INDIANA UNIVERSITY BLOOMINGTON

Trustworthiness in the Edge-Cloud-HPC Continuum

Beth Plale
Burns McRobbie Chair of Computer Engineering
Chair, Dept of Intelligent Systems Engineering
Executive Director, Pervasive Technology Institute
Indiana University, Bloomington, Indiana USA

PERVASIVE TECHNOLOGY INSTITUTE

First International Workshop on Conversational AI Interfaces for HPC
July 24, 2023

ICICLE
DEMOCRATIZING AI

With thanks to

- Sachith Withana, ISE PhD student
- Sadia Khan, Informatics PhD student
- Yu Luo, Postdoctoral Scholar
- Julie Wernert, PTI

PERVASIVE
TECHNOLOGY INSTITUTE

ICICLE
DEMOCRATIZING AI

NSF

The Common Good

Computing for the common good

PEARC23

The “common good” refers to those facilities— material, cultural or institutional—that the members of a community provide to all members in order to fulfill a relational obligation they all have to care for certain interests that they have in common.

e.g., the road system; public parks; museums and cultural institutions; public transportation; civil liberties; clean air and clean water

Stanford Encyclopedia of Philosophy

Computing for the common good

"members of a community provide to all members in order to fulfill a relational obligation they all have to care for certain interests that they have in common"

NATIONAL AI RESEARCH INSTITUTES

$500 million dollar investment in AI Institutes research network

ACCESS

National cyberinfrastructure a hundreds of millions of dollars investment

Common good asks that we as cyberinfrastructure researchers care for the shared interest - that is, the innovations, the infrastructure

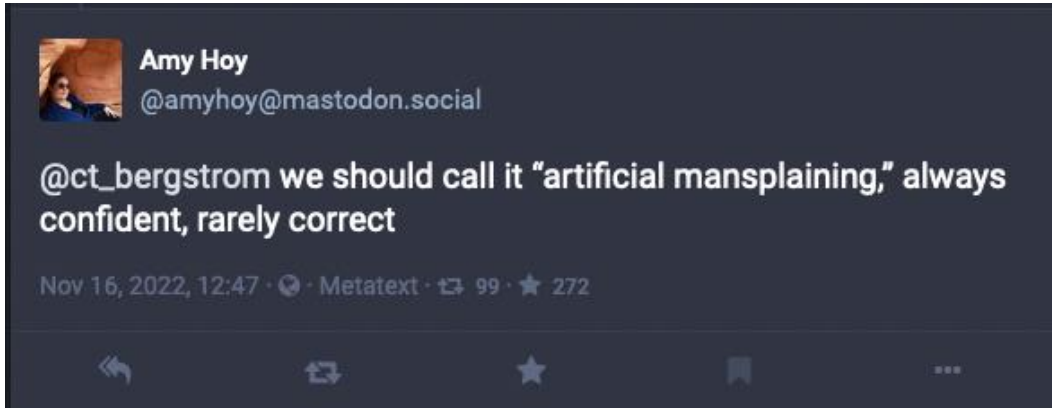
It is an obligation to 1) not add to misinformation, 2) not add AI innovations without assessment of potential harms, 3) not add to a resource depleted planet

Misinformation can, over time, erode the confidence that citizens have in the scientific methodology, and reduce confidence in scientists' commitment to acting in the public interest.

***All* use of generative AI (e.g., ChatGPT[1] and other LLMs) is banned when posting content on Stack Overflow.**

This includes "asking" the question to an AI generator then copy-pasting its output *as well as* using an AI generator to "reword" your answers.

