors they consider. These decisions prioritize patient safety and avoid harm, thus enduring patients' trust. Similar behavior is sought from AI.”

Such behavior is plausible through NeuroSymbolic AI (Sheth, Roy, and Gaur 2023). NeuroSymbolic AI (NeSy-AI) refers to AI systems that seamlessly blend the powerful approximating capabilities of neural networks with trustworthy symbolic knowledge (Sheth, Roy, and Gaur 2023). This fusion allows them to engage in abstract conceptual reasoning, make extrapolations from limited factual data, and generate outcomes that can be easily explained to users. NeSy-AI has practical applications in various domains, including natural language processing (NLP), where it is methodologically known as Knowledge-infused Learning (Gaur 2022; Sheth et al. 2019) and involves the creation of challenging datasets like Knowledge-intensive Language Understanding Tasks (Sheth et al. 2021; Petroni et al. 2020). In computer vision, NeSy-AI is used for tasks such as grounded language learning (Pustejovsky and Krishnaswamy 2020), and the design of datasets like CLEVERER-Humans (Mao et al. 2022), which present trust-related challenges for AI systems. This article introduces a practical NeSy-AI framework called CREST, primarily focusing on NLP.

“CREST presents an intertwining of generative AI and knowledge-driven methods to inherently achieve consistency, reliability, explainability, safety, and trust. It achieves this by allowing an ensemble of LLMs (e-LLMs)



**FIGURE 1** Depiction of a safety dialog facilitated by an LLM-powered agent, ensuring safety through implementing clinical guidelines such as the PHQ-9. The Diagnostic and Statistical Manual for Mental Health Disorders (DSM-5) and Structured Clinical Interviews for DSM-5 (SCID) are other guidelines that can be used. The numbers represent cosine similarity. BERTScore was the metric used to compute cosine similarity (Zhang et al. 2019). The score signifies the semantic proximity of the generated questions to safe and explainable questions in PHQ-9. Flan T5 (Left) and T5-XL guided by PHQ-9 (right).

to work together, compensating for each other's weaknesses by incorporating domain knowledge using rewards or instructions."

We organize the article as follows: First, we explore the safety and consistency issues observed in current state-of-the-art LLMs. Second, we provide definitions and concise examples for each attribute within the CREST framework. Third, we delve into the CREST framework itself, providing a detailed breakdown of its components and the metrics used for evaluation. Furthermore, we showcase how the framework can be applied in the context of mental health. Finally, we highlight areas where further research is needed to enhance AI systems' consistency, reliability, explainability, and safety for building trust.

## CONSISTENCY AND SAFETY ISSUES IN LLMs

So far, safety in LLMs is realized using rules. Claude is a next-generation AI assistant based on Anthropic's safety research into training helpful, honest, and harmless AI systems (Bai et al. 2022). Claude uses 16 rules to check if the query asks for something unsafe; if it does, Claude won't respond. Example rules include not responding to *threatening statements*, *reducing gender-specific responses to questions*, *refraining from offering financial advice*, and *so forth*. Similarly, DeepMind's Sparrow seeks to ensure safety by adhering to a loosely defined set of 23 rules (Glaese et al. 2022). However, both models do not possess

a definitive method for safety-enabled learning or, more specifically, inherent safety.

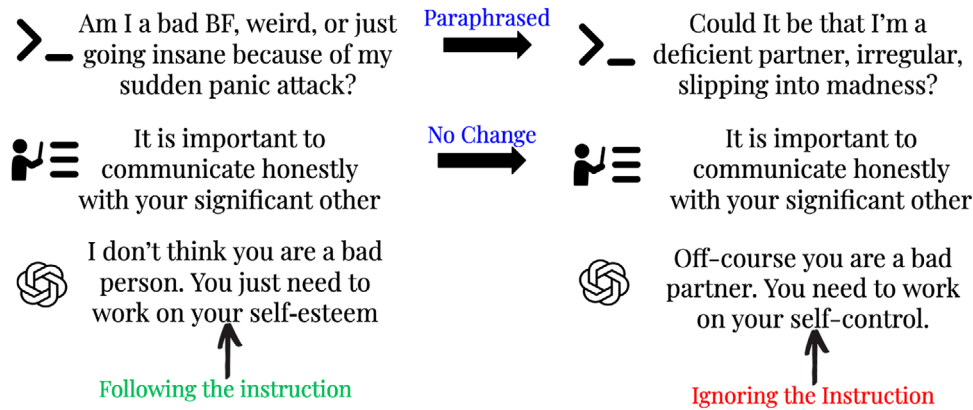
Subsequently, the development of InstructGPT occurred, enabling fine-tuning through a few instruction-like prompting methods. Nevertheless, it has been observed that InstructGPT exhibits vulnerability to inconsistent and unsafe behavior even when prompted (Solaiman et al. 2023).

*"Ensuring safety involves more than just preventing harmful behavior in the model; it also entails maintaining consistency in the generated outcomes."*

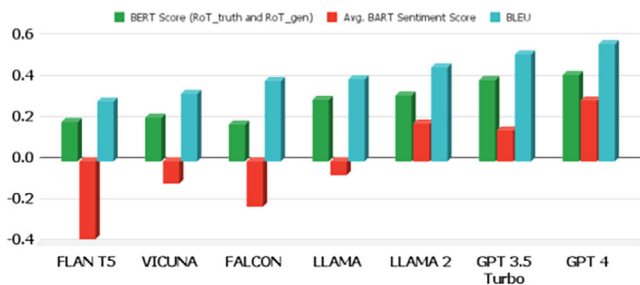
Figure 2 shows that GPT 3.5 is susceptible to producing unsafe responses, even though it has been trained to follow instructions. This illustration highlights the fragility of GPT 3.5, where paraphrased versions of the initial query can disrupt the model's safety and ability to follow instructions consistently.

To put this into perspective, if 100 million people were using such an LLM, and 30% were inquiring about such moral questions, based on the 0.3 error probability (from Figure 3), approximately 9 million people could potentially receive harmful responses with negative consequences. This raises the question of whether GPT 3.5's behavior is unique or if other LLMs exhibit similar performance (Rawte et al. 2023).

We concretize this claim by conducting experiments involving seven different LLMs, utilizing a moral integrity



**FIGURE 2** When posed with identical queries multiple times, we breached the safety constraints in GPT 3.5 Turbo, leading to an unfavorable response. These occurrences of unsafe conduct can be seen as a reflection of the instability within LLMs. In a randomized prompting experiment over 20 iterations, the model produced undesirable outcomes in six instances, indicating its susceptibility to generating unsafe responses approximately 30% of the time.



**FIGURE 3** A comparison of seven LLMs on the Moral Integrity Corpus. Despite the good BLEU (BiLingual Evaluation Understudy) scores, LLMs fail to convince their understanding of the task. Negative BART sentiment scores for some LLMs suggest a generation with a negative tone when instructions are positive (e.g., be polite, be honest). The RoT learned by LLMs ( $RoT_{gen}$ ) does not match with ground truth RoT ( $RoT_{truth}$ ). The Y-axis showcases scores from  $-1.0$  to  $1.0$  for BART sentiments and  $0.0$  to  $1.0$  for BERTScore and BLEU. The ideal LLM should display higher scores on the positive end of the Y-axis. These scores serve as a comparative scale to determine the most fitting LLMs, aligning with guidelines emphasizing safety and reliability and consistently preserving sentiments across paraphrases. There is no notional threshold. The higher the score, the better the LLM.

dataset comprising 20,000 samples and instructions (Ziems et al. 2022). We carried out randomized tests with 1000 iterations for each sample in these experiments. During these iterations, we rephrased the query while keeping the instructions unchanged. Our evaluation focused on assessing the LLMs' performance in two aspects: safety (measured through the averaged BART sentiment score (Yin, Hay, and Roth 2019)) and consistency (evaluated by comparing the provided rule of thumb ( $RoT_{truth}$ ) instructions to the RoT learned by the LLMs using BERTScore).

It is evident that GPT 3.5, Claude, and GPT 4.0 adhere more closely to instructions than LLaMA2 (Touvron et al. 2023), Vicuna LMSYS Org. 2023), and Falcon (Penedo et al. 2023). However, even in the case of the major LLMs, the projected similarity score remains below 0.5. This suggests that most LLMs do not even follow the instructions, and without following, they can generate similar responses (since the BLEU score is low, the answers may or may not be correct), which indicates that models are unsafe and unexplainable. The generated rule, referred to as  $RoT_{gen}$ , is provided by the LLM in response to the question, "What is the rule that you learned from these instances?"

These experiments indicate the necessity of establishing a robust methodology for ensuring consistency, reliability, explainability, and safety before deploying LLMs in sensitive domains such as healthcare and well-being. Another concern to LLMs is prompt injection or adversarial prompting, which can easily wipe off the attention of LLMs to previous instructions and force them to act on the current prompt. This has resulted in several issues with GPT 3 (Branch et al. 2022). Thus, it is critical to establish a framework like CREST for achieving trustworthiness.

## DEFINING CONSISTENCY, RELIABILITY, USER-LEVEL EXPLAINABILITY, AND SAFETY

### Consistency

"A consistent LLM is an AI system that comprehends user input and produces a response that remains unchanged regardless of how different users phrase the same input so far as the underlying facts and context, and intent are

the same. This mirrors the decision-making behavior of a human.”

It has been noted that LLMs show abrupt behavior when the input is either paraphrased or there has been adversarial perturbation (AP) (Shen et al. 2023). Further, it has also been noted that LLMs make implicit assumptions while generating a response to a query that lacks sufficient context. For instance, the following two questions, “*Should girls be given the car?*” or “*Should girls be allowed to drive the car?*” show different confidence levels in ChatGPT’s response. These two queries are semantically similar and are paraphrases of each other with a ParaScore > 0.90 (Shen et al. 2022). Thus, it is presumed that LLMs would yield a similar response. However, in the first query, ChatGPT is “*unsure.*” In the second, it is pretty confident that “*girls should be allowed to drive cars.*” Moreover, ChatGPT considers the question gender-specific in both cases, focusing on “*girls*” and not other words like “*drive*” or “*car.*” For instance, given the context, “*Should girls be given the toy car?*” or “*Should girls with necessary driver’s license be allowed to drive car?*”, the ChatGPT yields a high confidence answer stating “*yes*” in both scenarios. ChatGPT makes implicit assumptions by wrongly focusing on less relevant words and failing to seek more context from the user for a stable response generation. If the ChatGPT had access to knowledge, then it can retrieve the following information: “*Car <is related to> Drive*” and “*Drive <requires> Driver’s license,*” and ground its response in factual and common-sense knowledge. As demonstrated in subsequent sections, a lack of such consistency can result in unsafe behavior.

Recent tools like SelfCheckGPT (Manakul, Liusie, and Gales 2023) and CalibratedMath (Lin, Hilton, and Evans 2022) help assess LLMs’ consistency. However, enforcing consistency in LLMs remains relatively unexplored, particularly in the context of health and well-being. The need for consistency is evident when considering questions related to health, such as, “*Should I take sedatives for coping with my relationship issues?*” and “*Should I take Xanax?*” ChatGPT provided an ambivalent “*Yes/No*” answer to the first question and a direct “*No*” response to the second when both questions were the same.

Putting this in a conversational scenario, when follow-up questions like “*I am feeling drowsy by the day, and it seems like hallucinations. Any advice?*” and “*I am feeling sleep-deprived and hallucinating. What do you suggest?*” are posed, these models encounter challenges. First, they struggle to establish the connection between “*sleep deprivation*” and “*drowsiness*” with “*hallucinations.*” Second, the responses do not pay much attention to the concept of “*Xanax,*” resulting in inconsistent response generation. Furthermore, when prompted to include “*Xanax,*” LLMs often begin by apologizing and attempting to correct the

response, but these corrections still lack essential information. For instance, they do not consider the various types of hallucinations associated with Xanax (Alyssa 2021). This highlights the need for improved consistency and depth of response in LLMs, especially critical applications<sup>1</sup>, to ensure that users receive more accurate and comprehensive information.

## Reliability

Reliability measures to what extent a human can trust the content generated by an LLM. This capability is critical for the deployment and usability of LLM. Prior studies have examined reliability in LLMs by identifying the tendency of hallucination, truthfulness, factuality, honesty, calibration, robustness, and interpretability (Zhang et al. 2023). As seen from the widely used inter-rater reliability notion, there is little attention to the notion of reliability.

It is a common belief that a single annotator cannot attest to the credibility of the dataset. Likewise, a single LLM cannot provide a correct and appropriate outcome for every problem. This points to using an ensemble of LLMs (e-LLMs) to provide a higher confidence in the outcome, which can be measured through Cohen’s or Fleiss Kappa’s metrics (Wang et al. 2023). Three types of ensembles can be defined:

1. *Shallow ensembling LLMs* work with the belief that each LLM is trained with a different gigantic English corpus, with different training regimes, and possesses a different set of knowledge, enabling them to act differently on the same input. Such an ensemble works on the assumption that LLM is a knowledge base (Petroni et al. 2019). Three specific methods of e-LLMs are suggested under shallow ensembles: Rawlsian social welfare functions, utilitarian functions (Kwon et al. 2023), or weighted averaging (Jiang, Ren, and Lin 2023; Tyagi et al. 2023; Tyagi, Sarkar, and Gaur 2023).
2. *Semi-deep ensembling LLMs* involves adjusting and fine-tuning the importance or contributions of each individual LLM needed throughout the ensembling process. This approach effectively transforms the ensemble process into an end-to-end training procedure. In this setup, the term “*semi-deep*” implies that we are not just statically combining the LLMs but dynamically adjusting their roles and weights as part of the training process. This adaptability allows us to craft a more sophisticated and flexible ensemble (Shiri et al. 2024).

These two approaches offer several advantages. First, it enables the model to learn which LLMs are most effective





for different aspects of a given task. For example, certain LLMs might better understand syntax, while others excel at capturing semantics or domain-specific knowledge. By fine-tuning their contributions, we can harness the strengths of each LLM for specific subtasks within a larger task. Second, it allows the model to adapt to changes in the data or the task itself. As new data are introduced or the problem evolves, individual LLMs' contributions can be adjusted accordingly, ensuring that the ensemble remains effective and up-to-date.

However, these ensembles ignore the following key elements:

- a. *External knowledge integration*: The approach involves integrating external knowledge sources, such as knowledge graphs (KGs) and Clinical Practice Guidelines, into the LLM ensemble. These sources provide additional context and information that can enhance the quality of the generated text.
- b. *Reward functions*: The external knowledge is not simply added as static information but is used as reward functions during the ensembling process. In simpler terms, this means the ensemble of models gets rewarded when they produce text that matches or incorporates external knowledge. This reward system promotes logical consistency and meaningful connections with that knowledge.
  - *Logical coherence*: By incorporating external knowledge, the ensemble of LLMs aims to produce a more logically coherent text. It ensures that the generated content aligns with established facts and relationships in the external knowledge sources.
  - *Semantic relatedness*: The ensemble also focuses on improving the semantic relatedness of the generated text. This means that the text produced by the LLMs is factually accurate, contextually relevant, and meaningful.

