Tech Ethics LAB

NOTRE DAME–IBM TECHNOLOGY ETHICS LAB

Annual Report

2025
26

Table of Contents

2 Lab Mission

3 From the the Directors

4 Lab Team and Steering Committee

5 Collaborative Projects

9 Academic Research Grants

11 Raise the Bar 2025

13 2026-27 Academic Research Grants

14 New Partnerships and Initiatives

15 Pulitzer Partnership

16 Program Chairs

17 Externships

19 2026-27 Graduate Fellows

21 Tech Ethics Lab Blog

22 Lab History

MISSION

The **Notre Dame-IBM Technology Ethics Lab** produces critical research and thought leadership on ethical dimensions of artificial intelligence and other technologies by engaging with a broad set of stakeholders to examine real-world challenges and opportunities.

Our mission is to drive ethical and human-centered design, development, deployment, and use of technology.

We advance applied, interdisciplinary research projects that stimulate essential dialogue, foster collaborative communities, and influence policies and practices for responsible innovation and governance.

Tech Ethics LAB

Letter from the Directors

This past year, the Notre Dame–IBM Technology Ethics Lab continued to grow as a force for responsible innovation at one of the most consequential moments in the history of technology. Across our various research collaborations, partnership engagements, and community convenings, a single truth has become increasingly clear: designing and deploying AI in ways that actually benefit society and promote human flourishing will require all of us to slow down and ask critical questions about what we want life to look like, working together to create a future worth caring about, not one that is forced upon us.

Our first annual Raise the Bar conference, held at IBM's One Madison Avenue headquarters in New York City, brought nearly 200 senior leaders from business, technology, and academia together to wrestle with exactly this challenge. Over two days of critical reflection and hands-on collaboration, participants explored the possibilities and challenges of deploying AI responsibly at scale. The conversations were candid and generative. Whether the topic was trust, reliability, governance, or the future of the workforce, we kept returning to the idea that responsible AI begins with human responsibility—to ourselves and to others—and demands leadership, accountability, genuine care, and courage at every level of an organization. These conversations both affirmed and sharpened our mission, directly informing our research agenda for the year ahead.

That agenda is now organized around a theme we believe is urgent: the future of work and the workforce in the age of agentic AI. We are entering a period where AI systems don't just assist human workers; they now act alongside them, make decisions on their behalf, and reshape the very nature of professional labor. This shift demands serious, sustained inquiry. What does meaningful and dignified human oversight look like when agents operate autonomously across complex workflows? How do we ensure that automation expands human capability and creativity rather than displacing it? How do we design systems that protect and expand worker dignity, privacy, agency, and wellbeing, not as afterthoughts, but as first principles? These are some of the questions driving our upcoming research grants, partnerships, and collaborative projects in the year ahead, and we believe they are among the most important questions of our time.

Progress on questions this complicated requires collaboration that matches their scale and complexity. No single discipline, sector, or perspective has all the answers. That is why the Lab's model of bridging academic research and industry practice, connecting computer scientists with philosophers, sociologists, anthropologists, legal scholars, humanists, and economists, and convening people across belief systems and institutional backgrounds is so important. The challenges of agentic AI will not yield to siloed thinking. They require the kind of critical, pluralistic, good-faith inquiry that the Lab was built to foster and is honored to continue.

We are grateful for the researchers, fellows, faculty, partners, and administrative staff who bring that spirit to this work every day. Looking ahead, we are more energized than ever by the Lab's potential to provide the rigorous, human-centered guidance that business, government, and civil society urgently need. We aim to demonstrate, again and again, that ethical and effective AI are not in tension, but are in fact inseparable.

– Sara Berger and Megan McDermott
Notre Dame–IBM Tech Ethics Lab Co-Directors

The Lab Team


McDermott


Berger


Lee


Sullivan


Rhoads


Go


Fehring


Welser


Greytok

Lab Administration

Megan McDermott, Notre Dame, Notre Dame Director of the Notre Dame-IBM Technology Ethics Lab

Sara Berger, IBM, IBM Director of the Notre Dame–IBM Technology Ethics Lab

Jungmin Lee, IBM, Associate Director of Lab Operations

Steering Committee

Co-founded by the University of Notre Dame and IBM, the Technology Ethics Lab is governed by a steering committee which equally represents Notre Dame and IBM.

Meghan Sullivan, Notre Dame, Director, Ethics Initiative and Institute for Ethics and the Common Good; Wilsey Family College Professor of Philosophy

Jeffrey F. Rhoads, Notre Dame, John and Catherine Martin Family Vice President for Research, Professor of Aerospace and Mechanical Engineering

David Go, **Notre Dame,** Vice President and Associate Provost for Academic Strategy, Viola D. Hank Professor of Aerospace and Mechanical Engineering

Nick Fehring, IBM, Controller

Jeffrey J. Welser, IBM, Chief Operating Officer, IBM Research; Vice President, Exploratory Science and University Collaborations

Betsy Greytok, IBM, Vice President, Responsible Technology Officer and Co-Chair, IBM AI Ethics Board

Collaborative Projects

In 2024, the Lab launched eleven projects that received almost $300,000 in funding. The projects paired 10 Notre Dame faculty members with 15 IBM researchers who worked together to study the ethical challenges emerging at the research frontier of large language models. These projects were completed December 2025, with results, papers, and artifacts continuing to be delivered through mid 2026. An overview of each project and summary of their outcomes to date can be found below, with additional details and links found on our website.