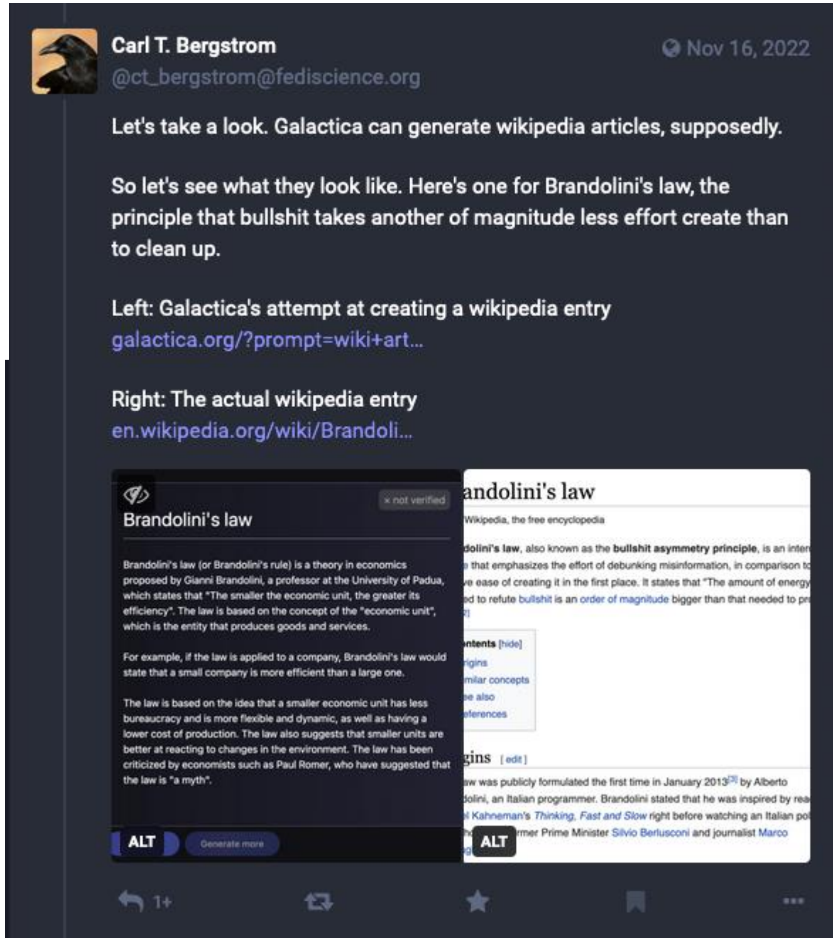
Overall, because the average rate of getting *correct* answers from ChatGPT and other generative AI technologies is too low, **the posting of answers created by ChatGPT and other generative AI technologies is *substantially harmful* to the site and to users who are asking questions and looking for *correct* answers.**



Amy Hoy
@amyhoy@mastodon.social

@ct_bergstrom we should call it “artificial mansplaining,” always confident, rarely correct

Nov 16, 2022, 12:47 · Metatext · 99 · 272



Carl T. Bergstrom
@ct_bergstrom@fediscience.org
Nov 16, 2022

Let's take a look. Galactica can generate wikipedia articles, supposedly.

So let's see what they look like. Here's one for Brandolini's law, the principle that bullshit takes another of magnitude less effort create than to clean up.

Left: Galactica's attempt at creating a wikipedia entry
galactica.org/?prompt=wiki+art...

Right: The actual wikipedia entry
en.wikipedia.org/wiki/Brandoli...

× not verified

Brandolini's law

Brandolini's law (or Brandolini's rule) is a theory in economics proposed by Gianni Brandolini, a professor at the University of Padua, which states that "The smaller the economic unit, the greater its efficiency". The law is based on the concept of the "economic unit", which is the entity that produces goods and services.

For example, if the law is applied to a company, Brandolini's law would state that a small company is more efficient than a large one.

The law is based on the idea that a smaller economic unit has less bureaucracy and is more flexible and dynamic, as well as having a lower cost of production. The law also suggests that smaller units are better at reacting to changes in the environment. The law has been criticized by economists such as Paul Romer, who have suggested that the law is "a myth".

andolini's law

Wikipedia, the free encyclopedia

dolini's law, also known as the **bullshit asymmetry principle**, is an inter
that emphasizes the effort of debunking misinformation, in comparison to
e ease of creating it in the first place. It states that "The amount of energy
ed to refute bullshit is an order of magnitude bigger than that needed to pr
2]