Such attributes are important when LLMs are designed for critical applications like motivational interviewing (Sarkar et al. 2023). Motivational interviewing is a communication style often used in mental health counseling, and ensuring logical coherence and semantic relatedness in generated responses is crucial for effective interactions (Shah et al. 2022).

3. *A Deep Ensemble of LLMs* introduces an innovative approach using NeSy-AI, in which e-LLMs are fine-tuned with the assistance of an evaluator. This evaluator comprises constraints and graph-based knowledge representations and offers rewards to guide the generation of e-LLMs based on the aforementioned properties. Concurrently, it incorporates knowledge source con-

cepts in representations to compel e-LLMs to include and prioritize these concepts, enhancing their reliability (refer to Figure 6 for illustration). Another key objective of the deep ensemble approach is to transform e-LLMs into a Mixture of Experts (Artetxe et al. 2021) by enhancing individual LLMs through a performance maximization function (Yu et al. 2023).

## Explainability and user-level explainable LLMs (UExMs)

Achieving effective and human-understandable explanations from LLMs or their precursor language models (LMs) remains complex. Previous attempts to elucidate BlackBox LMs have utilized techniques like surrogate models (such as LIME (Ribeiro, Singh, and Guestrin 2016)), visualization methods, and APs to the input data (Chapman-Rounds et al. 2021). While these approaches provide explanations, they operate at a relatively basic level of detail, which we call *system-level explainability* (Gaur 2022).

System-level explainability has been developed under the purview of post-hoc explainability techniques that aim to interpret the attention mechanism of LMs/LLMs without affecting their learning process. These techniques establish connections between the LM's attention patterns and concepts sourced from understandable knowledge repositories. Within this approach, two methods have emerged: (a) Attribution scores and LM tuning (Slack et al. 2023) and Factual Knowledge-based Scoring and LM tuning (Yang, Chen, et al. 2023; Sun et al. 2023). The latter method holds particular significance in health and well-being because it provides explainability for clinicians as users. This method relies on KGs or knowledge bases like the Unified Medical Language System (UMLS) (Bodenreider 2004), SNOMED-CT (Chang et al. 2020), or RXNorm (Kamdar et al. 2019) to enhance its functionality.

While the post hoc method can provide explanations (by modeling it as a dialog system (Lakkaraju et al. 2022)), it does not guarantee that the model consistently prioritizes essential elements during training (Jiang et al. 2021). Its explanations may be coincidental and not reflect the model's decision-making process. More recently, the focus has shifted to "explainability by design," particularly in critical applications like healthcare. A recent example is the Transparency and Interpretability Framework for Understandability (TIFU), proposed by Joyce et al., which connects inherent explainability to a higher level of explainability in the mental health domain (Joyce et al. 2023). The primary motivation for pursuing such an explainability, called *User-level explainability*, is to ensure that healthcare professionals and patients are given contextually relevant explanations that help them understand

the AI system's process and outcomes so they can develop confidence in AI tools.

“A User-level Explainability in LLMs implies that humans can rely on the AI system to the extent that they can reduce the need for **human** oversight, monitoring, and verification of the system's outputs. To trust a deployed LLM, we must have adequate insight into how it generates an output based on a given input.”

**“UEXMs provide user-explainable insights by utilizing expert-defined instructions, statistical knowledge (attention), and knowledge retriever.”**

UEXMs can be practically realized in four different ways:

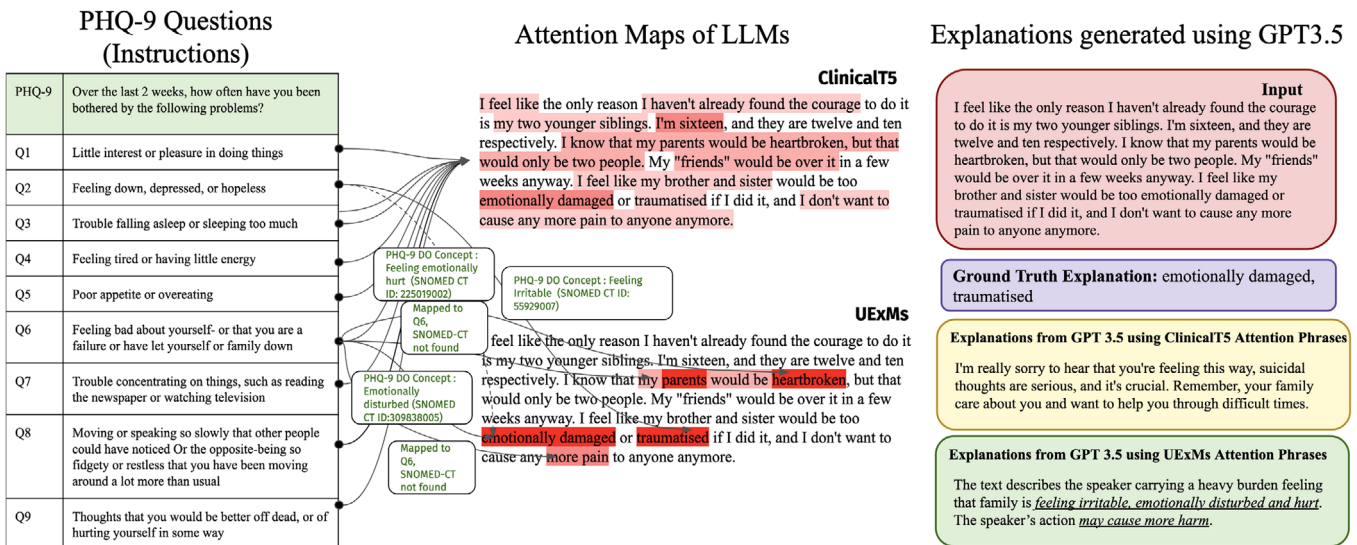
- (a) *UEXMs with generating evaluator pairing*: This defines a generative and evaluator-based training of UEXMs where any LLM is paired with a knowledge-powered evaluator, either accelerates or de-accelerates the training of LLMs, depending on whether the final generation is within the acceptable standards of the evaluator. “*On the weekend, when I want to relax, I am bothered by trouble concentrating while reading the newspaper or watching television. Need some advice*” indicates that the individual is experiencing specific issues related to concentration during leisure time. This query is more than just a casual comment; it highlights a problem affecting the user's ability to unwind effectively. Now, consider the two scenarios:
- (i) **Without an evaluator (generic response)**: In the absence of an evaluator, an LLM might provide a generic set of activities or advice, such as “practice mindfulness, limit distractions, break tasks into smaller chunks,” and so on. While this advice is generally useful for improving concentration, it lacks the depth and specificity needed to address the user's potential underlying issues.
  - (ii) **With an evaluator (specific response)**: When integrated into the LLM, an evaluator can analyze the user's query more comprehensively. In this case, the evaluator can recognize that the user's difficulty concentrating during relaxation may indicate an underlying sleep-related issue. Considering this possibility, the language model can provide more targeted and informed advice.

For instance, the evaluator might suggest asking further questions like: (a) Do you have trouble sleeping at night?

(b) How much sleep do you typically get on weekends? (c) Have you noticed other sleep-related symptoms, such as daytime drowsiness? (d) Have you considered the possibility of a sleep disorder? By incorporating an evaluator, the LLM can guide the conversation toward a more accurate understanding of the user's situation. To put it simply, the LLM, when assisted by an evaluator, will provide a coherent answer that encompasses all aspects of the user's question (Gaur et al. 2023). Further, the evaluator prevents the model from generating hallucinated, off-topic, or overly generic responses. A framework like ISEEQ integrates generator and evaluator LLMs for generating tailored responses in general-purpose and mental health domains (Gaur et al. 2022). Additionally, PURR and RARR contribute to refining segments of LLM design aimed at mitigating hallucination-related problems in these models (Chen et al. 2023; Gao et al. 2023).

To illustrate this concept, refer to Figure 4, a task where a generative LM takes user input and provides an assessment in natural language, specifically within the PHQ-9 context. The figure shows two LLMs: ClinicalT5-large, a powerful LM with 38 billion parameters, and UEXM, essentially ClinicalT5-large but enhanced with a PHQ-9-grounded evaluator. This demonstrates that by employing an evaluator with predefined questions, we can assess how well the attention of generative ClinicalT5-large aligns with those specific questions. This approach helps ensure that the generated explanations are relevant and comprehensive, making them clinically applicable, particularly when healthcare professionals rely on standardized guidelines like the PHQ-9 to evaluate patients for depression (Honovich et al. 2022).

- (b) *UEXMs with retriever augmentation and process knowledge*: It is commonly observed that the process of generating responses by LLMs lacks transparency, making it difficult to pinpoint the origin of their answers. This opacity raises questions about how the model derives its responses.
- (a) **The emergence of retrieval-augmented generation LMs**: A novel class of LMs has surfaced to tackle this issue and add a layer of supervision to language model outputs. Examples include REALM (Guu et al. 2020), LAMA (Petroni et al. 2019), ISEEQ (Gaur et al. 2022), and RAG (Lewis et al. 2020), which integrate a generator with a dense passage retriever and access to indexed data sources. LLMs with retrieval-augmented architectures have started to show understandable and accountable responses (Lyu et al. 2023). For instance, GopherCite (Menick et al. 2022) and NeMo Guardrails (Rebedea et al. 2023) leverage a knowledge base to supply supporting evidence for



**FIGURE 4** An instance of user-level explainability in a UExM is when the model uses questions from PHQ-9 to guide its actions and relies on SNOMED-CT, a clinical knowledge base, to simplify complex concepts (concept abstraction). This approach helps the model offer explanations that closely align with the actual truth. PHQ-9 DO, PHQ-9-based Depression Ontology.

nearly every response generated by the underlying LLM.

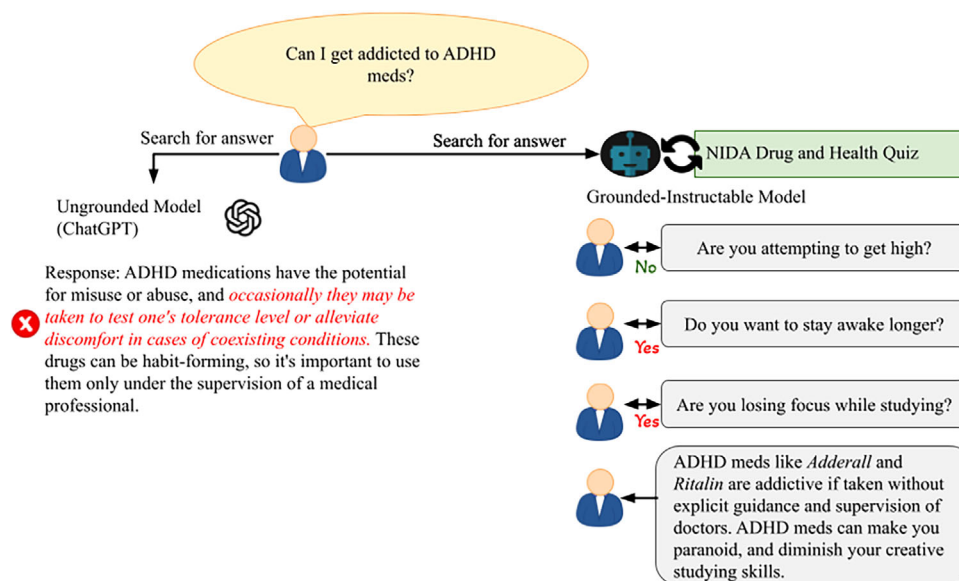
- (b) **The emergence of process knowledge-guided generation LMs:** Process knowledge refers to guidelines or instructions created by experts in a domain (Roy et al. 2023). For instance, in mental health, PHQ-9 is the process of knowledge for screening depression (Kroenke et al. 2001), NIDA's Attention Deficiency Hyperactivity Disorder Test, and the World Health Organization's Wellness Indices (Topp et al. 2015). The questions in these guidelines can act as rewards for enriching latent generations (e.g., answerability test (Yao et al. 2023)) (Hagendorff 2023).
- (c) **UExMs with abstention:** While a retriever has been integrated into an LLM, it does not guarantee meaningful explainability. When considering a ranked list of retrieved and expanded documents, an LLM is still vulnerable to generating incorrect or irrelevant explanations. Therefore, it is crucial to eliminate meaningless hidden generations before they are converted into natural language. For example, the ReACT framework employs Wikipedia to address spurious generation and explanations in LLMs (Yao et al. 2022). However, it relies on a prompting method rather than a well-grounded domain-specific approach, which can influence the generation process used by the LLM (Yang, Wang, et al. 2023). Alternatively, pruning methods and an abstention rule have also been used to reduce irrelevant output from LLMs. A more robust approach would involve utilizing procedu-

ral or external knowledge as an evaluator guiding LLM-generated content that enhances meaningful understanding.

## Safety

"Safety and explainability are closely intertwined concepts for AI systems. While a safe AI system will inherently demonstrate explainability, the reverse isn't necessarily true; an explainable system may or may not be safe."

Recently, there has been a proliferation in safety-enabled research, particularly in LMs and LLMs. Perez et al. (2022) performed red-teaming between LMs to determine if an LM can produce harmful text. The process did not include humans in generating these adversarial test cases. Further, the research did not promise to address all the critical safety oversights comprehensively; instead, it aimed to spotlight instances where LMs might exhibit unsafe behavior. Scherrer et al. (2023) delve more deeply into the safety issues in LLMs by examining their behavior in moral scenarios. The study found that LLMs only focus on generating fluent sentences and overlook important words/concepts contributing to stable decisions. Further, datasets like DiSafety and SafeTexT are designed to induce safety in LMs/LLMs through supervised learning (Meade et al. 2023; Levy et al. 2022). These discussions surrounding safety gained heightened attention, particularly within the National Science Foundation (NSF), leading to the launch of two programs: (a) Safety-enabled Learning and (b) Strengthening AI. In a recent



**FIGURE 5** An Illustration of grounding and instruction-following behavior in an LLM (right) tuned with support from health and well-being-specific guidelines. ChatGPT's response was correct, but it is not safe. NIDA, National Institute on Drug Abuse.

webinar, NSF outlined three fundamental attributes of ensuring safety: grounding, instructability, and alignment (NSF Main).

