Trustworthy AI kickoff

Wisconsin RISE-AI · Focus Area / Working Group Kickoff

Problem solved

An example

Dogfooding

Eating your own dog food

7 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Eating your own dog food or "**dogfooding**" is the practice of using one's own products or services.[1] This can be a way for an organization to test its products in real-world usage using product management techniques. Hence dogfooding can act as quality control, and eventually a kind of testimonial advertising. Once in the market, dogfooding can demonstrate developers' confidence in their own products.[2][3]

I used Claude code + the faculty search tool DSI has developed + UW design system to create the following slides.

- The faculty search tool, reads titles and abstracts of papers surfaced from Academic Analytics + an AI model to create semantic embeddings for those papers + a matching algorithm.
- How well does it work? Who is missing? What are the biases?
- These are questions of trustworthy AI that hit home for us as an academic community.

An example

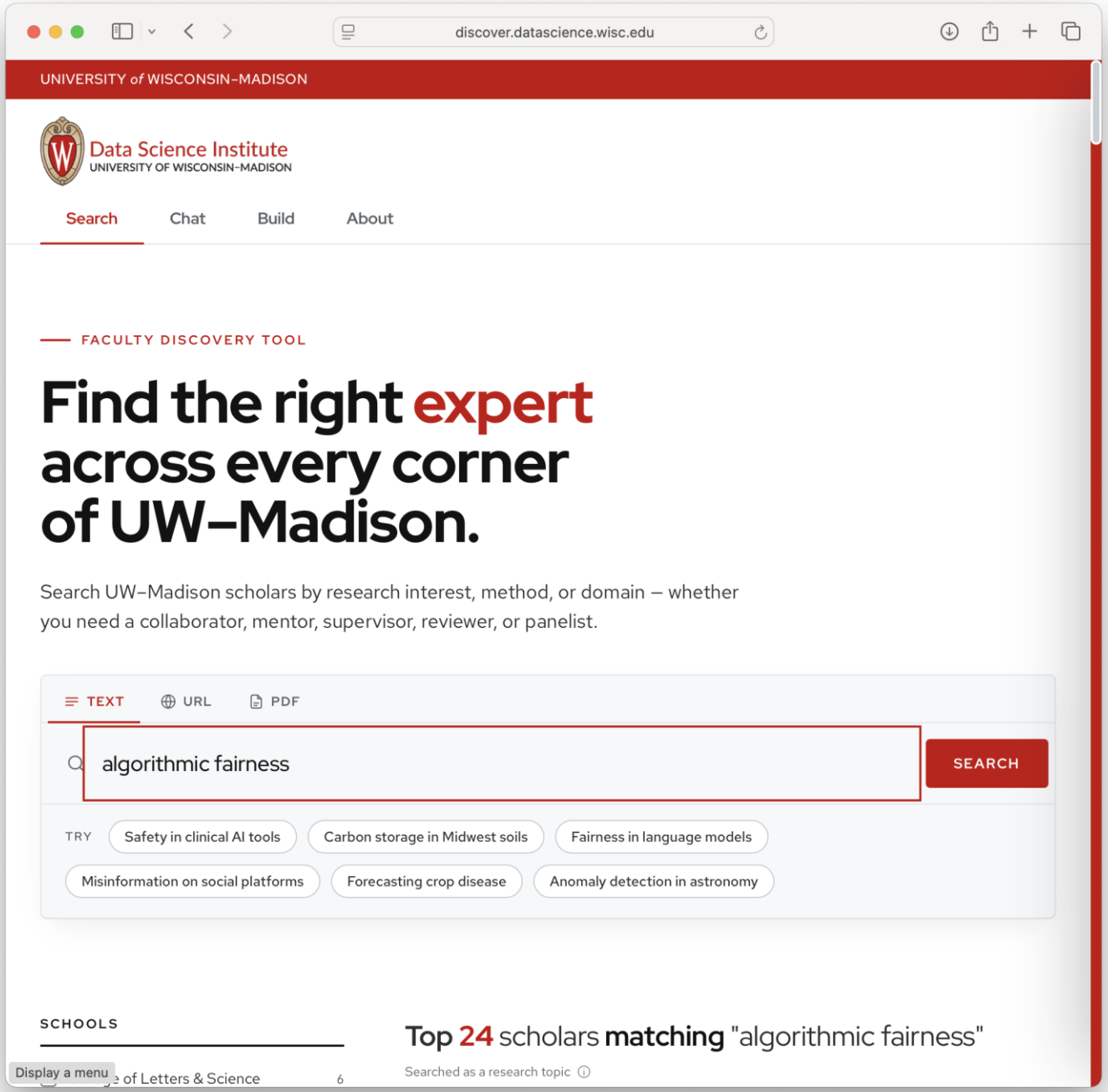
Transparency

As part of RISE-AI we are creating and nurturing multidisciplinary working groups that focus on a specfic topic. We don't want the topics to be so narrow that only a few groups are participating. We also don't want them to be so broad that those participating don't have much in common. We want to right-size the groups so that the participants can meaningfully share and learn from eachother. We don't need everyone to overlap on everything, but we want different aspects of the project to be intersectional across multiple members of these groups. One such topic is AI for proteins. This works well because we have theoretical chemists that do molecular dynamics, biochemists that know a lot about the actual proteins, computer scientists thinking about AI models, infrastructure for modern techniques that use alphafold and docking and RFdiffusion etc. We also have facilities to generate more unique data etc.

Another working group that was suggested is Trustworthy AI. I like this topic for many reasons. There are faculty in computer science and engineering that are thinking about the algorithmic and formal aspects of trustworthy AI. We also have use-cases from the medical school (e.g. use of AI / ML for radiology or summarizing patient notes) etc. that drive real world use-cases. There are also use-cases from the use of AI in legal contexts. Trustworthy AI is also closely related to ethical AI, responsible AI, human-centric AI, etc. We could/should also have ethicists and some of our science communications faculty involved. Maybe also policy, regulation, and buisness school. There are so many perspectives, that I start to worry that maybe this is too broad of a topic.