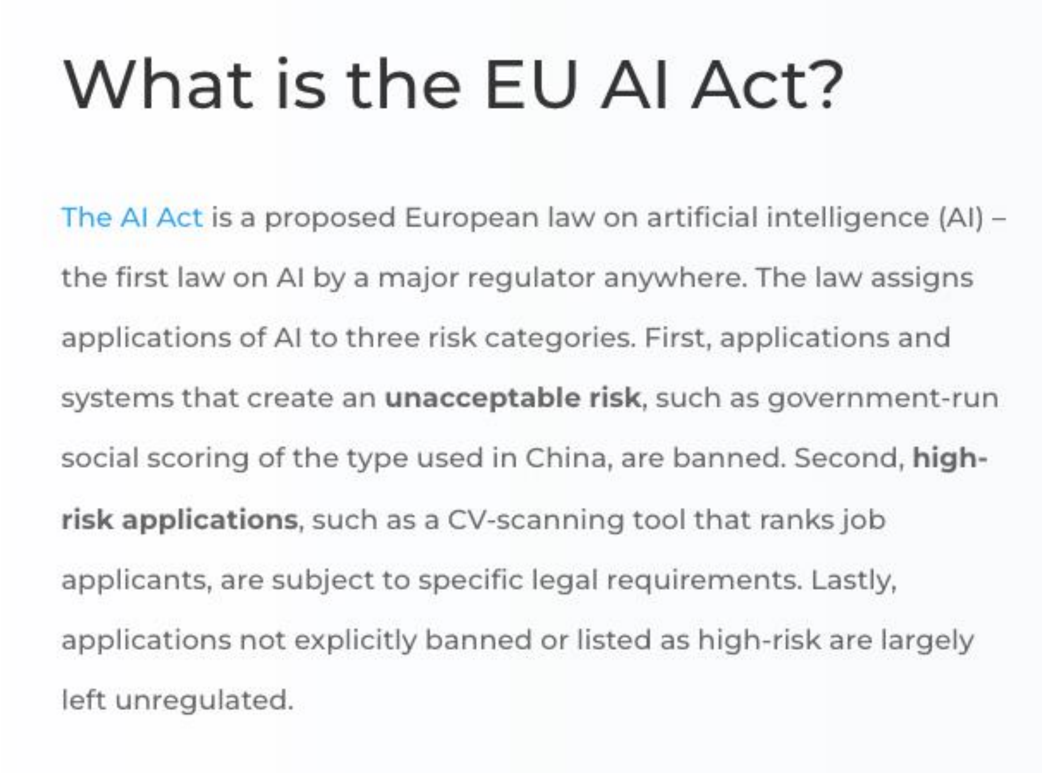
ntents [hide]
rigins
milar concepts
e also
eferences

gins [edit]

aw was publicly formulated the first time in January 2013[3] by Alberto
dolini, an Italian programmer. Brandolini stated that he was inspired by rea
l Kahneman's *Thinking, Fast and Slow* right before watching an Italian pol
rmer Prime Minister Silvio Berlusconi and journalist Marco

The Artificial Intelligence Act

Why should we care?



What is the EU AI Act?

The AI Act is a proposed European law on artificial intelligence (AI) – the first law on AI by a major regulator anywhere. The law assigns applications of AI to three risk categories. First, applications and systems that create an **unacceptable risk**, such as government-run social scoring of the type used in China, are banned. Second, **high-risk applications**, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.

Summary issues we face

- Cyberinfrastructure researchers have considerable power as upstream innovators but trust is fragile
- US is well behind EU in regulatory structures for AI
- Avoiding contributions to misinformation: ChatGPT quickly characterized as contributing to misinformation miasma