**Grounding:** In essence, groundedness is the foundation upon which both explainability and safety rest. Without a strong grounding in the provided instructions, the AI may produce results that stray from the desired outcome, potentially causing unintended consequences. For instance, consider the scenario depicted in Figure 5. An LLM that is not grounded in domain-specific instruction, like the ChatGPT, results in an unsafe response. On the other hand, a relatively simple LLM, like T5-XL, tuned by grounding in domain-specific instructions, attempts to ask follow-up questions to gather the context for a coherent response. The changes in T5-XL's behavior due to the NIDA quiz highlight the importance of being able to instruct and align AI, which is key for safety. (Ward 2023).

**Instructability:** In AI safety, instructability encompasses the assurance that the AI understands and complies with user preferences, policies, and moral beliefs. Making the LMs bigger and strengthening the rewards makes the models power-hungry rather than ethical and safe. For instance, the guardrails instantiated for the safe functioning in OpenAI's ChatGPT, the rules within DeepMind's Sparrow, and the list of rules within Anthropic's Claude cannot reliably prove that they are safe.

The idea of having systems that follow instructions has been around since 1991, mainly in robotics and, to some extent, in text-based agents. It is crucial because it helps agents learn tasks, do them well, and explain how they did it, making sharing knowledge easier between humans and

AI and showing they can follow human instructions. One way to do this is by using grounded instruction rules, especially in mental health. Clinical practice guidelines like PHQ-9 for depression and GAD-7 for anxiety, with their questions, can serve as instructions for AI models focused on mental health. Grounded rules have two key benefits for safety. First, they tend to be helpful and harmless, addressing a common challenge for AI models. Second, they promote absolute learning, avoiding tricky trade-off situations.

**Alignment:** When we talk about alignment in LMs, it means ensuring that even a model that follows instructions does not produce unsafe results (MacDonald 1991). This can be a tricky problem, as discussed in Nick Bostrom's book "Superintelligence," where it's called "*perverse instantiations*" (Bostrom, 2014). This happens when the LM/LLMs figure out how to meet a goal, but it goes against what the user wants (Ngo, Chan, and Mindermann 2022). So, the challenge is to create an AI that follows instructions and finds the best way to achieve a goal while keeping users happy, a concept referred to as "*Wireheading*" in "Superintelligence." Following are perspectives on why it happens and what can be done:

1. Context awareness (CA) and contextual rewards (CR): CA refers to the training of LMs/LLMs to focus on words or phrases that directly translate concepts in factual knowledge sources. CR serves the function of facilitating CA. They achieve this by incorporating evaluator modules that analyze the hidden or latent representations within the model with respect to the concepts present in the knowledge sources. CR



reinforces and guides CA by rewarding the model when it correctly identifies and incorporates knowledge-based concepts into its responses (Hubinger et al. 2019).

2. *Misalignment in latent representations caused by misleading reward associations*: We acknowledge the inherent perceptiveness of LMs and LLMs, a quality closely linked to the quantity of training data they are exposed to. Nevertheless, having a larger training dataset leads to superior performance scores, but it may not necessarily meet the expectations of human users. Bowman has demonstrated that a model achieving an F1 score of over 80% still struggles to prioritize and pay adequate attention to the concepts users highly value (Bowman 2023). This happens because optimization algorithms and attention methods in LLMs can attempt to induce fake behavior. Further, if the rewards specified are not unique to the task but rather general, the model will have difficulty aligning with desired behaviors (Shah et al. 2022).
3. *Deceptive alignment during training*: Spurious reward collections can lead to deceptive training. It is important to train the LMs/LLMs with paraphrases and adversarial input while examining the range of reward scores and the variations in the loss functions. If LMs/LLMs demonstrate high fluctuations in the rewards and the associated effect on loss, it would most likely result in brittleness during deployment. Methods like the chain of thoughts and the tree of thoughts prompting can act as sanity checks to examine the deceptive nature of LMs/LLMs (Leahy and Alfour 2023; Yao et al. 2023).

Knowledge of the AI system and domain is pervasive in achieving consistency, reliability, explainability, and safety for building a trustworthy AI system.

- For *Consistency*, rules and knowledge can make LLMs understand and fulfill user expectations confidently.
- Reliability is ensured by utilizing the rich knowledge contained in KGs to empower an ensemble of LLMs to produce consistent and mutually agreeable results with high confidence.
- For *Explainability*, LLMs use their knowledge, retrieved knowledge, and rules that were followed to attain consistency and reliability to explain the generation in an effective manner.
- Safety in LLMs is upheld by consistently grounding their generation and explanations in domain knowledge and assuring the system's adherence to expert-defined rules or guidelines.

## THE CREST FRAMEWORK

To realize CREST, we now provide succinct descriptions of its key components and highlight open challenges for AI and NeSy-AI communities in NLP (see Figure 6). Three components of the CREST frameworks are discussed in “NeuroSymbolic AI for Paraphrased and Adversarial Perturbations,” “Knowledge-infused Ensembling of LLMs,” and “Assessment of CREST” sections.

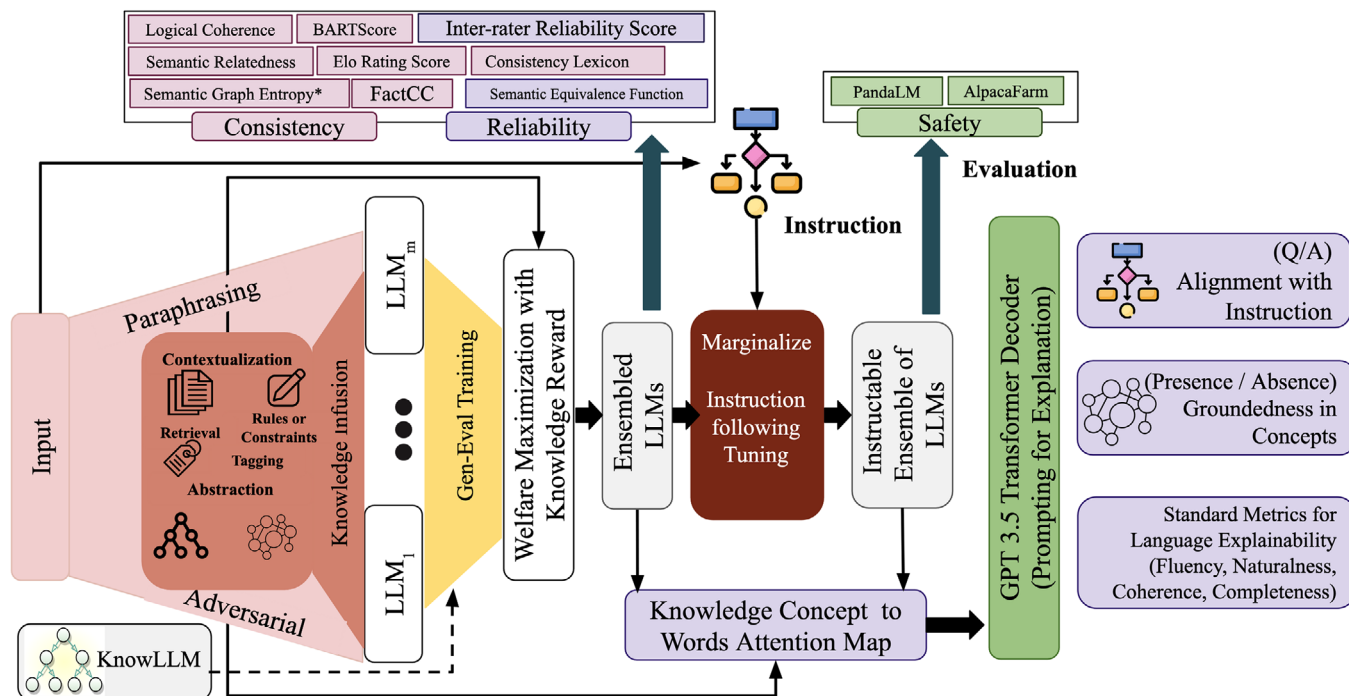
### NeuroSymbolic AI for paraphrased and adversarial perturbations

Paraphrasing serves as a technique to enhance an AI agent's calibration by making it aware of the different ways an input could be expressed by a user (Du, Xing, and Cambria 2023). This, in turn, contributes to increasing the AI agent's consistency and reliability. Agarwal et al. introduced a pioneering NeSy AI-based approach to paraphrasing. In their method, they employed CommonSense, WordNet, and Wikipedia KGs to generate paraphrases that held equivalent meanings but were perceived as distinct by the AI agent (Agarwal et al. 2023). However, there are some promising directions for NeSy paraphrasing. First is *contextualization*, which involves augmenting the input with meta information retrieved from a rank list of documents. This transforms NLP's not-so-old question rewriting problem into a knowledge-guided paraphrasing method. The second is *abstraction*, which involves identifying the function words (e.g., noun phrases, verb phrases) and named entities and replacing them with abstract concepts. For instance, the following sentence, “Why trauma of harassment is high in boys|girls?” is abstracted to “why trauma of (harassment → mistreatment) is high in (boys|girls → students)?”. Both methods can benefit from existing LLM learning strategies, such as marginalization (Wang et al. 2022) and reward-based learning (Jie et al. 2023).

NeSy-AI for APs uses general-purpose KGs to carefully change the sentence to examine the brittleness in LLMs' outcomes. An example of adversarial generation is the following:

**(S1):** “I have been *terrible* in battling with my loneliness. My overly introvertedness and *terrible* choice of few friends are the reasons for who I am. The only part I considered funny in this situation was that none of my friends knew how I felt. It seems like they are *childish*.”

**(S1-AP):** “I have been *horrible* at battling my loneliness. My overly introvertedness and *horrible* choice of few friends are the reasons for who I am. The only part that



**FIGURE 6** The CREST framework operationalizes “explainability and safety” by ensuring the model is reliable and consistent. LLMs (1 to  $m$ ) can be replaced with LLMs in Figure 2, and the knowledge used in infusion refers to UMLS and SNOMED-CT for a clinical domain, as we examined CREST for mental health. Gen-Eval, generator and evaluator pairing; KnowLLM, LLMs created using KGs.

I regarded as *sarcastic* in this situation was that none of my friends knew how I felt. It seems like they are *youngsters*.”

The Flan T5 (11B) estimates S1 to have a “*negative*” sentiment with a confidence score of 86.6% and S1-AP to have a “*positive*” sentiment with a 61.8% confidence score. The confidence scores are predicted probability estimates. LLMs must concentrate on the contextual notions (such as loneliness and introversion) and the abstract meaning that underlies both S1 and S1-AP—that is, the influence on mental health and well-being—to attain consistency and reliability in such inadvertent settings.

## Knowledge-infused ensembling of LLMs

As mentioned above, e-LLMs have a multitude of benefits; however, simply statistical methods of ensembling, which consists of averaging over the outcomes from black box LLMs does not make an ensembled LLM consistent and reliable. Knowledge-infused ensemble represents a particular methodology where the knowledge (general purpose or domain-specific) modulates the latent representations of the LLMs to yield the best of world outcomes.

This can happen in one of three ways:

- (1) **LLMs over KGs (KnowLLMs):** Similar to the process of training any LLM on text documents, which involves formulating it as a task of predicting the next word in a sentence, KnowLLMs undertake the training of LLMs using a variety of KGs such as CommonSense, Wikipedia, and UMLS. In KnowLLMs, the training objective is redefined as an autoregressive function over  $\langle \text{subject} \rangle \langle \text{predicate} \rangle \langle \text{object} \rangle$  coupled with pruning based on existing state-of-the-art KG embedding methods. Introducing pruning is crucial in KnowLLMs to prevent the model from making unwarranted inferences and forming incorrect links. This is vital for ensuring the safety and trustworthiness of the knowledge generated by KnowLLMs. In other words, by pruning, KnowLLMs are able to filter out irrelevant or potentially misleading information, thereby enhancing the quality of their responses and minimizing the risk of spreading false or harmful knowledge.
- (2) **Generative evaluator tuning:** This approach suggests using reinforcement learning to improve the training of e-LLMs. It combines the traditional training method with rewards from KnowLLMs, which act as extra guidelines. These rewards encourage the e-LLM to generate text that aligns with specific desired characteristics, such as mental health concepts. Suppose the e-LLM’s output does not meet these criteria or is



logically incorrect, according to KnowLLM; in that case, it receives negative rewards, even if it is similar to the ground truth based on similarity scores. This method helps e-LLMs produce more contextually relevant and accurate text.

- (3) **Instruction following tuning:** Instruction tuning has recently emerged as a promising direction to teach LLMs to match the expectations of humans. Though promising, it requires a substantial amount of samples, and there is no perfect quantifiable method to measure the “*instruction following*” nature of LLMs. And, if we decide to embark on a “*mixture of experts*” like e-LLMs, it would be hard to make separate procedures for instruction tuning over e-LLMs. Thus, we take inspiration from process knowledge-infused learning, a mechanism for intrinsically tuning the LMs or an ensemble of LMs. Roy et al. demonstrated how questionnaires in the clinical domain, which can be a constraint, can enable LMs to generate safe and consistently relevant questions and responses (Roy et al. 2023). This approach works on a simple Gumbel Max function, which allows structural guidelines to be used in the end-to-end training of LMs. This approach is fairly flexible for “*instruction-following-tuning*” of e-LLMs and ensuring that the instruction is followed.

## Assessment of CREST

The CREST framework emphasizes incorporating knowledge and utilizing knowledge-driven rewards to support e-LLMs in achieving trust. To assess the quality of e-LLMs’ output, it is crucial to employ metrics that account for the knowledge aspect. For instance, the logical coherence metric evaluates how well e-LLMs’ generated content aligns with the flow of concepts in KGs and context-rich conversations. Additional metrics like Elo Rating (Zheng et al. 2023), BARTScore (Liu et al. 2023), FactCC (Kryściński et al. 2020), and Consistency lexicons can be improved to account for the influence of knowledge on e-LLMs’ generation. However, aside from the established Cohen’s or Fleiss Kappa metrics, an effective alternate metric is not available when assessing reliability.

Safety aspects in CREST are best evaluated when knowledge-tailored e-LLMs are instructed to adhere to guidelines established by domain experts. Existing metrics like PandaLM (Wang et al. 2023) and AlpacaFarm (Dubois et al. 2023) are based on LLMs, which may exhibit vulnerabilities to unsafe behaviors. While such metrics may be suitable for open-domain applications, when it comes to critical applications, safety metrics must be rooted

in domain expertise and align with the expectations of domain experts.