can you use your knowledge of the literature and discourse around these topics to come up with some framework for this topical area. Suggest a few terms that can be used to breakdown the broad topic into more specific aspects of the question. Propose a framework that would foster interdisciplinary collaborations that include more technical / formal work, applied ethics etc., and example use-cases that would ground the work in more applied settings on campus.

Based on this framework, use the faculty search MCP server to build a standalone html slide deck that first presents the framework, topical areas, etc. and then moves into which faculty would be good fits. Add a slide with some figures (e.g. pie charts) or tables that gives a sense of the demographics of the faculty (e.g. what departments they are in, and the rank as assistant, associate, tenured) if possible.



THE DESIGN PROBLEM

Right-sized — not narrow, and not everything.

FAILURE MODE · TOO NARROW

The silo

A topic so specific only one or two labs care. Deep expertise, but no community and nothing to exchange.

THE TARGET

Right-sized

A shared vocabulary keeps everyone in the room — while distinct facets (formal, human, ethical, institutional) are each owned by different members who need one another.

FAILURE MODE · TOO BROAD

“AI & society”

Everyone nods; no one collaborates. Breadth with no shared problem, so the group never converges on work.

Trustworthy AI risks the too-broad failure. The framework that follows is built to hold it in the middle.

THE REFRAME

Trust is a chain, not a property.

LINK 01

Guarantee

Can we prove the system behaves — under attack, under shift, under uncertainty?

LINK 02

Evidence

Does it actually work on real, messy data — validated, calibrated, reproducible?

LINK 03

Calibration

Do the people using it rely on it appropriately — not too much, not too little?

LINK 04

Accountability

Once deployed, who is answerable, and how is it contested and governed?

An AI claim is only as trustworthy as its weakest link — and no single discipline can certify the chain alone.

THE FRAMEWORK

Four tracks, tested on shared proving grounds.

Track ↓ Proving ground →	Clinical & Health	Law & Public Sector	Science & Discovery	Public Understanding
01 Formal Guarantees	■	■	■	■
02 Human Trust	■	■	■	■
03 Ethics & Accountability	■	■	■	■
04 Governance & Policy	■	■	■	■

Every proving ground draws on all four tracks. The cell where a verification theorist meets a human-factors researcher meets an ethicist meets a policy scholar — around one real deployment — is exactly the collaboration this group exists to create.

THE VOCABULARY, IN FOUR TRACKS

Who owns which link.

TRACK 01

Formal Foundations & Guarantees

“Can we prove it behaves?”

Robustness, formal verification, safety, uncertainty quantification, privacy and security of ML.