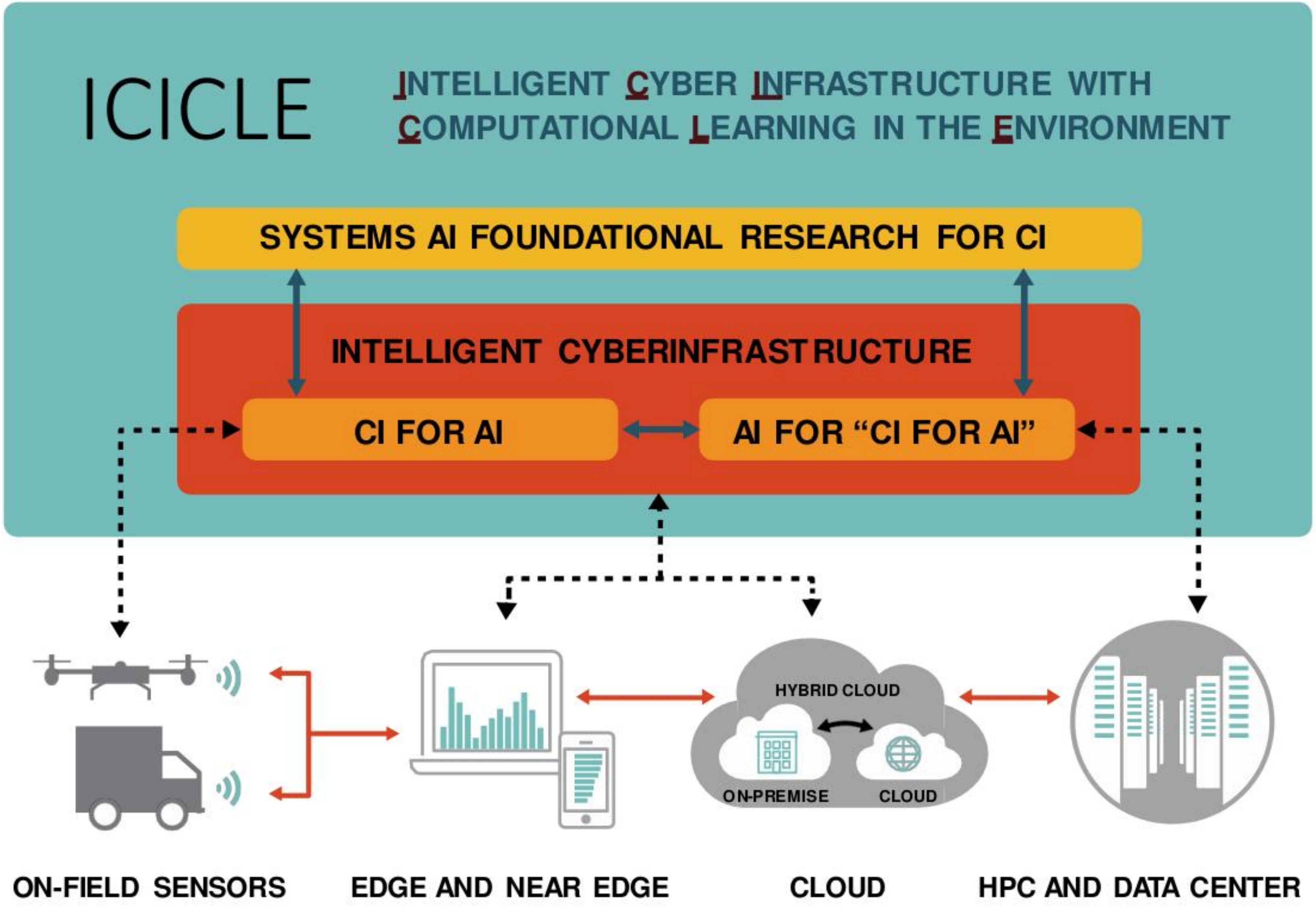
Cholera "tramples the victors & the vanquished both." Robert Seymour. 1831. U.S. National Library of Medicine / Wikipedia, Public Domain.

plale@indiana.edu

How ICICLE frames AI Ethics and Democratization

Intelligent Cyberinfrastructure with Computational Learning in the Environment (ICICLE)

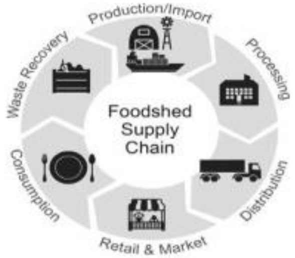
- US National Science Foundation Funded AI Institute
- http://icicle.ai