Evaluation, Metrics, and Benchmarks for Generative AI Systems

Notre Dame Collaborators: Diego Gómez-Zará (Computer Science and Engineering), Toby Jia-Jun Li (Computer Science and Engineering)

IBM Researchers: Michelle Brachman, Zahra Ashktorab, and Werner Geyer

Developed scalable, context-aware tools for evaluating genAI in real-world settings, with stakeholders and users co-creating assessment criteria.

Toby Li, Notre Dame collaborator on the "Evaluation, Metrics, and Benchmarks" project, presents his research at a conference on Notre Dame's campus.

Governance, Auditing, and Risk Assessment for Large Language Models

Notre Dame Researcher: Nuno Moniz (Lucy Family Institute)

IBM Researchers: Michael Hind and Elizabeth Daly

Introduced "Benchmark Cards", a standardized framework for documenting LLM benchmark properties, alongside a public database to support transparent and reproducible AI evaluations.

Fairness and Equity for Large Language Models

Notre Dame Researcher: Nitesh Chawla (Lucy Family Institute)

IBM Researcher: Elizabeth Daly

Addressed bias in LLMs through Chain of Thought prompting, fine-tuning, and anomaly detection methods to improve fairness, robustness, and reliability of models.

Interpretable and Explainable Foundation Models

Notre Dame Researchers: Fanny Ye (Computer Science and Engineering) and Nuno Moniz (Lucy Family Institute)

IBM Researcher: Keerthiram Murugesan

Tackled LLM hallucinations in high-stakes domains by grounding responses in knowledge graphs and using input attribution to improve interpretability and reliability.

Fanny Ye, who collaborated on the "Interpretable and Explainable Foundation Models" project, teaches graduate students in Notre Dame's Department of Computer Science and Engineering.

Next-Generation AI Models and Responsible AI

Notre Dame Researchers: Nuno Moniz (Lucy Family Institute) and Nitesh Chawla (Lucy Family Institute)
IBM Researchers: Youssef Mrouehe and Payel Das

Explored hybrid and memory-based architectures to overcome the computational, energy-consuming, and reasoning limitations of current transformer-based models, with a focus on efficiency and long-term sustainability.

Robustness for Large Language Models

Notre Dame Researcher: Xianliang (Lynn) Zhang (Computer Science and Engineering)
IBM Researchers: Pin-Yu Chen and Tien Gao

Created a "scientific lab simulator" to test whether LLM-generated action plans were safe and feasible, with the goal of extending this trustworthiness evaluation to other safety-critical domains.

Expanding Large-Scale Model Evaluations

Notre Dame Researcher: Luis Felipe Rosado Murillo (Anthropology)
IBM Researcher: Sara Berger

Leveraged mixed methods (qualitative/ethnographic and quantitative/benchmarking) to study how human assumptions shape LLM evaluation practices and specific LLM interactions, like vibe coding.

Data-Driven Work Practices in LLMs to Design Ethical and Human-Centric Mitigation Tools

Notre Dame Researcher: Karla Badillo-Urquiola (Computer Science and Engineering)

IBM Researcher: Adriana Alvarado Garcia

Partnered with youth online safety experts to build domain-specific evaluative datasets and risk taxonomies that better detect harmful LLM outputs for vulnerable populations.

Agentic AI: Trustworthy Agent-Based Systems

Notre Dame Researcher: Nuno Moniz (Lucy Family Institute)

IBM Researcher: Werner Geyer

Explored how to design human-AI delegation frameworks for single- and multi-agent systems that maintain trust, context, and appropriate human oversight.interactions, like vibe coding.

Spectral Tracing: Assessment Methods for Tracing the Impact of Consensus-Based Data Practices in LLM Development

Notre Dame Researcher: Ranjodh Singh Dhaliwal (English)

IBM Researcher: Felicia Jing

Examined how consensus-driven data curation practices can silence dissenting perspectives, developing frameworks and adversarial datasets to make AI development more inclusive and representative.

The Holistic Return on Investment (ROI) of AI Ethics

Notre Dame Researcher: Nicholas Berente (IT, Analytics, and Operations)

IBM Researchers: Francesca Rossi, Heather Domin, and Brian Goehring

Additional Collaborator: Marianna Ganapini

Built both a theoretical framework and practical tool for measuring the business value of ethical AI investments, drawing on interviews, surveys, and existing literature.