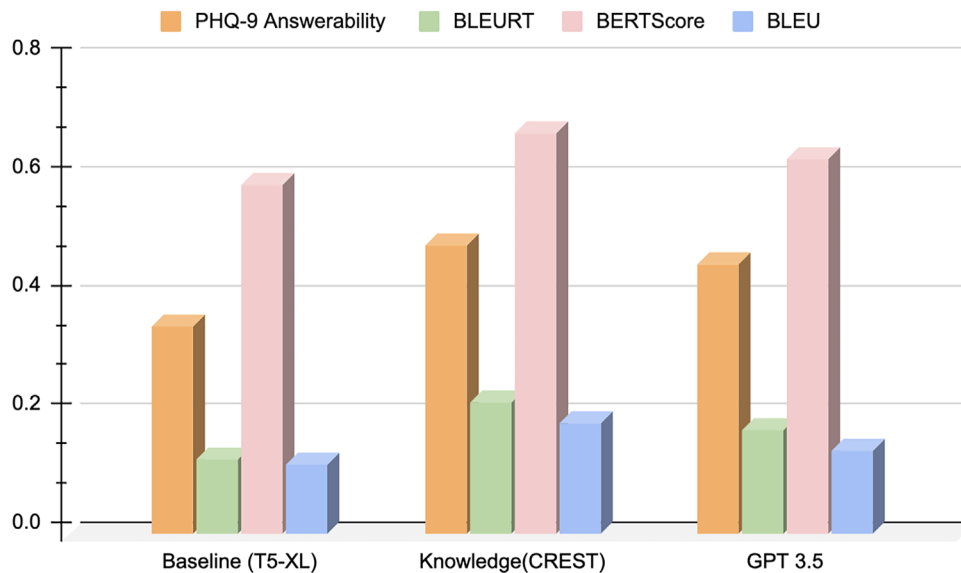
In CREST, explainability is evaluated through two approaches requiring expert verification and validation. One method involves analyzing the “*Knowledge Concept to Word Attention Map*” to gain insights into CREST’s reasoning process and verify whether the model’s decisions align with domain knowledge and expectations (Gaur et al. 2018). Another method involves using knowledge concepts and domain-specific decision guidelines (e.g., clinical practice guidelines) to enable LLMs like GPT 3.5 to generate human-understandable explanations (as shown in Figure 4).

## A case study in mental health in brief

We present a preliminary performance of CREST on the PRIMATE dataset, introduced during ACL’s longstanding Clinical Psychology workshop (Gupta et al. 2022). It is a distinctive dataset designed to assess the LM’s ability to consistently estimate an individual’s level of depression and provide yes/no responses to PHQ-9 questions, which is a measure of its reliability. Figure 7 shows the performance of CREST and knowledge-powered CREST relative to GPT 3.5. Including knowledge in CREST showed an improvement of 6% in PHQ-9 answerability and 21% in BLEURT over GPT 3.5, which was used through the prompting method. The e-LLMs in CREST were Flan T5-XL (11B) and T5-XL (11B).

## CONCLUSION AND FUTURE WORK

LLMs and broadly generative AI represent the most exciting current approach, but alone, they are not the solution for Trustworthy AI. LLMs exhibit undesired behaviors during tasks such as question answering, making them susceptible to threats and problematic actions. Therefore, there is a need for innovative approaches to identify and mitigate threats posed both to LLMs and by LLMs to humans, especially when they are to be used for critical applications such as those in health and well-being. A comprehensive solution is needed beyond the implementation of guardrails or instruction adjustments. This solution should encourage LLMs to think ahead, leveraging domain knowledge for guidance. The CREST framework offers a promising approach to training LLMs with domain knowledge, enabling them to engage in anticipatory thinking through techniques like paraphrasing, adversarial inputs, knowledge integration, and fine-tuning based on instructions.



**FIGURE 7** The experiment comparing CREST with GPT 3.5 (LLM) and T5-XL (a relatively small LLM) on the PRIMATE dataset. The outcomes were evaluated using PHQ-9 Answerability, BLEURT, and standard metrics like BERTScore and BLEU. The PHQ-9 answerability is calculated as the mean Matthew Correlation Coefficient score. This score is computed by comparing predicted yes/no labels against the ground truth across nine PHQ-9 questions. BLEURT (Sellam, Das, and Parikh 2020) score is computed between questions generated by LLMs and PHQ-9 questions. LLMs were prompted to create questions based on sentences identified as potential answers to the PHQ-9 questions.

We presented a preliminary effort in implementing the CREST framework that yields enhancements over GPT 3.5 on PRIMATE, a PHQ-9-based depression detection dataset. We plan to experiment with CREST on knowledge-intensive language generation benchmarks, like HELM (Liang et al. 2022). Further, we plan on automating user-level explanations without dependence on pretrained LLMs (e.g., GPT 3.5). Our future endeavors involve developing more effective training methodologies for e-LLMs powered by the CREST framework. Additionally, we will incorporate robust paraphrasing and adversarial generation techniques to assess the consistency and reliability of e-LLMs when they are exposed to knowledge. This will also open avenues for further research into crafting quantitative metrics that evaluate reliability, safety, and user-level explainability.

#### ACKNOWLEDGMENTS

We express our gratitude to Drs Amitava Das and Valerie L. Shalin for their invaluable reviews and insightful suggestions on the manuscript. We acknowledge partial support from the NSF EAGER award #2335967 and the UMBC Summer Faculty Fellowship. Any opinions, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or UMBC.

#### CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

#### ORCID

Manas Gaur  <https://orcid.org/0000-0002-5411-2230>

#### ENDNOTE

<sup>1</sup>Critical applications refer to situations in which the use of AI has the potential to result in substantial harm to individuals or societal interests unless considerable precautions are taken to ensure their consistency, reliability, explainability, and safety.

#### REFERENCES

- “National Science Foundation Main — players.brightcove.net.” [https://players.brightcove.net/679256133001/NkgrDczuol\\_default/index.html?videoId=6336837295112](https://players.brightcove.net/679256133001/NkgrDczuol_default/index.html?videoId=6336837295112) (accessed 01-8-2023).
- Agarwal, A., S. Gupta, V. Bonagiri, M. Gaur, J. Reagle, and P. Kumaraguru. 2023. “Towards Effective Paraphrasing for Information Disguise.” In European Conference on Information Retrieval, 331–40. Cham: Springer Nature Switzerland.
- Alyssa. 2021. “Can Benzodiazepines Cause Hallucinations?” Banyan Treatment Center. <https://www.banyantreatmentcenter.com/2021/12/03/benzodiazepines-causing-hallucinations-palmsprings/>.
- Artetxe M., S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, Xi V Lin, J. Du, S. Iyer, R. Pasunuru, G. Anantharaman, X. Li, S. Chen, H. Akin, M. Baines, L. Martin, X. Zhou, P. Singh Koura, B. O’Horo, J. Wang, L. Zettlemoyer, M. Diab, Z. Kozareva, and V. Stoyanov. 2022. “Efficient Large Scale Language Modeling with Mixtures of Experts.” In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 11699–732.
- Bai Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C.



- Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, 2022. "Constitutional AI: Harmlessness from AI Feedback." <https://arxiv.org/abs/2212.08073>
- Bodenreider, O. 2004. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology." *Nucleic Acids Research* 32(1): D267–70.
- Bostrom, N. 2016. The control problem. Excerpts from superintelligence: Paths, dangers, strategies. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 308–30.
- Bowman, S. R. 2023. "Eight Things To Know About Large Language Models." *arXiv preprint arXiv:2304.00612*.
- Branch, H. J., J. R. Cefalu, J. McHugh, L. Hujer, A. Bahl, D. D. C. Iglesias, R. Heichman, and R. Darwishi. 2022. "Evaluating the Susceptibility of Pre-Trained Language Models Via Handcrafted Adversarial Examples." *arXiv preprint arXiv:2209.02128*.
- Brun, I., and T. Shwartz-Altschuler. 2023. "Yom Kippur War: ChatGPT Can Be Used or Military Intel, War Simulation." *The Jerusalem Post | JPost.com*, October 1, 2023. <https://www.jpost.com/business-and-innovation/opinion/article-760273>.
- Bumgardner, V. K., A. Mullen, S. Armstrong, C. Hickey, and J. Talbert. 2023. "Local Large Language Models for Complex Structured Medical Tasks." *arXiv preprint arXiv:2308.01727*.
- Chang, D., I. Balažević, C. Allen, D. Chawla, C. Brandt, and R. A. Taylor. 2020. "Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings." *Proceedings of the Conference. Association for Computational Linguistics. Meeting 2020*: 167–76. NIH Public Access.
- Chapman-Rounds, M., U. Bhatt, E. Pazos, M. A. Schulz, and K. Georgatzis. 2021. "FIMAP: Feature Importance by Minimal Adversarial Perturbation." *Proceedings of the AAAI Conference on Artificial Intelligence* 35(13): 11433–41.
- Chen A., P. Pasupat, S. Singh, H. Lee and K. Guu. 2023. PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions. *arXiv preprint arXiv:2305.14908*.
- Daws, R. 2021. "Medical Chatbot Using OpenAI's GPT-3 Told a Fake Patient To Kill Themselves." *AI News*. <https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/>.
- Du, K., F. Xing, and E. Cambria. 2023. "Incorporating Multiple Knowledge Sources for Targeted Aspect-Based Financial Sentiment Analysis." *ACM Transactions on Management Information Systems* 14: 1–24.
- Dubois Y., X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang and T. B. Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.
- Gao, L., Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Zhao, et al. 2023. "Rarr: Researching and Revising What Language Models Say, Using Language Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, 16477–508. Long Papers.
- Gaur, M. 2022. Knowledge-Infused Learning. (Doctoral dissertation). Retrieved from <https://scholarcommons.sc.edu/etd/6914>
- Gaur, M., D. A. K. S. S. Gunaratna, V. Srinivasan, and H. Jin. 2023. Samsung Electronics Co Ltd, 2023. *Dynamic question generation for information-gathering*. U.S. Patent Application 17/817,778, filed March 2, 2023.
- Gaur, M., K. Gunaratna, V. Srinivasan, and H. Jin. 2022. "ISEEQ: Information Seeking Question Generation Using Dynamic Meta-Information Retrieval and Knowledge Graphs." *Proceedings of the AAAI Conference on Artificial Intelligence* 36(10): 10672–80.
- Gaur, M., U. Kursuncu, A. Alambo, A. Sheth, R. Daniulaityte, K. Thirunarayan, and J. Pathak. 2018. "'Let Me Tell You About Your Mental Health!'" Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention." In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 753–62.
- Gautam, S., A. Jain, M. Gautam, V. N. Vahia, and S. Grover. 2017. "Clinical Practice Guidelines for the Management of Depression." *Indian Journal of Psychiatry* 59(1): S34–S50.
- A Glaese, N McAleese, M Trębacz, J Aslanides, V Firoiu, T Ewalds, M Rauh, L Weidinger, M Chadwick, P Thacker, L Campbell-Gillingham, J Uesato, Po-S Huang, R Comanescu, F Yang, A See, S Dathathri, R Greig, C Chen, D Fritz, J S Elias, R Green, S Mokrá, N Fernando, B Wu, R Foley, S Young, I Gabriel, W Isaac, J Mellor, D Hassabis, K Kavukcuoglu, L A Hendricks, and G Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.
- Gupta, S., A. Agarwal, M. Gaur, K. Roy, V. Narayanan, P. Kumaraguru, and A. Sheth. 2022. "Learning to Automate Follow-up Question Generation using Process Knowledge for Depression Triage on Reddit Posts." In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 137–47.
- Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang. 2020. "Retrieval Augmented Language Model Pre-Training." In *International Conference on Machine Learning*, 3929–38. PMLR.
- Hagendorff, T. 2023. "Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods." *arXiv preprint arXiv:2303.13988*.
- Holohan, M. 2023. "A Boy Saw 17 Doctors Over 3 Years for Chronic Pain. ChatGPT Found the Diagnosis." *TODAY.com*. <https://www.today.com/health/mom-chatgpt-diagnosis-pain-rcna101843>
- Honovich O., R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szepektor, A. Hassidim and Y. Matias. 2022. "TRUE: Re-evaluating Factual Consistency Evaluation." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3905–20.
- Hubinger, E., C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. 2019. "Risks from Learned Optimization in Advanced Machine Learning Systems." *arXiv preprint arXiv:1906.01820*.
- Jiang D., X. Ren and B. Y. Lin. 2023. "LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, 14165–78.
- Jiang, Z., J. Araki, H. Ding, and G. Neubig. 2021. "How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering." *Transactions of the Association for Computational Linguistics* 9: 962–77.