ICICLE

INTELLIGENT CYBER INFRASTRUCTURE WITH COMPUTATIONAL LEARNING IN THE ENVIRONMENT

USE INSPIRED SCIENCE CASES

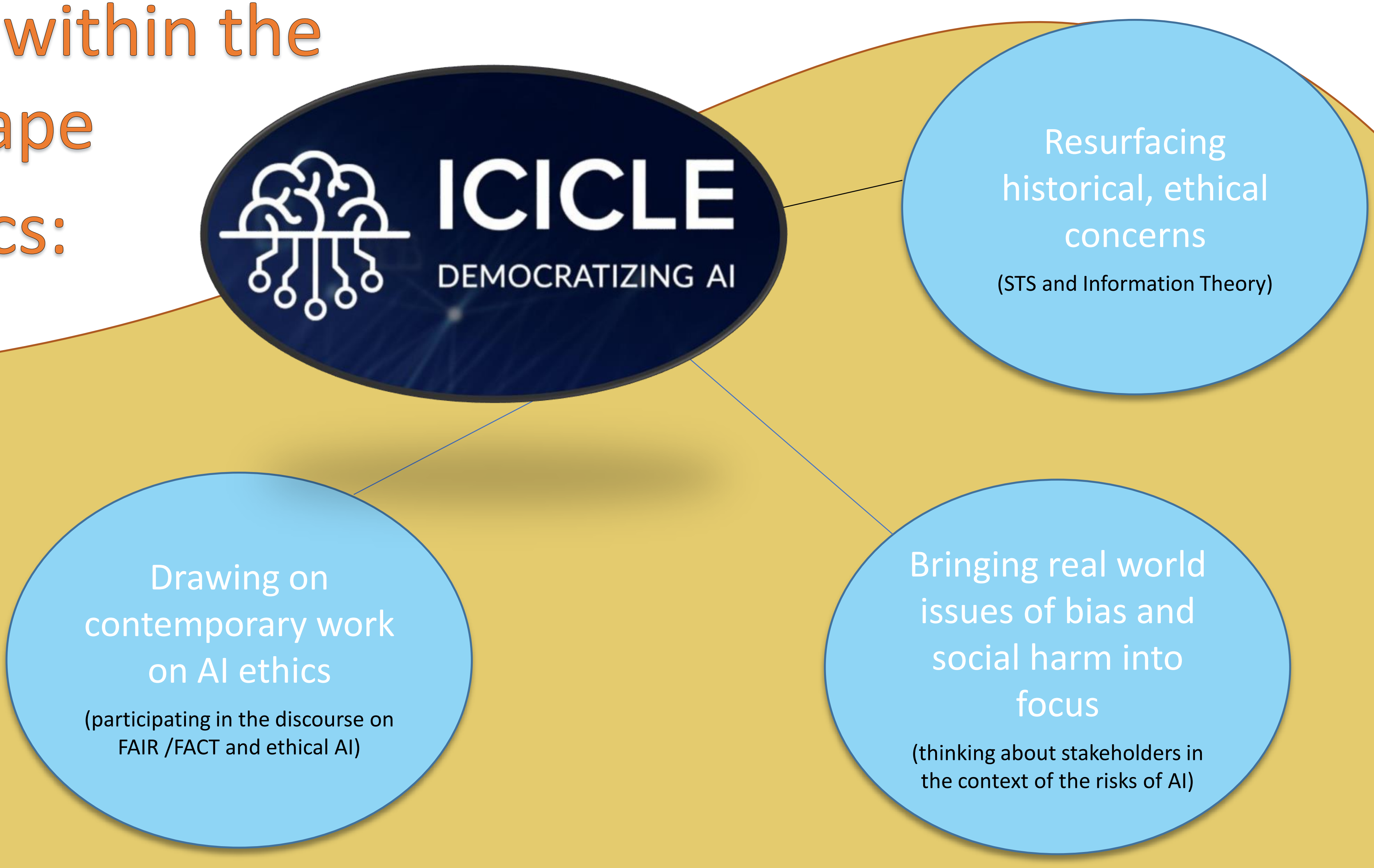


SMART FOODSHEDS

ANIMAL ECOLOGY

DIGITAL AGRICULTURE

Centering ICICLE design within the landscape of ethics:



Resurfacing historical, ethical concerns

(STS and Information Theory)

Drawing on contemporary work on AI ethics

(participating in the discourse on FAIR /FACT and ethical AI)

Bringing real world issues of bias and social harm into focus

(thinking about stakeholders in the context of the risks of AI)

How we do it?

ICICLE

DEMOCRATIZING AI

I. Workforce Development

A. *challenge inevitability··*

B. *encourage forethought··*

II. DEI and BPC (*democratization of AI devt.* to minimize bias)

Build a workforce that considers ethical implications

III. Democratization

A. *Engage end-users* to maximize accessibility & minimize risk

IV. Trustworthiness

A. Use *model cards* to build trust (through accountability and contextuality)

Focus on stakeholders and use-inspired science!

V. Privacy

A. Employ privacy preserving techniques

B. Apply *contextual integrity*/evaluate privacy tradeoff

VI. Fairness

VII. Accountability (accountable to & accountable for)

A. Governance and reporting

B. Utilize KG with *FAIR/FACT principles*

Harness the best methods in privacy, accountability, transparency, and more!