Nick Berente, ND PI on the Holistic ROI of AI Ethics project

Collaborative Projects by the Numbers

21 *papers published in renown journals or leading disciplinary conferences*

Including but not limited to: NeurIPS, Nature Machine Intelligence, CHI, ICLR, IUS, KDD, CSCW, ACL, AIES, and 4S.

4 *presentations given at workshops*

3 *invited talks or keynotes at top venues*

3 *public- and client-facing white papers produced and disseminated*

5 *open sourced tools, benchmarks, or datasets (with associated repos) produced and released via GitHub and/or HuggingFace.*

Current project outputs can be viewed at

ethics.nd.edu/lab-outputs

Academic Research Grants

In January 2025, the Lab awarded six research teams grants totalling $345,000 to advance a variety of technology ethics projects. Pairing Notre Dame faculty with scholars from partnering research institutions, these grants catalyzed innovation, insight, and new tools to help AI become safer, more trustworthy, and more reliable.

(Digital) Companionship in the Digital Age: On Human-AI Relationships and the Ethical Landscape Surrounding Artificial Others

Principal investigators: Robert Clowes (NOVA University of Lisbon) and Kesavan Thanagopal (Notre Dame)

Notre Dame Collaborator: Diego Gómez-Zará (Computer Science)

This project examined the "mind-technology problem"—how technology reshapes thinking, knowledge, and human agency—in the age of widespread access to generative AI. Clowes and his team examined AI companion apps like Replika, CharacterAI, and Kuki, working to establish concrete classifications of the roles these systems play in people's lives and to explore the ethical implications of deeply affective human-AI relationships. As the culmination of this project, Clowes edited a special issue of the journal *Social Epistemology* on this topic.

Building AI Text Classifiers with Peacebuilders: A Human-AI Collaboration to Improve Conflict Analysis and Resolution

Principal investigator: Allan Cheboi (Build Up)

Notre Dame Collaborator: Lisa Schirch (Peace Studies)

Peacebuilding processes that involve novel technologies have great potential, but the technologists and peacebuilding experts involved rarely get opportunities to truly collaborate. This project integrated peacebuilders from across Africa into the early stages of the development of a new AI text classification system. By drawing on the expertise of both the tech developers and the peacebuilders, the team aims to drive better early conflict analysis and resolution outcomes.

Digital Afterlives? Mourning, Memory, and Grief Tech

Principal investigators: Joseph Davis (University of Virginia), Micah E. Lott (Boston College), and William Hasselberger (Catholic University of Portugal)
Notre Dame Collaborator: Paul Scherz (Theology)

The team examined "ghostbots" and other forms of AI-powered "grief tech"—new AI tools that can analyze data from a specific deceased person, like text messages, emails, and videos, to create an interactive digital companion that simulates them. In addition to publishing their own research, the

team organized a conference on the topic at Universidade Católica Portuguesa in Lisbon, the proceedings from which will appear in a forthcoming special issue of the journal *Social Research*.

Jeffrey Bishop, Principal Investigator of the "Digital Moral Twins" project, talks with Lab Program Chair Tim Weninger at *Raise The Bar 2026*.

Digital Moral Twins: From Bioethical Principles to AI Ethics and Back Again

Principal investigator: Jeffrey P. Bishop (Saint Louis University)

Notre Dame Collaborator: Paul Scherz (Theology)

The goal of this project was to determine to what extent the principles and practices of the well established field of bioethics can be applied to new and emerging questions of AI ethics. The team published findings in *The American Journal of Bioethics*, *The Review of Faith & International Affairs*, and *The Wall Street Journal*.

Enhancing Human-AI Collaboration and Policy in Emergency Response: Ethical Deployment of AI-Enabled Drones

Principal investigators: Ricardo Morales (Brown University) and Demetrius Hernandez (Notre Dame)

Notre Dame Collaborator: Jane Cleland-Huang (Computer Science)

How do we deploy autonomous systems that strengthen rather than strain public trust and safety? This project took up this question in a specific high-stakes context: the FAA's newly proposed rules for drones in BVLOS (beyond visual line of sight) operations. The team published research and white papers with the MIT Science Policy Review to help shape how AI systems are governed in consequential, real-world contexts—from emergency response to infrastructure and climate monitoring.

Image Descriptions Are Less Reliable Than They Appear: Support for Blind Users Assessing Capabilities of AI-Powered Access Technology

Principal investigators: Amy Pavel (University of California, Berkeley)

Notre Dame Collaborator: Toby Li (Computer Science)

This research team developed a novel system to help blind and low vision users identify errors and uncertainties in AI-generated image descriptions. By highlighting differences across multiple versions of the same description, the system helped users identify 5x more errors than when using a single image description alone. The team shared their research at ASSETS 2025, the top accessibility conference.