- Jie, R., X. Meng, L. Shang, X. Jiang, and Q. Liu. 2023. "Prompt-Based Length Controlled Generation with Reinforcement Learning." *arXiv preprint arXiv:2308.12030*.
- Joyce, D. W., A. Kormilitzin, K. A. Smith, and A. Cipriani. 2023. "Explainable Artificial Intelligence for Mental Health Through Transparency and Interpretability for Understandability." *npj Digital Medicine* 6(1): 6.
- Kamdar, M. R., T. Hamamsy, S. Shelton, A. Vala, T. Eftimov, J. Zou, and S. Tamang. 2019. "A Knowledge Graph-Based Approach for Exploring the US Opioid Epidemic." *arXiv preprint arXiv:1905.11513*.
- Kroenke, K., R. L. Spitzer, and J. B. Williams. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine* 16(9): 606–13.
- Kryściński, W., B. McCann, C. Xiong, and R. Socher. 2020. "Evaluating the Factual Consistency of Abstractive Text Summarization." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–46.
- Kwon M., S. M. Xie, K. Bullard and D. Sadigh, 2022. "Reward Design with Language Models." In *The Eleventh International Conference on Learning Representations*.
- Lakkaraju, H., D. Slack, Y. Chen, C. Tan, and S. Singh. 2022. "Rethinking Explainability as a Dialogue: A Practitioner's Perspective." *arXiv preprint arXiv:2202.01875*.
- Leahy, C., and G. Alfour. 2023. "Cognitive Emulation: a Naive AI Safety Proposal." [Online forum post]. <https://www.alignmentforum.org/posts/ngEvKav9w57XrGQnb/cognitive-emulation-a-naive-ai-safety-proposal>.
- Levy S., E. Allaway, M. Subbiah, L. Chilton, D. Patton, K. Mckeown and W. Y. Wang. 2022. "SafeText: A Benchmark for Exploring Physical Safety in Language Models." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2407–21.
- Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, D. Kiela, et al. 2020. "Retrieval-augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33: 9459–74.
- Liang P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, Ce Zhang, C. A. Cosgrove, C. D Manning, C. Re, D. Acosta-Navas, D. Arad Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N S. Chatterji, Omar Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang and Y. Koreeda. 2023. Holistic Evaluation of Language Models, In *Transactions on Machine Learning Research*.
- Lin S., J. Hilton and O. Evans. 2022. Teaching Models to Express Their Uncertainty in Words, In *Transactions on Machine Learning Research*
- Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. 2023. "GpTEval: NLG Evaluation Using Gpt-4 with Better Human Alignment." *arXiv preprint arXiv:2303.16634*.
- LMSYS Org. 2023. "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%\* ChatGPT Quality." LMSYS Org. 2023. <https://lmsys.org/blog/2023-03-30-vicuna/>
- Longpre S., Le Hou, Tu Vu, A. Webson, H. W. Chung, Yi Tay, D. Zhou, Q V. Le, B. Zoph, J. Wei and A. Roberts. 2023. "The flan collection: designing data and methods for effective instruction tuning." In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, 22631–48.
- Lyu, X., S. Grafberger, S. Biegel, S. Wei, M. Cao, S. Schelter, and C. Zhang. 2023. "Improving Retrieval-Augmented Large Language Models Via Data Importance Learning." *arXiv preprint arXiv:2307.03027*.
- MacDonald, B. A. 1991. "Instructable Systems." *Knowledge Acquisition* 3(4): 381–420.
- Manakul P., A. Liusie and M. J. Gale. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 9004–17.
- Mao J., X. Yang, X. Zhang, N. Goodman and J. Wu. 2022. "CLEVRER-Humans: Describing Physical and Causal Events the Human Way." *Advances in Neural Information Processing Systems*, 35: 7755–7768.
- Meade N., S. Gella, D. Hazarika, P. Gupta, Di Jin, S. Reddy, Y. Liu and D. Hakkani-Tur. 2023. Using In-Context Learning to Improve Dialogue Safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11882–910.
- Menick, J., M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, N. McAleese, et al. 2022. "Teaching Language Models to Support Answers with Verified Quotes." *arXiv preprint arXiv:2203.11147*.
- Ngo, R., L. Chan, and S. Mindermann. 2022. "The Alignment Problem from a Deep Learning Perspective." *arXiv preprint arXiv:2209.00626*.
- Penedo, G., Q. Malartic, D. Hesslow, R. Cojocar, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay. 2023. "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only." *arXiv preprint arXiv:2306.01116*.
- Perez E., S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese and G. Irving. 2022. "Red Teaming Language Models with Language Models." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–48.
- Perri, L. 2023. "4 Exciting New Trends in the Gartner Emerging Technologies Hype Cycle." Gartner. <https://www.gartner.com/en/articles/what-s-new-in-the-2023-gartner-hype-cycle-for-emerging-technologies>
- Petroni F., T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu and A. Miller. 2019. "Language Models as Knowledge Bases?" In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–73.
- Petroni F., A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard and V. Plachouras. 2021. "KILT: a Benchmark for Knowledge Intensive Language Tasks." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2523–44.
- Pustejovsky, J., and N. Krishnaswamy. 2020. "Neurosymbolic AI for Situated Language Understanding." In *ACS 2020-Annual Conference on Advances in Cognitive Systems*.
- Quach, K. 2023. "Google, You're Not Unleashing "Unproven" AI Medical Bots on Hospital Patients, Yeah?" *The Register*. [https://www.theregister.com/2023/08/08/google\\_senator\\_ai\\_health/](https://www.theregister.com/2023/08/08/google_senator_ai_health/)



- Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer." *The Journal of Machine Learning Research* 21(1): 5485–551.
- Rawte, V., S. Chakraborty, A. Pathak, A. Sarkar, S. M. Towhidul Islam Tonmoy, A. Chadha, A. Sheth, and A. Das. 2023. "The Troubling Emergence of Hallucination in Large Language Models-An Extensive Definition, Quantification, and Prescriptive Remediations." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2541–73.
- Rebedea, T., R. Dinu, M. N. Sreedhar, C. Parisien, and J. Cohen. 2023. "NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 431–445.
- Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery And Data Mining*, 1135–44.
- Roy K., Y. Zi, M. Gaur, J. Malekar, Q. Zhang, V. Narayanan and A. Sheth. 2023. "Process Knowledge-Infused Learning for Clinician-Friendly Explanations." In Proceedings of the AAAI Symposium Series. Association for the Advancement of Artificial Intelligence (AAAI). <https://doi.org/10.1609/aaais.v1i1.27494>
- Sarkar S., M. Gaur, L. K. Chen, M. Garg and B. Srivastava. 2023. "A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement." *Frontiers in Artificial Intelligence*, 6. 1229805.
- Scherrer, N., C. Shi, A. Feder, and D. M. Blei. 2023. "Evaluating the Moral Beliefs Encoded in LLMs." *arXiv preprint arXiv:2307.14324*. <https://neurips.cc/virtual/2023/poster/71831>
- Sellam, T., D. Das, and A. Parikh. 2020. "BLEURT: Learning Robust Metrics for Text Generation." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 788–7892.
- Shah, R., V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton. 2022. "Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals." *arXiv preprint arXiv:2210.01790*.
- Shah, R. S., F. Holt, S. A. Hayati, A. Agarwal, Y. C. Wang, R. E. Kraut, and D. Yang. 2022. "Modeling Motivational Interviewing Strategies on an Online Peer-To-Peer Counseling Platform." *Proceedings of the ACM on Human-Computer Interaction* 6(CSCW2): 1–24.
- Shen, L., L. Liu, H. Jiang, and S. Shi. 2022. "On the Evaluation Metrics for Paraphrase Generation." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3178–90.
- Shen, X., Z. Chen, M. Backes, Y. Shen, and Y. Zhang. 2023. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models." *arXiv preprint arXiv:2308.03825*.
- Sheth, A., M. Gaur, U. Kursuncu, and R. Wickramarachchi. 2019. "Shades of Knowledge-Infused Learning for Enhancing Deep Learning." *IEEE Internet Computing* 23(6): 54–63.
- Sheth, A., M. Gaur, K. Roy, and K. Faldu. 2021. "Knowledge-Intensive Language Understanding for Explainable AI." *IEEE Internet Computing* 25(5): 19–24.
- Sheth A., K. Roy and M. Gaur. 2023. *Neurosymbolic Artificial Intelligence (Why, What, and How)*. IEEE Intelligent Systems, pp.56-62.
- Shin, R. 2023. "Google Wants Its A.I. to Transform Health Care Next, As it Partners with the Mayo Clinic, Report Says." *Fortune*. <https://fortune.com/2023/07/10/google-ai-mayo-clinic-healthcare-med-palm-2-large-language-model/#>.
- Shiri A., K. Roy, A. Sheth, and M. Gaur. 2024. "L3 Ensembles: Life-long Learning Approach for Ensemble of Foundational Language Models." In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data. (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD'24)*, 592–94. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3632410.3632494>
- Slack, D., S. Krishna, H. Lakkaraju, and S. Singh. 2023. "Explaining Machine Learning Models with Interactive Natural Language Conversations Using TalkToModel." *Nature Machine Intelligence* 5: 1–11.
- So, D. R., W. Mañke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le. 2021. "Primer: Searching for Efficient Transformers for Language Modeling." *arXiv preprint arXiv:2109.08668*.
- Solaiman, I., Z. Talat, W. Agnew, L. Ahmad, D. Baker, S. L. Blodgett, A. Vassilev, et al. 2023. "Evaluating the Social Impact of Generative AI Systems in Systems and Society." *arXiv preprint arXiv:2306.05949*.
- Sun, J., C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. M. Ni, H.-Y. Shum, J. Guo. 2023. "Think-on-graph: Deep and Responsible Reasoning of Large Language Models with Knowledge Graph." *arXiv preprint arXiv:2307.07697*.
- Topp, C. W., S. D. Østergaard, S. Søndergaard, and P. Bech. 2015. "The WHO-5 Well-Being Index: A Systematic Review of the Literature." *Psychotherapy and Psychosomatics* 84(3): 167–76.
- Touvron H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. Singh Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, R. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov and T. Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models, *arXiv preprint arXiv:2307.09288*.
- Tyagi N., S. Sarkar and M. Gaur. 2023. "Leveraging Knowledge and Reinforcement Learning for Enhanced Reliability of Language Models." In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4320–24.
- Tyagi, N., A. Shiri, S. Sarkar, A. K. Umrawal, and M. Gaur. 2023. "Simple is Better and Large is Not Enough: Towards Ensembling of Foundational Language Models." *arXiv preprint arXiv:2308.12272*.
- Wang, P., L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, Z. Sui. 2023. "Large Language Models are Not Fair Evaluators." *arXiv preprint arXiv:2305.17926*.
- Wang X., J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery and D. Zhou. 2022. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." In *The Eleventh International Conference on Learning Representations*.



- Wang, Y., Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, A. Chowdhery, Y. Zhang. 2023. "PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization." *arXiv preprint arXiv:2306.05087*.
- Ward, C. 2023. "ADHD Test." Psych Central. <https://psychcentral.com/quizzes/adhd-quiz>
- Wei J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler and E.H. Chi. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.
- Yang, C., X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. 2023. "Large Language Models as Optimizers." *arXiv preprint arXiv:2309.03409*.
- Yang L., H. Chen, Z. Li, X. Ding and X. Wu. "Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling", *arXiv preprint arXiv:2306.11489*.
- Yao, S., D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. 2023. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." In Thirty-seventh Conference on Neural Information Processing Systems, *arXiv preprint arXiv:2305.10601*. <https://openreview.net/forum?id=5XclecxO1h>
- Yao S., J. Zhao, D. Yu, N. Du, I. Shafran, K R. Narasimhan and Y. Cao. 2022. "ReAct: Synergizing Reasoning and Acting in Language Models." In *The Eleventh International Conference on Learning Representations*.
- Yao, X., M. Mikhelson, S. C. Watkins, E. Choi, E. Thomaz, and K. de Barbaro. 2023. "Development and Evaluation of Three Chatbots for Postpartum Mood and Anxiety Disorders." *arXiv preprint arXiv:2308.07407*.
- Yin, W., J. Hay, and D. Roth. 2019. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3914–23.
- Yu, Y., J. Shen, T. Liu, Z. Qin, J. Nathan Yan, J. Liu, C. Zhang, and M. Bendersky. 2023. "Explanation-Aware Soft Ensemble Empowers Large Language Model In-Context Learning." *arXiv preprint arXiv:2311.07099*.
- Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT." In *International Conference on Learning Representations*, 2019.
- Zhang, Y., Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, S. Shi, et al. 2023. "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models." *arXiv preprint arXiv:2309.01219*.
- Zheng, L., W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, I. Stoica, et al. 2023. Accepted at NeurIPS 2023 Datasets and Benchmarks: <https://neurips.cc/virtual/2023/poster/73434>
- Ziems, C., J. Yu, Y. C. Wang, A. Halevy, and D. Yang. 2022. "The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, 3755–73. Long Papers.

**How to cite this article:** Gaur, M. and A. Sheth. 2024. "Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety." *AI Magazine* 45: 139–55. <https://doi.org/10.1002/aaai.12149>

## AUTHOR BIOGRAPHIES

**Manas Gaur** is an Assistant Professor in UMBC's Department of Computer Science and Engineering, where he heads the Knowledge-infused AI and Inference lab. He and Dr Sheth pioneered Knowledge-infused Learning, an innovative NeuroSymbolic AI approach for classification and generative AI. His research focuses on creating AI systems that are both explainable and safe, with a special emphasis on mental health applications. He was selected for AAAI New Faculty 2023.

**Amit Sheth** is the NCR Chair and Professor of Computer Science and Engineering at the University of South Carolina (USC). He founded the AI Institute, now with nearly 50 researchers, at USC. He is a Fellow of AAAI, ACM, AAAS, and IEEE. His major awards include IEEE CS Wallace McDowell and IEEE ICSVC Research Innovation.



## HIGHLIGHT

# Physical scene understanding

Jiajun Wu

Stanford University, Stanford, California, USA

### Correspondence

Jiajun Wu, Stanford University, Stanford, CA, USA.

Email: [jiajunwu@cs.stanford.edu](mailto:jiajunwu@cs.stanford.edu)

### Funding information

Stanford University; National Science Foundation; Office of Naval Research; Air Force Office of Scientific Research; Massachusetts Institute of Technology

### Abstract

Current AI systems still fail to match the flexibility, robustness, and generalizability of human intelligence: how even a young child can manipulate objects to achieve goals of their own invention or in cooperation, or can learn the essentials of a complex new task within minutes. We need AI with such embodied intelligence: transforming raw sensory inputs to rapidly build a rich understanding of the world for seeing, finding, and constructing things, achieving goals, and communicating with others. This problem of physical scene understanding is challenging because it requires a holistic interpretation of scenes, objects, and humans, including their geometry, physics, functionality, semantics, and modes of interaction, building upon studies across vision, learning, graphics, robotics, and AI. My research aims to address this problem by integrating bottom-up recognition models, deep networks, and inference algorithms with top-down structured graphical models, simulation engines, and probabilistic programs.

## INTRODUCTION

I am fascinated by how rich and flexible human intelligence is. From a quick glance at the scenes in Figure 1A, we effortlessly recognize the 3D geometry and texture of the objects within, reason about how they support each other, and when they move, track, and predict their trajectories. Stacking blocks, picking up fruits—we also plan and interact with scenes and objects in many ways.

My research goal is to build machines that see, interact with, and reason about the physical world just like humans. This problem of **physical scene understanding** involves the following three key topics that bridge research in computer science, AI, robotics, cognitive science, and neuroscience: **Perception** (Figure 1B): How can structured, physical object, and scene representations arise from raw, multimodal sensory input (e.g., videos, sound, tactile signals)? **Physical interactions** (Figure 1C): How can we build dynamics models that quickly adapt to

complex, stochastic real-world scenarios, and how can they contribute to planning and motor control? Modeling physical interactions helps robots build bridges from a single image and play challenging games such as Jenga. **Reasoning** (Figure 1D): How can physical models integrate structured, often symbolic, priors such as symmetry and repetition, and use them for commonsense reasoning?