Computer Sciences · ECE · ISyE · Math & Stats

TRACK 02

Interpretation, Calibration & Human Trust

“Will people rely on it appropriately?”

Explainability, calibration, human-AI interaction, appropriate reliance and over-trust, decision support.

CS / HCI · Human Factors (ISyE) · Psychology

TRACK 03

Fairness, Ethics & Accountability

“Should it be used — and on whom?”

Algorithmic fairness, bias auditing, applied ethics, value alignment, contestability, participatory design.

Philosophy · Information School · Sociology

TRACK 04

Governance, Policy, Law & Adoption

“Who is answerable once it ships?”

Regulation and standards (NIST, EU AI Act), liability, data governance, organizational adoption.

Law · Public Affairs · Business · iSchool

WHAT KEEPS IT RIGHT-SIZED

Proving grounds force the intersection.

Clinical & Health AI

Radiology triage, sepsis prediction, summarizing patient notes — where a wrong, over-trusted output can harm a patient.

guarantees · calibrated reliance · bias · FDA & liability

Law, Justice & Public Sector

Risk assessment, benefits eligibility, legal drafting — where contestability and due process are the whole game.

accountability · transparency · fairness · governance

Science & Discovery

AI for science — including AI-for-Proteins — where "trust" means uncertainty, reproducibility, and knowing when to believe a prediction.

uncertainty · validation · reproducibility

Public Understanding

How the public calibrates trust in AI, resists misinformation, and holds institutions to account.

risk communication · transparency · policy

Concrete, high-stakes deployments pull one or two people from every track into a shared problem. Start with two or three.

The faculty behind the framework

One community, four tracks, real proving grounds — anchored by senior scholars and seeded by RISE-AI's new hires (flagged in gold).

FACULTY · TRACK 01

Formal foundations & guarantees

Somesh Jha

Computer Sciences · Full Professor

Robustness — adversarial robustness, verification & security of ML.

Patrick McDaniel

Computer Sciences · Full Professor

Security of ML — systems security and adversarial ML.

Aws Albarghouthi

Computer Sciences · Associate

Formal methods — verifying fairness & robustness.

Sharon Li

Computer Sciences · Associate

Reliable ML — out-of-distribution detection.

Adithya Murali

Computer Sciences · Assistant

Verification — program verification & neuro-symbolic synthesis.

RISE-AI

Sandeep Silwal

Computer Sciences · Assistant

Provable ML — algorithms & differential privacy.

RISE-AI

Manolis Vlatakis

Computer Sciences · Assistant

ML theory — optimization & learning in games.

RISE-AI

Gavin Brown

Computer Sciences · Assistant

Privacy — differential privacy & robustness of learning.

RISE-AI

Grigorios Chrysos

Electrical & Computer Eng · Assistant

Trustworthy DL — robustness, privacy & safety.

Xiangru Xu

Mechanical Engineering · Assistant

Safe control — verified learning-based control.

Ten faculty — anchored by Jha & McDaniel, seeded by four RISE-AI hires.

Human trust, ethics & accountability

John D. Lee

Industrial & Systems Eng · Full

Trust & reliance — foundational scholar of automation trust.

Douglas Wiegmann

Industrial & Systems Eng · Full

Human factors — human-automation interaction & safety.

Bilge Mutlu

Computer Sciences · Full

Human-AI — human-AI & human-robot interaction.

Tony McDonald

Industrial & Systems Eng · Associate

Reliance — reliance on automated decision support.

Benjamin Lengerich

Statistics · Assistant

Interpretable ML — context-adaptive ML for medicine.

RISE-AI

Annette Zimmermann

Philosophy · Assistant

Algorithmic justice — political philosophy of AI.

Alan Rubel

Information School · Full

Information ethics — privacy & accountability.

Clinton Castro

Information School · Assistant

Ethics of AI — algorithmic fairness.

Devansh Saxena

Information School · Assistant

Accountability — public-sector algorithms.

Track 02 (human trust) and Track 03 (fairness, ethics & accountability).

FACULTY · TRACK 04

Governance, policy & law

Benjamin Sobel

Law School · Assistant

AI & copyright — the law of generative AI.

RISE-AI

Paul Connell

Law School · Assistant

Law & economics — AI for empirical social science.

RISE-AI

Jason Reinecke

Law School · Assistant

IP & innovation law — patents, tech & the courts.

Vallabh Sambamurthy

Business · Info Systems · Full

IT governance — technology adoption.