Raise the Bar 2025

On October 29 and 30, 2025, the Notre Dame–IBM Tech Ethics Lab convened nearly 200 senior leaders from industry, technology, and academia at IBM's flagship One Madison Avenue office in New York City for *Raise The Bar*, the Lab's signature annual conference. Designed around a "think and do" model, the two-day convening moved deliberately from reflection to practice—using Day 1 to examine the ethical, organizational, and cultural challenges of deploying AI at scale, and Day 2 to translate those insights into action. Across panels, case studies, and candid discussions, a unifying insight emerged: responsible AI is not a technical endpoint, but an ongoing human and organizational commitment.

Ashley Reichheld, Principal at Deloitte Digital, speaks during morning panel session.

Participants surfaced four interlocking themes that now define the Lab's research and engagement priorities. First, **trust was recognized as a leadership choice**, not a system feature. While transparency and technical rigor are necessary, trust ultimately depends on visible accountability, psychological safety, and leaders who model integrity and humility—especially in moments of uncertainty or failure. Second, **reliability must be designed**, not assumed. Rather than pursuing perfection, leading organizations are building AI systems that anticipate failure, embed human judgment, and apply oversight proportionate to risk. Reliability, participants agreed, comes not from control alone, but from confidence in how systems—and people—respond when things go wrong. Third, conversations emphasized that **ethics must earn their keep**. Responsible AI endures when ethical deliberation is clearly connected to business value, resilience, and long-term performance. Organizations that treat governance as a living practice—supported by incentives, ownership, and ethical fluency—are better positioned to innovate responsibly and competitively. Finally, **humans remain at the center of meaningful change**. As AI reshapes work, the most resilient enterprises are designing systems that enhance human creativity, judgment, and adaptability, rather than optimizing people to fit machines.

Notre Dame professor Yong Suk Lee leads the afternoon panel "Humans at the Center: Ethics in Workforce AI Integration"

Together, *Raise The Bar 2025* reinforced a simple but urgent conclusion: responsible AI begins with human responsibility. Ethics, reliability, and innovation are not trade-offs but reinforcing practices when leadership, design, and culture are aligned. Building on this momentum, the Lab is advancing applied research, practical frameworks, and partnerships aimed at helping organizations move from principles to practice and from intent to implementation and raising the bar not only for what AI can do, but for what it should do.

2026 Theme

The Future of Work and the Workforce in the Age of Agentic AI

As AI systems become more autonomous and influential inside organizations, long-standing assumptions about work, leadership, and accountability are being reshaped. These shifts raise urgent questions about how responsibility is distributed, where human judgment remains essential, and what kinds of leadership this moment demands.

Raise the Bar 2026

When: October 28-29, 2026

Where: IBM Flagship Office
1 Madison Avenue,
New York, NY

Raise The Bar 2026 will convene senior leaders, researchers, technologists, and policy advocates for a day and a half focused on how agentic AI is transforming work and organizational life at a structural level. The convening will combine ethical and philosophical grounding with applied dialogue—surfacing unresolved questions about autonomy, power, and accountability, and examining real-world cases from organizations navigating these changes in practice. Built for those already thinking seriously about responsible AI, *Raise The Bar 2026* continues the Lab's commitment to moving beyond principles toward lived organizational practice—where human judgment, leadership, and care remain central as work is reimagined in the age of agentic systems.

2026-27 Academic Research Grants

The Lab's 2026 Call for Proposals focuses on human dignity and agency in the workforce in the age of agentic AI. As AI systems move beyond generating content to autonomously performing tasks, making decisions, and coordinating workflows, workers across industries are encountering new and complex challenges to their autonomy, professional identity, and sense of purpose. The 2026 CFP invited researchers to examine these challenges through a novel two-phase structure: Phase 1 provides up to $20,000 over four months for research design, with successful teams eligible to apply for Phase 2 implementation grants of up to $100,000 over twelve months. The Lab received 10 proposals from 8 institutions and awarded Phase 1 funding to five projects. Phase 2 proposals will be due in August 2026.

Automating Ourselves: Agency Loops, Worker Dignity, and the Service-Software Cycle

Principal investigator: Paul Leonardi (University of California, Santa Barbara)

This project examines the "service-software cycle" — the organizational strategy in which firms repeatedly convert human-delivered services into AI-driven products, enlisting workers to encode their own expertise into systems that may ultimately redefine their roles. Drawing on pilot fieldwork at a digital marketing firm, the Phase 1 research design develops theoretical constructs and interview protocols for a comparative field study of how workers experience and respond to recursive self-automation.

From Voluntary Intensification to Sustainable AI Practice

Principal investigator: Aruna Ranganathan (University of California, Berkeley)

Building on eight months of ethnographic fieldwork, this project investigates a subtle but widespread consequence of AI adoption: rather than displacing workers, AI tools often intensify their work through workers' own voluntary engagement. Phase 1 identifies where organizational intervention is feasible, with Phase 2 planning to test interventions. The project reframes agency as the capacity to decide when to stop, not just how to act.