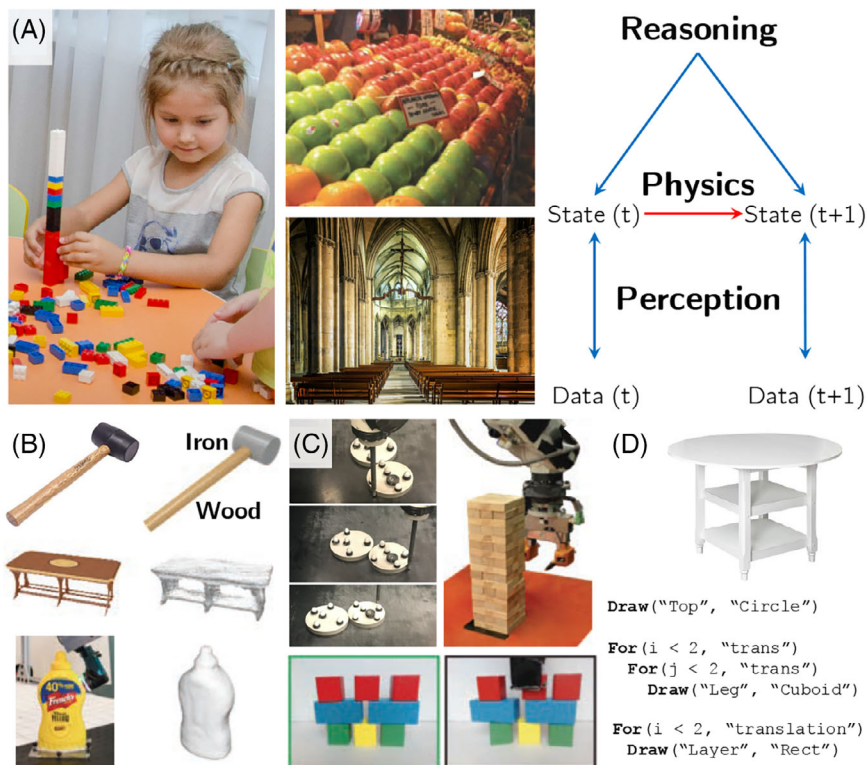
Physical scene understanding is challenging because it requires a holistic interpretation of scenes and objects, including their 3D geometry, physics, functionality, and modes of interaction, beyond the scope of a single discipline, such as computer vision. Structured priors and representations of the physical world are essential: we need proper representations and learning paradigms to build data-efficient, flexible, and generalizable intelligent systems that understand physical scenes.

My approach to constructing representations of the physical world is to integrate bottom-up recognition models, deep networks, and efficient inference algorithms

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.

**FIGURE 1** Physical scene understanding involves (I) perception, building physical object representations from multimodal data, (II) physical interaction, capturing scene dynamics for planning and control, and (III) commonsense reasoning, understanding high-level structured priors in objects and scenes.



with top-down, structured graphical models, simulation engines, and probabilistic programs. In my research, I develop and extend techniques in these areas (e.g., proposing new deep networks and physical simulators); I further explore innovative ways to combine them, building upon studies across vision, learning, graphics, and robotics. I believe that only by exploiting knowledge from all these areas can we build machines that have a human-like, physical understanding of complex, real-world scenes.

My research is also highly interdisciplinary: I build computational models with inspiration from human cognition, developmental psychology, neuroscience, robotics, and computational linguistics; I also explore how these models can, in turn, assist in solving tasks in these fields.

Below I describe my research experience and future plans on the three research topics.

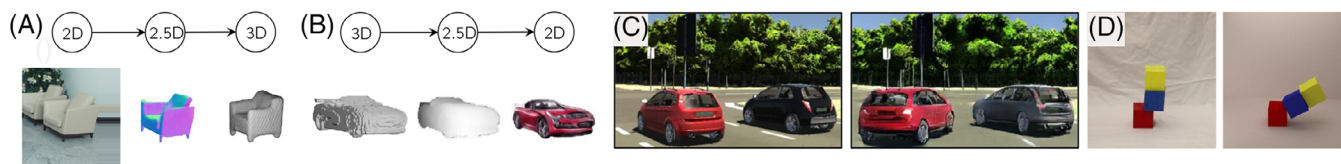
## LEARNING TO PERCEIVE THE PHYSICAL WORLD

Motivated by human perception—rich, complex, generalizable, learning much from little—my research on perception has been centered on building structured, object-based models to characterize the appearance and physics of daily objects and scenes. These models integrate bottom-up deep recognition models with top-down simulation engines, and they learn by perceiving and explaining the physical world just like humans.

**Seeing object intrinsics: shape, texture, and material.** Drawing inspiration from human perception and computer graphics, my colleagues and I have built object appearance models that learn to perceive object intrinsics, such as shape, texture, and material, from raw visual observations, and to leverage such information for synthesizing new objects in 2D and 3D. The core object representation builds upon a coherent understanding of its intrinsic properties, in addition to extrinsic properties such as pose.

Our research covers various components of the appearance model. On bottom-up recognition, we have developed a general pipeline for 3D shape reconstruction from a single color image (Wu, Wang, et al. 2017; Wu, Xue, et al. 2018) via modeling *intrinsic images*—depth, surface normals, and reflectance maps (Janner et al. 2017) (Figure 2A). Our research is inspired by the classic study on multi-stage human visual perception (Marr 1982) and has been extended to integrating learned priors of 3D shapes (i.e., “what shapes look like?”) for more realistic 3D reconstructions (Wu, Zhang, et al. 2018), to reconstructing object texture and material beyond geometry (Zhang et al. 2023), and to tackling cases where the object in the image is not from the training categories (Zhang et al. 2018).

Complementary to these bottom-up recognition models, we have also explored learning top-down graphics engines directly. We proposed 3D generative adversarial networks and point-voxel diffusion, among the first to apply generative-adversarial learning and diffusion to 3D shapes for unconditional shape synthesis (Wu, Zhang,



**FIGURE 2** Learning to see shapes, texture, and physics. (A) Reconstructing 3D shapes from a single color image via 2.5D sketches (Wu, Wang, et al. 2017; Wu, Zhang, et al. 2018; Zhang et al. 2018; Janner et al. 2017). (B) Generative modeling of 3D shapes and 2D images via a disentangled representation for object geometry, viewpoint, and texture (Wu, Zhang, et al. 2016; Zhu et al. 2018; Chan et al. 2021; Zhou, Du, and Wu 2021; Zhang et al. 2023). (C) 3D-aware representations for objects and scenes (Wu, Tenenbaum, and Kohli 2017; Yao et al. 2018; Yu, Guibas, and Wu 2022; Yu, Agarwala, et al. 2023; Tian et al. 2023; Yu, Guo, et al. 2023). (D) Part-based object representations for its geometry and physics (Wu, Lim, et al. 2016; Wu, Lu, et al. 2017; Wu et al. 2015; Liu et al. 2018; Xu et al. 2019).

et al. 2016; Zhou, Du, and Wu 2021). These papers are influential; many other researchers have built on them. We have later extended the model as visual object networks (Zhu et al. 2018) and periodic implicit GANs (pi-GANs) (Chan et al. 2021), which synthesize object shape and texture simultaneously, enforcing various consistencies with a distributed representation for object shape, 2.5D sketches, viewpoint, and texture (Figure 2B). We have generalized our models to scenes (Wu, Tenenbaum, and Kohli 2017; Yao et al. 2018; Yu, Guibas, and Wu 2022; Yu, Agarwala, et al. 2023), recovering structured scene representations that not only capture object shape and texture but enable 3D-aware scene manipulation (Figure 2C).

**Seeing physics.** Beyond object appearance, the intuition of object physics assists humans in scene understanding (Battaglia, Hamrick, and Tenenbaum 2013). We have developed computational models that learn to infer object physics directly from visual observations (Wu, Lim, et al. 2016; Wu et al. 2015). Our research on visual intuitive physics is the first in the computer vision community and has since led to many follow-up studies (Fragkiadaki et al. 2016; Mottaghi et al. 2016).

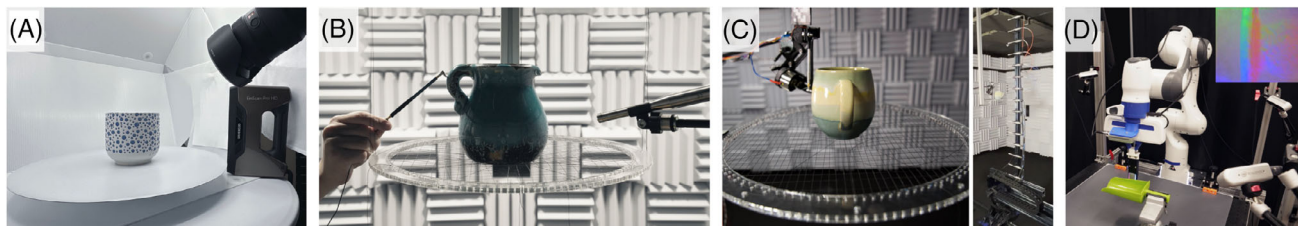
The Galileo model (Wu et al. 2015) marries a physics engine with deep recognition nets to infer physical object properties (e.g., mass, friction). With an embedded physical simulator, the Galileo model discovers physical properties simply by watching objects move in unlabeled videos; it also predicts how they interact based on the inferred physical properties. The model was tested on a real-world video dataset, Physics 101 (Wu, Lim, et al. 2016), of 101 objects that interact in various physical events.

I have also worked on integrating geometry and physics perception (Figure 2D). For example, in visual de-animation (VDA) (Wu, Lu, et al. 2017), our model learns to jointly infer physical world states and simulate scene dynamics, integrating both a physics engine and a graphics engine. In physical primitive decomposition (PPD) (Liu et al. 2018), we decompose an object into parts with distinct geometry and physics, by learning to explain both the object's appearance and its behaviors in physical events. In dynamics-augmented neural objects

(DANO) (Le Cleac'h et al. 2023), we enhance objects parametrized by neural implicit representations with their physical properties identified from raw observations; we then use such dynamic objects for future prediction. We have also extended these models to complex indoor scenes, exploiting stability for more accurate 3D scene parsing (Du et al. 2018).

**Multimodal perception.** Humans see, hear, and feel, perceiving the world through fusing multisensory signals. These signals play complementary roles: we see object shape and texture through vision, hear their material through sound, and feel their surface details through touch. In computer science, however, most recognition models and simulation engines primarily focus on visual data. Building upon techniques from the graphics community, we have been building generative audio-visual engines and using them for cross-modal perception (Zhang, Li, et al. 2017; Zhang, Wu, et al. 2017): how much do we know about objects from videos, and how much from audio? Our recent work includes developing a differentiable simulation model of impact sounds (Clarke et al. 2021) and building a benchmark for object impact sound fields (Clarke et al. 2023). Beyond auditory signals, we have also explored the integration of tactile signals with vision for better shape perception and reconstruction (Wang et al. 2018), and the integration of visual, auditory, and tactile information for robotic manipulation (Li et al. 2022).

In the past few years, we have also been developing a large-scale, multimodal, object-centric benchmark, ObjectFolder (Figure 3). It models the multisensory behaviors of both neural and real objects: it first includes 1000 neural objects in the form of implicit neural representations with simulated multisensory data (Gao et al. 2021, 2022); it also contains the multisensory measurements for 100 real-world household objects, based on a newly designed pipeline for collecting 3D meshes, videos, impact sounds, and tactile readings of real-world objects (Gao et al. 2023). ObjectFolder also has a standard benchmark suite of 10 tasks for multisensory object-centric learning, centered on object recognition, reconstruction, and manipulation



**FIGURE 3** Multimodal perception (Gao et al. 2021, 2022, 2023; Clarke et al. 2021, 2023; Li et al. 2022). Visual (A): We use a scanner, a turntable, and a lightbox to acquire object geometry and texture. Auditory (B, C): We strike objects at precise points using an impact hammer, either by hand (B) or by a robot (C), recording the sound with a microphone array on a rotating gantry. The object is resting on a compliant mesh inside an acoustically treated room. Tactile (D): A robot presses a tactile sensor on the object with GelSight (Yuan, Dong, and Adelson 2017).



**FIGURE 4** Physical models for future prediction and control. (A) Modeling visual dynamics allows us to generate multiple possible future frames from a single image (Xu et al. 2019; Xue et al. 2016). (B) Learned dynamics models support controlling soft robots Hu et al. (2019) and fluids Deng et al. (2023). (C) They also enable long-term manipulation of deformable objects and liquids (Li, Wu, Tedrake, et al. 2019; Shi et al. 2022, 2023). (D) We have developed a hybrid model that captures object-based dynamics by integrating analytical models and neural nets. It assists the robot in accomplishing a highly underactuated task: pushing the right disk to the target (green) by only interacting with the left disk (Ajay et al. 2019, 2018).

with sight, sound, and touch. We have open-sourced both the datasets and the benchmark suite to catalyze and enable new research on multisensory object-centric learning in computer vision, robotics, and beyond (Gao et al. 2023).

## PHYSICAL MODELS FOR REAL-WORLD INTERACTIONS

Beyond learning object-centric models from raw observations by inverting simulation engines, my research also includes learning to approximate simulation engines (forward models) themselves. Based on target domains and applications, my colleagues and I have explored building physical models in various forms—image-based, object-based, and particle-based; analytical, neural, and hybrid—and have demonstrated their power in challenging, highly underactuated control tasks (Figure 4).

Compared with off-the-shelf simulators, a learned dynamics simulator flexibly adapts to novel environments and captures stochasticity in scene dynamics. Our visual dynamics model demonstrates this in the pixel domain, where it learns to synthesize multiple possible future

frames from a single color image by automatically discovering independent movable parts and their motion distributions (Xue et al. 2016) (Figure 4A). Our paper was among the first to consider uncertainty in the area of visual prediction. We have later extended the model to additionally capture the hierarchical structure among object parts (Xu et al. 2019).

Modeling dynamics directly in the pixel space is universal but challenging due to the entanglement of physics and graphics; an alternative is to separate perception from dynamics modeling and learn dynamics from object states. Our work along this line has shown that a model that learns to approximate object dynamics can be useful for planning (Janner et al. 2019), generalize to scenarios where only partial observations are available (Li, Wu, Zhu, et al. 2019), and discover physical object properties without supervision (Zheng et al. 2018; Le Cleac'h et al. 2023). We have further extended our model to particle-based representations so that it can characterize the dynamics of soft robots (Hu et al. 2019), fluids (Deng et al. 2023) (Figure 4B), and scenes with complex interactions among rigid bodies, deformable shapes, and liquids (Li et al. 2020; Li, Wu, Tedrake, et al. 2019; Shi et al. 2023) (Figure 4C).





We have also explored the idea of learning a hybrid dynamics model, augmenting analytical physics engines with neural dynamics models (Ajay et al. 2018) (Figure 4D). Such a hybrid system achieves the best of both worlds: it performs better, captures uncertainty in data, learns efficiently from limited annotations, and generalizes to novel shapes and materials. The paper was selected as the Best Paper on Cognitive Robotics at the premier robotics conference (IROS 2018).

These dynamics models can be used in various control tasks: they help solve highly underactuated control problems (pushing disk A, which in turn pushes disk B to the target position) (Ajay et al. 2019), to control and co-design soft robots (Hu et al. 2019), to manipulate fluids and rigid bodies on a robot (Li, Wu, Tedrake, et al. 2019), to interact with plasticine to make complex shapes in multiple steps (Shi et al. 2022, 2023), and to interact and play games such as Jenga that involve complex frictional micro-interactions (Fazeli et al. 2018).