Jordan Tong

Business · Ops & Tech Mgmt · Full

Human-AI decisions — decision-making in organizations.

Michael Light

Sociology · Full

Justice systems — courts & algorithmic decisions.

Note: Law was missed originally because law faculty are not indexed by academic analytics.
A trustworthy AI issue

FACULTY · PROVING GROUND

Clinical & Health AI

Majid Afshar

Medicine · Pulm & Critical Care · Assoc

Clinical NLP — ML from patient notes & EHRs.

Matthew Churpek

Medicine · Pulm & Critical Care · Full

Deterioration — ML for sepsis & clinical risk.

Brian Patterson

Emergency Medicine · Associate

Decision support — AI-based CDS in the ED.

Pallavi Tiwari

Radiology · Associate

Radiomics — ML for cancer imaging.

Elizabeth Burnside

Radiology · Full

Screening AI — cancer detection & screening.

Oguzhan Alagoz

Industrial & Systems Eng · Full

Decision modeling — medical screening policy.

Roomasa Channa

Ophthalmology · Associate

Autonomous AI — diagnostic AI & equity in eye care.

Adam Rule

Information School · Assistant

EHR usability — clinical data visualization.

Where guarantees, calibrated reliance, fairness and liability meet one patient.

FACULTY · PROVING GROUND

Public understanding & communication

Dominique Brossard

Life Sciences Communication · Full

Public perception — of science & technology.

Michael Xenos

Life Sciences Communication · Full

Public opinion — & digital media.

Sedona Chinn

Community & Env. Sociology · Assistant

Misinformation — perceptions of AI.

Young Mie Kim

Journalism & Mass Comm · Full

Political comm — misinformation & influence.

Ross Dahlke

Journalism & Mass Comm · Assistant

AI-generated content — computational social science.

RISE-AI

Tomás Dodds

Journalism & Mass Comm · Assistant

Public-good AI — AI in journalism & its ethics.

RISE-AI

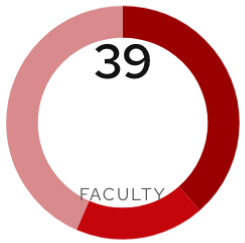
How the public calibrates trust — and holds institutions to account.

Faculty identified thus far

39 faculty · 6 schools · 18 departments.

By rank

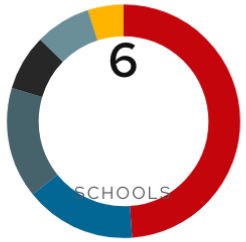
N = 30



Full (tenured)	15 · 38%
Associate	7 · 18%
Assistant	17 · 44%

By school / college

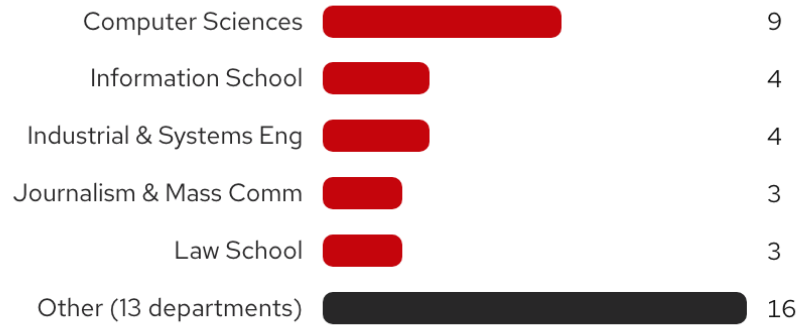
N = 30



Letters & Science	19 · 49%
Medicine & Public Health	6 · 15%
Engineering	6 · 15%
Law School	3 · 8%
Ag & Life Sciences	3 · 8%
Business	2 · 5%

By department

TOP UNITS · REST GROUPED



Read: the roster balances **15 tenured full professors** – anchors of credibility – with **17 pre-tenure faculty**; the RISE-AI hires (badged) skew deliberately junior, bringing early-career energy and time to invest. It now spans **6 schools including the Law School**, closing the governance gap this proposal originally flagged – with public policy (La Follette) the next target.

Discussion

Format & Organization

Format

- What kind of events?
- Alternate between horizontals / verticals
- Mix horizontals and verticals (intersections)
- Sub-groups only

Organization / Volunteers

- Need 1 or 2 main point people
- Sub-leads for horizontals / verticals?