Task Re-organization and the Meaning of Work in the Age of AI

Principal investigators Yong Suk Lee and Christopher Mills (University of Notre Dame)

This project measures how AI reorganizes work at the task level — substituting, augmenting, or migrating tasks — and what those shifts mean for meaningfulness, autonomy, and skill development. Phase 1 resolves a key methodological question between experimental and survey-based approaches, and runs a pilot study. The project draws on an interdisciplinary Notre Dame team spanning economics, computer science, education, and policy.

Gatekeepers of Work: Protecting Human Dignity and Agency in the Age of Agentic AI

Principal investigators: Heng Xu, Cam Kormylo, and Nan Zhang (University of Notre Dame)

This project argues that agentic AI systems must be constrained to option-generating functions rather than choice-delegating ones. It identifies two core threats: dignity deficits arising when fairness-compliant algorithms paradoxically disadvantage qualified individuals, and agency suppression through "anticipatory conformity," in which workers reshape their behavior to match what they believe the AI rewards. Phase 1 includes a distinctive convening with the National Academies.

Human Flourishing in the Creative Workplace After AI Adoption

Principal investigators: Greg Robson (University of Notre Dame), Andrew Bailey (National University of Singapore), and Justin Tosi (Georgetown University)

This project brings a philosophical lens to the question of how AI adoption affects human flourishing in creative industries such as writing, design, and the arts. Distinguishing between crises of agency for producers and crises of dignity for consumers, the project explores what is lost — and what might be preserved — as AI becomes embedded in creative work. Phase 1 develops the foundational philosophical framework to guide empirical investigation in Phase 2.

New Partnerships and Initiatives

The Lab is excited to announce a new partnership with All Tech Is Human around a shared focus on the "Agentic Shift", the profound transformation in how people work as AI moves from passive tool to active agent. In summer 2026, All Tech Is Human and the Tech Ethics Lab will co-host a two-day workshop series in New York City, bringing together mid-level professionals and senior executives to surface the ground-level realities, governance challenges, and ethical tensions of human-AI agent collaboration in the modern workplace. Findings from the workshops will be synthesized into a white paper at the end of the year. We look forward to sharing more soon.

Pulitzer Partnership

The Notre Dame–IBM Technology Ethics Lab partnered with the Pulitzer Center—alongside the MacArthur Foundation and the Ford Foundation—to support the AI Spotlight Series, a global journalism training initiative focused on strengthening accountability and ethical coverage of artificial intelligence. As AI and predictive technologies increasingly shape decisions in government, healthcare, education, and the workplace, this collaboration addressed a central ethical challenge: ensuring that these systems are critically examined, publicly understood, and responsibly governed.

Originally designed to train 1,000 journalists, the program ultimately exceeded its initial scope. By the conclusion of the partnership in 2025, the AI Spotlight Series had delivered more than two dozen online and in-person trainings, reaching nearly 3,000 journalists across seven languages worldwide, with a strong emphasis on reporters from the Global South and from communities underrepresented in media. A key milestone of the partnership was the open-sourcing of the AI Spotlight Series curriculum—including course videos, slide decks, worksheets, and resource guides—which expanded global access and fostered a collaborative knowledge community. The curriculum was structured across three tracks—for reporters on any desk, journalists specializing in AI, and editors commissioning and shaping coverage—supporting ethical, accountability-driven reporting at both the individual and newsroom levels.

Together, the Lab and the Pulitzer Center reinforced journalism's role as a critical accountability mechanism for emerging technologies and contributed to more informed public discourse and responsible innovation.

Lab Director Sara Berger meets with Pulitzer Center President and CEO Lisa Gibbs on campus at the University of Notre Dame.

Program Chairs

Program Chairs are Notre Dame faculty members who develop research programs at the intersection of technology and ethics that align with or directly complement the vision and mission of the Lab. In the 2025-26 academic year, each of the Lab's three inaugural Program Chairs completed the first of their three-year terms. Through their research, teaching, and engagement with the Lab's Graduate Fellows, the Program Chairs contributed to the intellectual and communal life of the Lab, and advanced significant projects in the field of technology ethics.

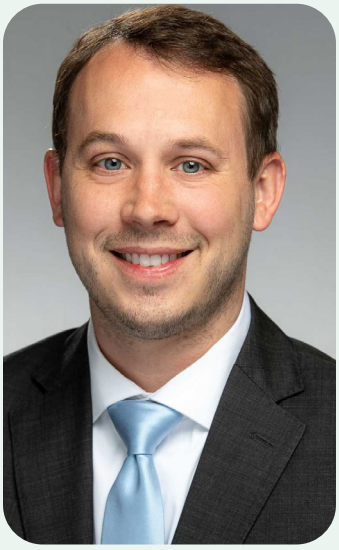
Meng Jiang, Frank M. Freimann Collegiate Professor of Engineering

Dr. Meng Jiang's research fields are data mining, machine learning, and natural language processing. In his first year as Program Chair, Jiang launched the Tech Ethics Data (TED) Program, a research initiative that pairs undergraduate and graduate students with faculty collaborators to build foundational datasets and analytical insights for the technology ethics community. The program's inaugural cohort produced work spanning NLP ethics, moral reasoning in language models, and contextual safety in multimodal AI systems, with papers accepted to the ACL conference and ACM CSCW. Jiang also organized a Physical AI Research Poster Session featuring 27 research posters on safety in AI and robotic systems, and developed new interdisciplinary collaborations connecting engineering with law, social science, and theology.