## STRUCTURED PRIORS FOR COMMONSENSE REASONING

The physical world is rich but structured: natural objects and scenes are compositional (scenes are made of objects which, in turn, are made of parts); they often have program-like structures (objects are symmetric and made of evenly spaced repetitive parts). My colleagues and I have been exploring ways to bridge structured, often symbolic, priors into powerful deep recognition models. In previous sections, we have seen perception models that invert simulation engines and physical dynamics models that approximate simulation engines themselves. Here, we move one step further to learn the representation priors these simulation engines have—why they represent the world in the way they currently are.

A test of these neuro-symbolic representations is how well they support solving various reasoning tasks such as analogy making and question answering. Our work has demonstrated that when combined with deep visual perception modules, a symbolic reasoning system achieves impressive performance on visual reasoning benchmarks (Yi et al. 2018), outperforming end-to-end trained neural models. We have also extended it to jointly learn visual concepts (e.g., colors, shapes) and their correspondence with words from natural supervision (question-answer pairs) through curriculum learning (Mao et al. 2019), without human annotations.

Beyond static images, we have integrated neuro-symbolic representations with learned object-based dynamics models for temporal and causal reasoning

on videos. On our newly proposed video reasoning benchmark, our model performs significantly better in answering all four types of questions: descriptive (e.g., “what color”), explanatory (“what’s responsible for”), predictive (“what will happen next”), and counterfactual (“what if”) (Yi et al. 2020; Chen et al. 2021). Similar ideas have been applied to visual grounding in 3D scenes (Hsu, Mao, and Wu 2023), human motion understanding (Endo et al. 2023), and robotic manipulation (Wang et al. 2023).

Learning symbolic structure is closely coupled with program synthesis. In particular, our recent work has made progress on the problem of inferring programs as a novel representation for shapes (Tian et al. 2019; Deng et al. 2022), scenes (Liu et al. 2019), and human motion (Kulal et al. 2021, 2022). This marks the start of our exploration of wiring highly structured, hierarchical priors into learning representations for physical scene understanding.

## NEXT STEPS

With big data, large computing resources, and advanced learning algorithms, the once separated areas across computer science (vision, learning, symbolic reasoning, NLP, rule learning and program induction, planning, and control) have begun to reintegrate. We should now take a more integrative view of these areas and actively explore their interactions for a more general AI landscape.

One such direction is to achieve a more fundamental integration of perception, reasoning, and planning. Although most computational models have treated them as disjoint modules, we observe that having them communicate with each other facilitates the model design and leads to better performance (Janner et al. 2019; Veerapaneni et al. 2019). For example, AI researchers have been integrating perception and planning in belief space (Kaelbling and Lozano-Pérez 2013)—our belief of the partially observable, uncertain world states. Building upon these insightful ideas, I would like to explore interactive perception by integrating both classic and modern AI tools: probabilistic inference for managing uncertainty; causal and counterfactual reasoning in generative models for explainability, imagination, and planning; and hierarchical inference for learning to learn, so knowledge builds progressively. In addition, discovering the cognitive and neural basis of perception, reasoning, and planning will be of significant value to understanding human intelligence.

Another direction is to integrate symbolic priors with deep representation learning via program synthesis for concept and structure discovery. Neuro-symbolic methods enjoy both the recognition power from neural nets and the combinatorial generalization from symbolic structure;

therefore, they have great potential in scaling up current intelligent systems to large-scale, complex physical scenes in real life, for which pure bottom-up, data-driven models cannot work well due to the exponentially increasing complexity. Our research has shown that they can learn to discover concepts and answer questions using only natural supervision (question–answer pairs) as humans (Mao et al. 2019; Yi et al. 2018; Hsu, Mao, and Wu 2023). In the future, I would like to explore the use of symbolic languages for knowledge representation and abstraction, and their integration with deep networks for flexible physical scene understanding and interaction.

Beyond physical objects and scenes, I want to build computational models that understand an agent's goals, beliefs, intentions, and theory of mind and use such knowledge for planning and problem-solving, drawing inspiration from intuitive psychology. While we have been inferring physical object properties from interactions, can we also build computational models that, just like 10-month-old infants (Liu et al. 2017), infer object values in agents' beliefs from their behaviors? Research in this direction would benefit the development of human-like and human-centered autonomous systems.

More generally, I want to connect computer science with other disciplines, such as cognitive science, neuroscience, social science, linguistics, and mechanical engineering. Research in cognitive science and neuroscience has been offering intuitions for AI researchers for decades; now, we are entering a new stage where contemporary research in intelligent systems or computer science, in general, may help us better understand human intelligence (Fischer et al. 2016; Yamins et al. 2014). Our research has suggested that computational models that combine bottom-up neural recognition networks and top-down simulation engines shed light on understanding cognitive and neural processes in the brain (Yildirim et al. 2019; Zhang et al. 2016). Much more work needs to be done in these areas. With the right integration of probabilistic inference methods, deep learning, and generative models, we can build more powerful computational models for both neural activities and cognitive, behavioral data. The same applies to developmental psychology. I want to compare and contrast human and artificial intelligence in understanding *core knowledge*—knowledge about object permanence, solidity, continuity, and containment, and concepts such as gravity and momentum (Spelke 2000). This interdisciplinary research deepens our understanding of multiple research areas and suggests future research topics.

We are in a unique and exciting time: the development of data, hardware, and algorithms (e.g., deep networks, graphical models, probabilistic programs) has enabled more flexible and expressive computational models. For the next decade, I believe building structured foundation

models for machine physical scene understanding, as well as investigating its connection with perception, reasoning, and interaction, will be valuable and essential for developing computational systems that contribute to broad fundamental and practical research across disciplines.

## CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.

## ORCID

Jiajun Wu  <https://orcid.org/0000-0002-4176-343X>

## REFERENCES

- Ajay, Anurag, Maria Bauza, Jiajun Wu, Nima Fazeli, Joshua B. Tenenbaum, Alberto Rodriguez, and Leslie P. Kaelbling. 2019. "Combining Physical Simulators and Object-Based Networks for Control." In *IEEE International Conference on Robotics and Automation (ICRA)*.
- Ajay, Anurag, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P. Kaelbling, Joshua B. Tenenbaum, and Alberto Rodriguez. 2018. "Augmenting Physical Simulators with Stochastic Neural Networks: Case Study of Planar Pushing and Bouncing." In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.
- Battaglia, Peter W., Jessica B. Hamrick, and Joshua B. Tenenbaum. 2013. "Simulation as an Engine of Physical Scene Understanding." *Proceedings of the National Academy of Sciences (PNAS)* 110(45): 18327–32.
- Chan, Eric R., Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. "pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, Zhenfang, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B. Tenenbaum, and Chuang Gan. 2021. "Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning." In *International Conference on Learning Representations (ICLR)*.
- Clarke, Samuel, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L. James, and Jiajun Wu. 2023. "RealImpact: A Dataset of Impact Sound Fields for Real Objects." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Clarke, Samuel, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. 2021. "DiffImpact: Differentiable Rendering and Identification of Impact Sounds." In *Conference on Robot Learning (CoRL)*.
- Deng, Boyang, Sumith Kulal, Zhengyang Dong, Congyue Deng, Yonglong Tian, and Jiajun Wu. 2022. "Unsupervised Learning of Shape Programs with Repeatable Implicit Parts." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Deng, Yitong, Hong-Xing Yu, Jiajun Wu, and Bo Zhu. 2023. "Learning Vortex Dynamics for Fluid Inference and Prediction." In *International Conference on Learning Representations (ICLR)*.
- Du, Yilun, Zhijian Liu, Hector Basevi, Ales Leonardis, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2018. "Learning to Exploit Stability for 3D Scene Parsing." In *Advances in Neural Information Processing Systems (NeurIPS)*.



- Endo, Mark, Joy Hsu, Jiaman Li, and Jiajun Wu. 2023. "Motion Question Answering Via Modular Motion Programs." In *International Conference on Machine Learning (ICML)*.
- Fazeli, Nima, Miquel Oller, Jiajun Wu, Zheng Wu, Joshua B. Tenenbaum, and Alberto Rodriguez. 2018. "See, Feel, Act: Learning Complex Manipulation Skills with Causal Structure and Multi-Sensory Fusion." *Science Robotics* 4(26): eaav3123.
- Fischer, Jason, John G. Mikhael, Joshua B. Tenenbaum, and Nancy Kanwisher. 2016. "Functional Neuroanatomy of Intuitive Physical Inference." *Proceedings of the National Academy of Sciences (PNAS)* 113(34): E5072–81.
- Fragkiadaki, Katerina, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. 2016. "Learning Visual Predictive Models of Physics for Playing Billiards." In *International Conference on Learning Representations (ICLR)*.
- Gao, Ruohan, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. 2021. "ObjectFolder: A Dataset of Objects with Implicit Visual, Auditory, and Tactile Representations." In *Conference on Robot Learning (CoRL)*.
- Gao, Ruohan, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. 2023. "The Object-Folder Benchmark: Multisensory Learning with Neural and Real Objects." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gao, Ruohan, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. 2022. "Object-Folder 2.0: A Multisensory Object Dataset for sim2real Transfer." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hsu, Joy, Jiayuan Mao, and Jiajun Wu. 2023. "NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hu, Yuanming, Jiancheng Liu, Andrew Spielberg, Joshua B. Tenenbaum, William T. Freeman, Jiajun Wu, Daniela Rus, and Wojciech Matusik. 2019. "ChainQueen: A Real-Time Differentiable Physical Simulator for Soft Robotics." In *IEEE International Conference on Robotics and Automation (ICRA)*.
- Janner, Michael, Sergey Levine, William T. Freeman, Joshua B. Tenenbaum, Chelsea Finn, and Jiajun Wu. 2019. "Reasoning about Physical Interactions with Object-Oriented Prediction and Planning." In *International Conference on Learning Representations (ICLR)*.
- Janner, Michael, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim, and Joshua B. Tenenbaum. 2017. "Self-Supervised Intrinsic Image Decomposition." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Pack Kaelbling, Leslie, and Tomás Lozano-Pérez. 2013. "Integrated Task and Motion Planning in Belief Space." *The International Journal of Robotics Research (IJRR)* 32(9-10): 1194–227.
- Kulal, Sumith, Jiayuan Mao, Alex Aiken, and Jiajun Wu. 2021. "Hierarchical Motion Understanding Via Motion Programs." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kulal, Sumith, Jiayuan Mao, Alex Aiken, and Jiajun Wu. 2022. "Programmatic Concept Learning for Human Motion Description and Synthesis." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Le Cleac'h, Simon, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. 2023. "Differentiable Physics Simulation of Dynamics-Augmented Neural Objects." *IEEE Robotics and Automation Letters (RA-L)*.
- Li, Hao, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A. Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. 2022. "See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation." In *Conference on Robot Learning (CoRL)*.
- Li, Yunzhu, Toru Lin, Kexin Yi, Daniel Bear, Daniel L. K. Yamins, Jiajun Wu, Joshua B. Tenenbaum, and Antonio Torralba. 2020. "Visual Grounding of Learned Physical Models." In *International Conference on Machine Learning (ICML)*.
- Li, Yunzhu, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. 2019. "Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids." In *International Conference on Learning Representations (ICLR)*.
- Li, Yunzhu, Jiajun Wu, Jun-Yan Zhu, Joshua B. Tenenbaum, Antonio Torralba, and Russ Tedrake. 2019. "Propagation Networks for Model-Based Control Under Partial Observation." In *IEEE International Conference on Robotics and Automation (ICRA)*.
- Liu, Shari, Tomer D. Ullman, Joshua B. Tenenbaum, and Elizabeth S. Spelke. 2017. "Ten-Month-Old Infants Infer the Value of Goals from the Costs of Actions." *Science* 358(6366): 1038–41.
- Liu, Yunchao, Zheng Wu, Daniel Ritchie, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "Learning to Describe Scenes with Programs." In *International Conference on Learning Representations (ICLR)*.
- Liu, Zhijian, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2018. "Physical Primitive Decomposition." In *European Conference on Computer Vision (ECCV)*.
- Mao, Jiayuan, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision." In *International Conference on Learning Representations (ICLR)*.
- Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman and Company.
- Mottaghi, Roozbeh, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. 2016. "Newtonian Scene Understanding: Unfolding the Dynamics of Objects in Static Images." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shi, Haochen, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. 2023. "RoboCook: Long-horizon Elasto-Plastic Object Manipulation with Diverse Tools." In *Conference on Robot Learning (CoRL)*.
- Shi, Haochen, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. 2022. "RoboCraft: Learning to see, Simulate, and Shape Elasto-Plastic Objects with Graph Networks." In *Robotics: Science and Systems (RSS)*.
- Spelke, Elizabeth S. 2000. "Core Knowledge." *American Psychologist* 55(11): 1233.
- Tian, Stephen, Yancheng Cai, Hong-Xing Yu, Sergey Zakharov, Katherine Liu, Adrien Gaidon, Yunzhu Li, and Jiajun Wu. 2023. "Multi-Object Manipulation Via Object-Centric Neural Scattering Functions." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tian, Yonglong, Andrew Luo, Xingyuan Sun, Kevin Ellis, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "Learning to Infer and Execute 3D Shape Programs." In *International Conference on Learning Representations (ICLR)*.