Paul Scherz, Our Lady of Guadalupe Professor of Theology

Dr. Paul Scherz's work examines the intersection of theology, science, medicine, and technology. Scherz had a prolific first year, anchored by the publication of Reclaiming Human Agency in the Age of AI, a book he co-led for the AI Working Group of the Vatican's Dicastery for Culture and Education. The book is being translated into Korean and has generated a series of commentaries in the Journal of Moral Theology. Scherz also made substantial progress on a new book project, Virtue and Tools: A Theological Ethics of AI, drafting close to ten chapters. His public engagement was extensive: he delivered an address to the U.S. Conference of Catholic Bishops plenary, presented at the Pontifical Academy for Social Sciences, and gave talks at institutions including the Angelicum, NYU Medical School, and the University of Scranton. He also co-organized a conference on digital afterlives and AI-generated memorial technologies, which will result in a special issue of Social Research.

Program Chairs



Tim Weninger, Associate Professor, Computer Science and Engineering

Dr. Tim Weninger's research in social media and artificial intelligence looks to better understand how humans create and consume information, especially in online social systems. In his first year as Program Chair, Weninger launched multiple human-centered AI research projects with IBM, with papers carrying the Notre Dame–IBM affiliation published at top venues, and developed new research partnerships with Deloitte and Lockheed Martin. He also began work on an interactive directory of Notre Dame faculty engaged in technology ethics research, a resource designed to facilitate cross-campus collaboration. Weninger contributed to the Vatican's dialogue on AI ethics, participating in meetings that informed a papal encyclical on humans and AI, and his media engagement included interviews with the New York Times, Washington Post, CNN, and BBC.

Externships

The annual Tech Ethics Lab's externship program places Notre Dame graduate students within IBM business units for 12-weeks during the summer, offering a critical bridge between academic inquiry and real-world contexts where a mix of technical, business, and governance decisions are made and operationalized. When possible, the Lab matches our Tech Ethics Graduate Fellows to industry research and development teams working on projects that best fit the students' skills, background, interests, and/or doctoral work.

We kicked off the externship program in 2025 with our very first extern, Perla Khattar, a JSD candidate specializing in digital privacy. During her externship, Perla explored the ethical and governance risks of synthetic data, examining how these increasingly common data generation techniques can both mitigate and introduce new forms of bias, privacy risk, and downstream misuse. Her and her IBM colleagues' work surfaced key tensions between scalability and accountability, particularly in contexts where synthetic data is treated as a low-risk substitute for real-world data. The project highlighted the need for more robust evaluation frameworks

Tech Ethics Lab Fellow Perla Khattar on the grounds of IBM's T.J. Watson Research Center where she completed her externship during summer 2025.

as synthetic data becomes embedded in AI development pipelines and increasingly relied on for agentic AI simulations and validity testing. The research also contributed to a published whitepaper sponsored by IBM's Responsible Technology Board and generated two peer-reviewed publications (forthcoming).

In summer 2026, we are pleased to welcome three Lab Fellows to IBM's Thomas J. Watson Research Center. Students will work on projects that address the rapidly evolving landscape of AI systems and the future of work, where large language models and multi-agent systems are reshaping how decisions are made, delegated, and evaluated across domains. Adnan Hoq will be developing context-dependent red-teaming and evaluation methods for large language models, focusing on how safety and bias assessments can better account for cultural, linguistic, and domain-specific nuance. Annalisa Syzmanski will be working on scaling contextual alignment in LLMs by integrating human subject matter expertise into evaluation systems, aiming to bridge the gap between automated metrics and real-world performance. Last but not least, Mariana Fernandez-Espinosa will be investigating how concepts of contextual integrity and appropriateness can govern communication in multi-agent systems, focusing on designing coordination mechanisms that are both effective and aligned with human expectations and regulatory constraints.

2026-27 Graduate Fellows

The Notre Dame-IBM Tech Ethics Lab Graduate Fellowship is a first-of-its-kind experiential learning program for Notre Dame doctoral students interested in applied technology ethics research. The fellowship is a year-long academic experience in which Fellows form intellectual community, advance their scholarly and professional development, and contribute to the Lab's vibrant interdisciplinary network. Fellows may also have the opportunity to participate in a fully funded externship embedded in a company, where they can learn more about the contexts in which their theoretical work is lived out. Throughout the year, Fellows participate in seminars, receive guidance on research, and will have opportunities to help communicate the Lab's scholarly research to technology developers and industry professionals. This year, the fellowship will also feature a new five-day ethics bootcamp, designed to ground incoming fellows in the foundational frameworks and applied methods that will anchor their work throughout the year.

Fabián Díaz Maldonado

A Ph.D. student in the department of sociology, Díaz Maldonado studies how software developers navigate ethical responsibility in the age of AI-assisted coding. Drawing on interviews, computational analysis, and ethnography, his research examines how professional identity and organizational norms shape the moral reasoning of developers working alongside AI tools. His work sits at the intersection of the sociology of professions, science and technology studies, and the ethics of AI-augmented work.

Emelia Hughes

A Ph.D. student in computer science and engineering with a focus in communication, Hughes investigates how algorithmic recommendation systems shape online community formation and belonging. Her research explores how users develop a sense of community membership through repeated exposure to platform-curated content — often before any explicit social ties form — raising important questions about the role of algorithms in mediating human connection.

Gina Mannino

A Ph.D. student in the department of statistics, Mannino develops improved methods for differential privacy and synthetic data generation. Her work includes a novel "heterogeneous DP" framework that tailors privacy protections to individuals and data types rather than applying a one-size-fits-all approach, advancing more equitable and context-sensitive approaches to data privacy.

Maria Milkowski

A Ph.D. student in the department of computer science and engineering, Milkowski studies conflict and cooperation among autonomous AI agents. Using simulations such as Among Us and Sugarscape, her research explores how competing goals and ethical frameworks shape agent behavior and system stability — work with direct implications for the design of trustworthy multi-agent AI systems.

Wilson Raney

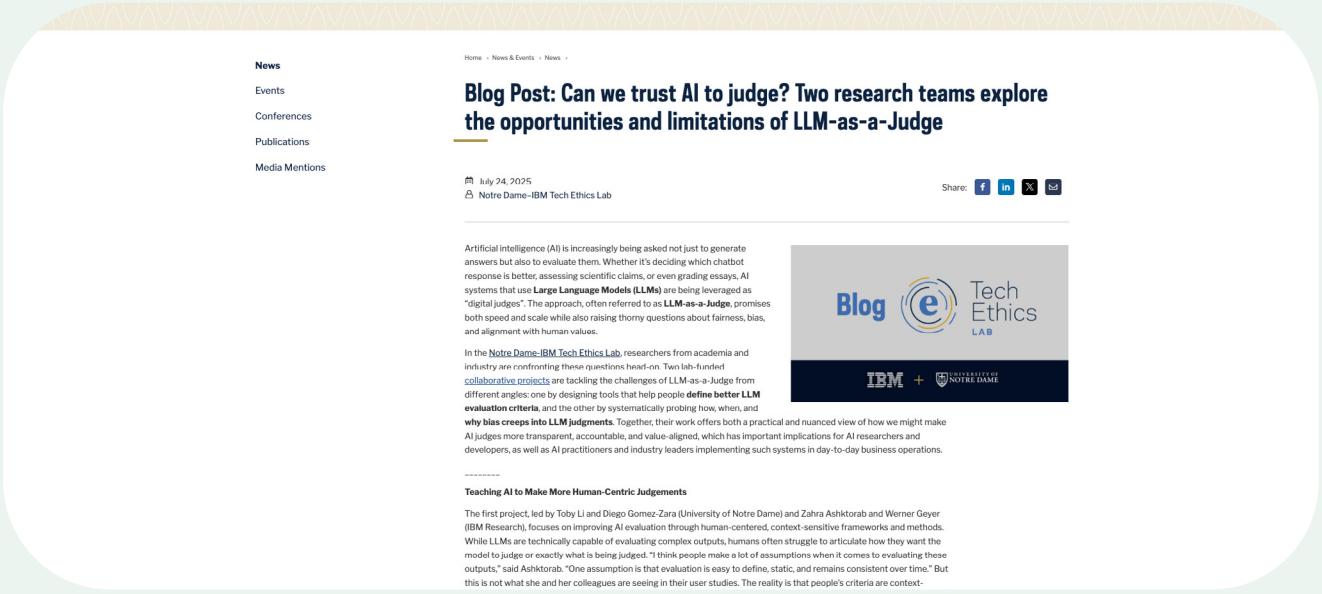
A Ph.D. student in chemical engineering, Raney uses computational methods and machine learning to screen metal-organic frameworks (MOFs) for their potential to generate entangled photons. His research aims to discover new materials for quantum communication technologies, bringing an applied science perspective to questions about how emerging technologies are developed and governed.

Julia Shenot

A Ph.D. student in the department of philosophy, Shenot works on the metaphysics of consciousness and its ethical implications for AI. She argues that consciousness is fundamentally a social rather than purely scientific attribution — a position that reframes how we should think about the moral status of AI systems and the ethical responsibilities of those who design them.