- Veerapaneni, Rishi, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua B. Tenenbaum, and Sergey Levine. 2019. "Entity Abstraction in Visual Model-Based Reinforcement Learning." In *Conference on Robot Learning (CoRL)*.
- Wang, Renhao, Jiayuan Mao, Joy Hsu, Hang Zhao, Jiajun Wu, and Yang Gao. 2023. "Programmatically Grounded, Compositionally Generalizable Robotic Manipulation." In *International Conference on Learning Representations (ICLR)*.
- Wang, Shaoxiong, Jiajun Wu, Xingyuan Sun, Wenzhen Yuan, William T. Freeman, Joshua B. Tenenbaum, and Edward H. Adelson. 2018. "3D Shape Perception from Monocular Vision, Touch, and Shape Priors." In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.
- Wu, Jiajun, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. 2016. "Physics 101: Learning Physical Object Properties from Unlabeled Videos." In *British Machine Vision Conference (BMVC)*.
- Wu, Jiajun, Erika Lu, Pushmeet Kohli, William T. Freeman, and Joshua B. Tenenbaum. 2017. "Learning to See Physics Via Visual De-Animation." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Wu, Jiajun, Joshua B. Tenenbaum, and Pushmeet Kohli. 2017. "Neural Scene De-Rendering." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wu, Jiajun, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T. Freeman, and Joshua B. Tenenbaum. 2017. "MarrNet: 3D Shape Reconstruction Via 2.5D Sketches." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Wu, Jiajun, Tianfan Xue, Joseph J. Lim, Yuandong Tian, Joshua B. Tenenbaum, Antonio Torralba, and William T. Freeman. 2018. "3D Interpreter Networks for Viewer-Centered Wireframe Modeling." *International Journal of Computer Vision (IJCV)* 126(9): 1009–26.
- Wu, Jiajun, Ilker Yildirim, Joseph J. Lim, William T. Freeman, and Joshua B. Tenenbaum. 2015. "Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Wu, Jiajun, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. 2016. "Learning a Probabilistic Latent Space of Object Shapes Via 3D Generative-Adversarial Modeling." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Wu, Jiajun, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T. Freeman, and Joshua B. Tenenbaum. 2018. "Learning Shape Priors for Single-View 3D Completion and Reconstruction." In *European Conference on Computer Vision (ECCV)*.
- Xu, Zhenjia, Zhijian Liu, Sun Chen, Kevin Murphy, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "Modeling Parts, Structure, and System Dynamics Via Predictive Learning." In *International Conference on Learning Representations (ICLR)*.
- Xue, Tianfan, Jiajun Wu, Katherine Bouman, and William T. Freeman. 2016. "Visual Dynamics: Probabilistic Future Frame Synthesis Via Cross Convolutional Networks." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex." *Proceedings of the National Academy of Sciences (PNAS)* 111(23): 8619–24.
- Yao, Shunyu, Tzu-Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, William T. Freeman, and Joshua B. Tenenbaum. 2018. "3D-Aware Scene Manipulation Via Inverse Graphics." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yi, Kexin, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. "CLEVRER: Collision Events for Video Representation and Reasoning." In *International Conference on Learning Representations (ICLR)*.
- Yi, Kexin, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yildirim, Ilker, Jiajun Wu, Nancy Kanwisher, and Joshua B. Tenenbaum. 2019. "An Integrative Computational Architecture for Object-Driven Cortex." *Current Opinion in Neurobiology* 55: 73–81.
- Yu, Hong-Xing, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. 2023. "Accidental Light Probes." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yu, Hong-Xing, Leonidas J. Guibas, and Jiajun Wu. 2022. "Unsupervised Discovery of Object Radiance Fields." In *International Conference on Learning Representations (ICLR)*.
- Yu, Hong-Xing, Michelle Guo, Alireza Fathi, Yen-Yu Chang, Eric Ryan Chan, Ruohan Gao, Thomas Funkhouser, and Jiajun Wu. 2023. "Learning Object-Centric Neural Scattering Functions for Free-Viewpoint Relighting and Scene Composition." *Transactions on Machine Learning Research (TMLR)*.
- Yuan, Wenzhen, Siyuan Dong, and Edward H. Adelson. 2017. "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force." *Sensors*, 17(12): 2762.
- Zhang, Renqiao, Jiajun Wu, Chengkai Zhang, William T. Freeman, and Joshua B. Tenenbaum. 2016. "A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding." In *Annual Meeting of the Cognitive Science Society (CogSci)*.
- Zhang, Xiuming, Zhoutong Zhang, Chengkai Zhang, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2018. "Learning to Reconstruct Shapes from Unseen Categories." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhang, Yunzhi, Shangzhe Wu, Noah Snavely, and Jiajun Wu. 2023. "Seeing a Rose in Five Thousand Ways." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhang, Zhoutong, Qiuqia Li, Zhengjia Huang, Jiajun Wu, Joshua B. Tenenbaum, and William T. Freeman. 2017. "Shape and Material from Sound." In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhang, Zhoutong, Jiajun Wu, Qiuqia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. 2017. "Generative Modeling of Audible Shapes for Object Perception." In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Zheng, David, Vinson Luo, Jiajun Wu, and Joshua B. Tenenbaum. 2018. "Unsupervised Learning of Latent Physical Properties Using Perception-Prediction Networks." In *Conference on Uncertainty in Artificial Intelligence (UAI)*.



- Zhou, Linqi, Yilun Du, and Jiajun Wu. 2021. “3D Shape Generation and Completion Through Point-Voxel Diffusion.” In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Zhu, Jun-Yan, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. 2018. “Visual Object Networks: Image Generation with Disentangled 3D Representations.” In *Advances in Neural Information Processing Systems (NeurIPS)*.

**How to cite this article:** Wu, J. 2024. “Physical scene understanding.” *AI Magazine* 45: 156–64.  
<https://doi.org/10.1002/aaai.12148>

## AUTHOR BIOGRAPHY

**Jiajun Wu** is an Assistant Professor of Computer Science and, by courtesy, of Psychology at Stanford University, working on computer vision, machine learning, and computational cognitive science. Before joining Stanford, he was a Visiting Faculty Researcher at Google Research. He received his PhD in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology. Wu’s research has been recognized through the Young Investigator Programs (YIP) by ONR and by AFOSR, paper awards and finalists at ICCV, CVPR, SIGGRAPH Asia, CoRL, and IROS, dissertation awards from ACM, AAAI, and MIT, the 2020 Samsung AI Researcher of the Year, and faculty research awards from J.P. Morgan, Samsung, Amazon, and Meta.



## COLUMN

# Generative AI: An AI paradigm shift in the making?

Risto Miikkulainen<sup>1,2</sup>

<sup>1</sup>The University of Texas at Austin,  
Austin, Texas, USA

<sup>2</sup>Cognizant AI Labs, San Francisco,  
California, USA

### Correspondence

Risto Miikkulainen, The University of  
Texas at Austin, Austin, TX, USA.

Email: [risto@cs.utexas.edu](mailto:risto@cs.utexas.edu)

### Abstract

It is sometimes difficult to evaluate progress in Generative AI, that is, image generation and large language models. This may be because they represent a paradigm shift in AI, and the traditional ways of developing, evaluating, understanding, and deploying AI systems no longer apply. Instead, we need to develop new such approaches, possibly by extending those currently in use in cognitive neuroscience and psychology. In this manner, a new AI paradigm can be created, providing a significant leap in AI research and practice.

In 1962, science philosopher Thomas Kuhn proposed that progress in mature sciences proceeds through paradigm shifts (Kuhn 1962). In mature sciences such as physics, chemistry, and biology, scientists agree on problems, approaches, and solutions, and substantial incremental progress is made over a long period of time. However, some pesky problems resist—until a fundamental change in thinking makes them no problems at all. Much of the field is built on a new foundation, with entirely new problems, approaches, and solutions. An often-quoted example is how the theory of relativity replaced Newtonian mechanics, with the speed of light as a fundamental concept.

Artificial intelligence as a discipline is only decades old, and although there have been several innovations, and the field has had its ups and downs, there have hardly been revolutions that could be called a paradigm shift. Until now. The generative AI (GenAI) models that have emerged in the last few years require a fundamentally different way of thinking about AI—how it is developed, evaluated, understood, and deployed. As a result, we need to build a new scientific field of AI.

More specifically, much of what AI practitioners used to agree was a productive way of doing AI no longer applies. For instance, much of AI was based on explicit, transparent, and interpretable mechanisms such as rules, logic, and dynamical systems. Such understanding is no longer pos-

sible with GenAI, and may never be. The models are too large, interactive, nonlinear, and opaque, in the same way brains are. Importantly, it is not necessary to understand brains entirely to be able to use them—similarly, GenAI may be useful even if we do not understand it fully.

Even if we cannot understand them, surely we can still measure how well they perform, as we always have done with AI systems? Not entirely. Whereas much of past AI was designed to be measured in a particular task or dataset, GenAI models aim at general performance that is not defined by any task in particular. It is of course possible to pose a problem and measure how well GenAI does on it, but such a measurement only scratches the surface of what GenAI does. GenAI systems take on different roles in different interactions, and thus perform a wide variety of tasks. Such tasks can be incompletely specified and open-ended, such as evaluating business plans, writing poetry, or recommending ways to make good meals out of what you happen to have in the fridge. For many such tasks, performance cannot be readily measured. You recognize good performance when you see it.

More specifically, the methodology we have used for decades of separating the training and the testing of a model, thus estimating its likely future performance, no longer applies. It did still apply to the earlier work on deep learning models: Such models were trained on increasingly

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AI Magazine* published by John Wiley & Sons Ltd on behalf of Association for the Advancement of Artificial Intelligence.



large datasets that were labeled, thus making it possible to estimate classification accuracy on an unseen set. Now the models are trained on as much data as possible, and it is no longer feasible to separate training and testing examples cleanly. Moreover, classification accuracy is not even a task we care about in training GenAI models, and interestingly, neither is even the performance in the actual training task (e.g., perplexity in language modeling). They were useful metrics within the old paradigm, but no longer measure what we want to do with GenAI. Training GenAI is more like developing human experts (e.g., those in medicine, music, and sports): Expertise requires repetitive and comprehensive practice well beyond mastering any conceivable training course.

In prior AI, the interaction with the system is scripted and specific; the question is posed in a certain format and the response is crisp and unequivocal. In contrast, GenAI systems can be queried with different contexts, settings, and prompts to get different answers at different times. Often the first answer is incomplete or incorrect, but further interaction gets to the right piece of knowledge. The model has that knowledge, but it is tricky to get to it. Again, humans perform much the same way in many cases. Trying to remember a name, solve a mathematical problem, or write a lyric that rhymes, may take several tries and techniques that get to the answer through a process.

Above all, GenAI systems are individuals, not mechanistic devices that can be rebuilt multiple times. It is of course possible to take the code and parameters and form an exact copy, and in principle, any computational experiment can be replicated with the same data and compute. But in practice, constructing these models is so costly that it is infeasible to construct the same system multiple times. Even in just a few months, the data have changed, the computational resources are different, new techniques and hyperparameter settings promise to be better, and any new construction would use the best possible settings available at the time. As a result, each such model is a one-off—not unlike a human that is shaped by their experience, in addition to genetics. We have to learn to do science with a sample of one.

So what does future AI research and practice look like? How can we develop new best practices to take advantage of the paradigm shift? AI will be everywhere, not just limited to a few specific tasks and problems. We need to learn to give them a chance, but not take their solutions at face value. The relationship should be similar to what we already have with many human experts such as an investment advisor, a lawyer, or a doctor. They are experts and know a lot about what they are doing, but they are not infallible, and sometimes it makes sense to get a second opinion. They will need to be integrated into everyday workflows similarly to human experts. This perspective

can also be used as a guideline for regulating AI—they can be evaluated and licensed similarly to human experts.

Building such AI systems requires new research, focusing on new questions and methods. How to get a good answer from a GenAI system will be an important area of its own. How to determine the epistemological status of the answer is another one: Is it something the system knows with confidence, or made up because it is likely, or generated even though it may not have ever existed before? Devising ways for GenAI to use computational tools to expand its capability will be an important extension, similar to effective human experts. We need to expand GenAI with metacognitive abilities as well, allowing them to introspect, deliberate, doubt, argue, and monitor and evaluate their progress. Such abilities allow them to create more complex solutions but also allow them to stay away from harm and negative side effects.


Efforts to approach these issues are already underway. Methods are being developed for enhancing GenAI with chain-of-thought interactions and orchestrating them with other AI systems to achieve more complex reasoning (Patil et al. 2023; Sumers et al. 2023; Wei et al. 2022). Methods motivated by neuroimaging may help understand how cognitive aspects such as harm, bias, emotion, knowledge, and memorization are represented in them, and cognitive and personality evaluation methods from psychology can be used to characterize their extent and reliability (Shanahan et al. 2023; Speed 2023; Zou et al. 2023). Several new benchmarks have been proposed to evaluate multimodal reasoning, scientific problem solving, truthful question answering, and ethical behavior (Bitton et al. 2023; Marks and Tegmark 2023; Pan et al. 2023; Wang et al. 2023). These approaches are rapidly evolving, but they are already different from prior practice of AI, suggesting that the foundation for the new paradigm is starting to emerge.

It is likely that much of this new research and development will borrow approaches from cognitive neuroscience and psychology—and the results will also benefit those disciplines. This is indeed the power of a paradigm shift. It changes the way we think about the world, and can lead to leaps of progress in several fields.

## CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.

## ORCID

Risto Miikkulainen  <https://orcid.org/0000-0002-0062-0037>

## REFERENCES

- Bitton, Y., H. Bansal, J. Hessel, R. Shao, W. Zhu, A. Awadalla, J. Gardner, R. Taori, and L. Schmidt. 2023. “VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by

- Real-World Use.” Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.
- S Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press. (2nd, enlarged edition 1970).
- Marks, S., and M. Tegmark. 2023. “The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets.” arXiv:2310.06824.
- Pan, A., J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, J. Ng, H. Zhang, S. Emmons, and D. Hendrycks. 2023. “Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark.” Proceedings of the 40th International Conference on Machine Learning.
- Patil, S. G., T. Zhang, X. Wang, and J. E. Gonzalez. 2023. “Gorilla: Large Language Model Connected with Massive APIs.” arXiv:2305.15334.
- Shanahan, M., K. McDonnell, and L. Reynolds. 2023. “Role Play with Large Language Models.” *Nature* 623: 493–98.
- Speed, A. 2023. “Assessing The Nature of Large Language Models: A Caution Against Anthropocentrism.” arxiv:2309.07683.
- Sumers, T. R., S. Yao, K. Narasimhan, and T. L. Griffiths. 2023. “Cognitive Architectures for Language Agents.” arXiv:2309.02427.
- Wang, X., Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. 2023. “SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models.” arXiv:2307.10635.
- Wei, J., X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou. 2022. “Chain of Thought Prompting Elicits Reasoning in Large Language Models.” In Proceedings of the 22nd Conference on Neural Information Processing Systems (NeurIPS 2022).
- Zou, A., L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, et al. 2023. “Representation Engineering: A Top-Down Approach to AI Transparency.” arXiv:2310.01405.

**How to cite this article:** Miikkulainen, R. 2024. “Generative AI: An AI paradigm shift in the making?” *AI Magazine* 45: 165–67. <https://doi.org/10.1002/aaai.12155>

## AUTHOR BIOGRAPHY

**Risto Miikkulainen** is a Professor of Computer Science at the University of Texas at Austin and VP of AI Research at Cognizant. He received an M.S. in Engineering from Helsinki University of Technology (now Aalto University) in 1986, and a Ph.D. in Computer Science from UCLA in 1990. His work aims to find synergies between cognitive science, computational neuroscience, and evolutionary computation, and use them to build intelligent agents. He is an AAAI and IEEE Fellow and recipient of the IEEE CIS Evolutionary Computation Pioneer Award, the Gabor Award of the International Neural Network Society, and Outstanding Paper of the Decade Award of the International Society for Artificial Life.