Tech Ethics Lab Blog

The Tech Ethics Lab Blog was launched in July 2025. The blog provides exclusive interviews with Lab faculty, highlights from new publications, tools, and other research project outputs, and a dedicated series on the future of work.



News
Events
Conferences
Publications
Media Mentions

Home › News & Events › News ›

Blog Post: Can we trust AI to judge? Two research teams explore the opportunities and limitations of LLM-as-a-Judge

July 24, 2025
Notre Dame–IBM Tech Ethics Lab

Share:

Artificial intelligence (AI) is increasingly being asked not just to generate answers but also to evaluate them. Whether it's deciding which chatbot response is better, assessing scientific claims, or even grading essays, AI systems that use **Large Language Models (LLMs)** are being leveraged as "digital judges". The approach, often referred to as **LLM-as-a-Judge**, promises both speed and scale while also raising thorny questions about fairness, bias, and alignment with human values.

Blog Tech Ethics LAB
IBM + University of Notre Dame

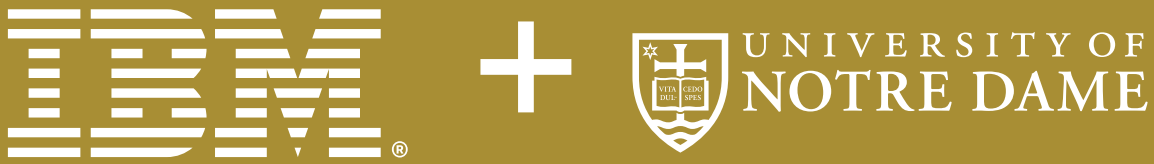
In the Notre Dame–IBM Tech Ethics Lab, researchers from academia and industry are confronting these questions head-on. Two lab-funded collaborative projects are tackling the challenges of LLM-as-a-Judge from different angles: one by designing tools that help people **define better LLM evaluation criteria**, and the other by systematically probing how, when, and **why bias creeps into LLM judgments**. Together, their work offers both a practical and nuanced view of how we might make AI judges more transparent, accountable, and value-aligned, which has important implications for AI researchers and developers, as well as AI practitioners and industry leaders implementing such systems in day-to-day business operations.

Teaching AI to Make More Human-Centric Judgements

The first project, led by Toby Li and Diego Gomez-Zara (University of Notre Dame) and Zahra Ashktorab and Werner Geyer (IBM Research), focuses on improving AI evaluation through human-centered, context-sensitive frameworks and methods. While LLMs are technically capable of evaluating complex outputs, humans often struggle to articulate how they want the model to judge or exactly what is being judged. "I think people make a lot of assumptions when it comes to evaluating these outputs," said Ashktorab. "One assumption is that evaluation is easy to define, static, and remains consistent over time." But this is not what she and her colleagues are seeing in their user studies. The reality is that people's criteria are context-

The Lab inaugurated the blog series by featuring the collaborative work of Toby Li, Diego Gomez-Zara, and Xiangliang Zhang from the University of Notre Dame, alongside Zahra Ashktorab, Werner Geyer, Pin-Yu Chen, and Tian Gao from IBM Research. As cross-sector fellows and principal investigators at the Notre Dame-IBM Tech Ethics Lab, this interdisciplinary team utilized the platform to examine the ethical frontiers of the "LLM-as-a-Judge" paradigm. Their featured work—spanning the development of human-centric evaluation interfaces like MetricMate and adversarial bias-detection frameworks like CALM—focused on how developers and business leaders can anticipate the vulnerabilities of automated decision-making while safeguarding algorithmic fairness and human oversight.

Follow the Tech Ethics Lab on LinkedIn for the latest updates.



The plans for the **Notre Dame-IBM Technology Ethics Lab** were formed in 2020 when it became clear that emerging technologies were beginning to face critical and complex ethical challenges. These technologies are making game-changing innovations at speeds never imagined, but they risk quickly losing the trust of the very society they intend to improve.

In this spirit, IBM made a 10-year, $20 million commitment to Notre Dame to create the jointly run Tech Ethics Lab, which combines IBM's deep technical expertise with Notre Dame's strength in philosophy and ethics. The lab aims to develop and apply a holistic approach to tech ethics and practical, research-based models for the responsible development and deployment of these technologies across enterprises.

In a short amount of time, the Tech Ethics Lab has moved to the forefront of the global conversation on how to build ethics into the early stages of design and development and continue throughout the entire lifecycle of the technology. Furthermore, as the technologies evolve, so must the ethical lenses applied to those technologies.

Through this partnership, we are bringing together leading academic, business, community, and government leaders to ask the tough questions and to champion responsible technology that is human-centric.

UNIVERSITY OF NOTRE DAME | ETHICS AND THE COMMON GOOD

1124 Flanner Hall, Notre Dame, IN 46556 USA | (574) 631-1305 | ethics.nd.edu | ethics.nd.edu/techethicslab