Democratizing Artificial Intelligence

Who gets access to the technologies and what do they do with them (Fischoff 2014)

ICICLE

DEMOCRATIZING AI

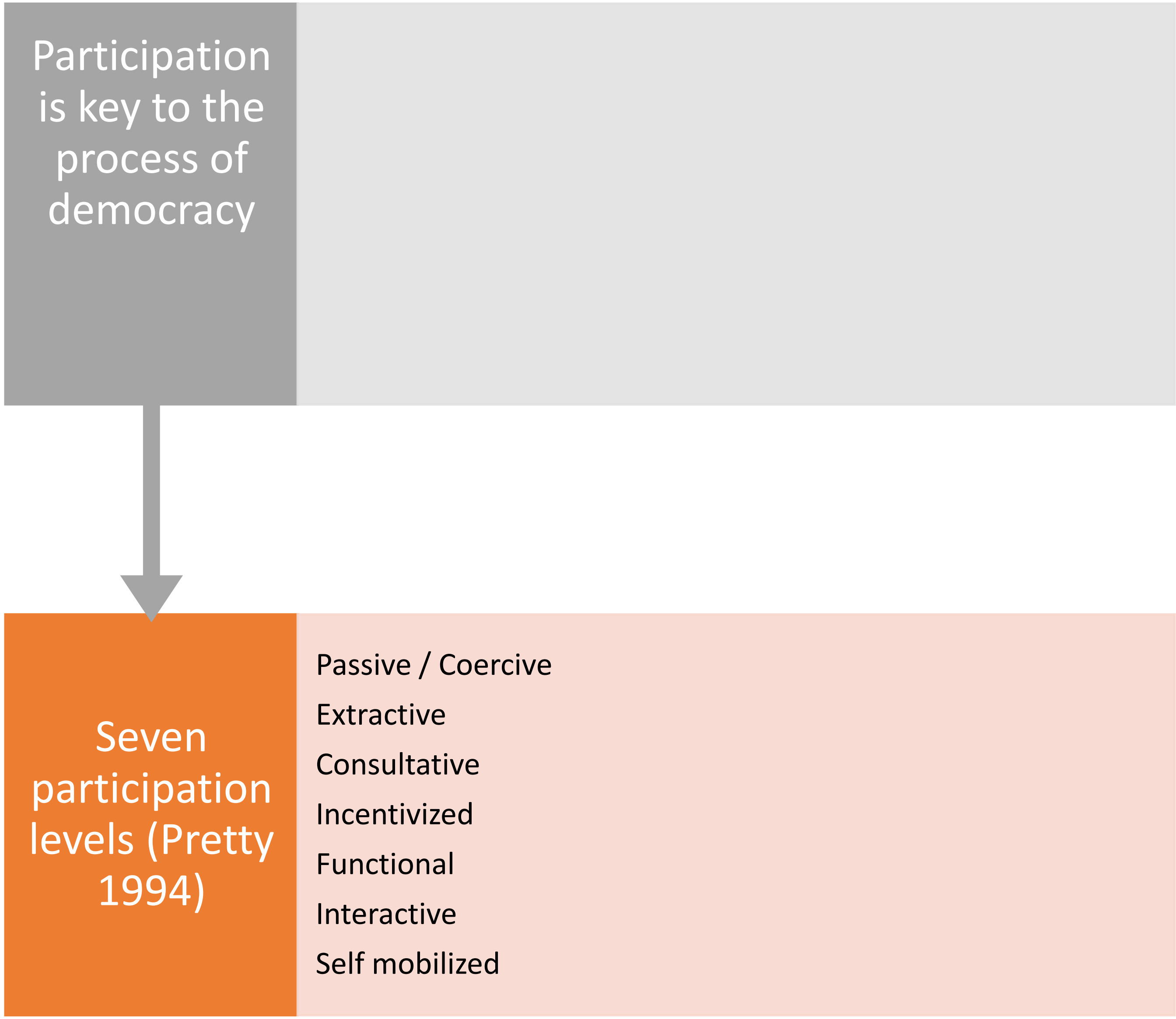
Democratizing AI

- ICICLE’s goal of Democratizing AI is for

  broad and just access to ICICLE AI technologies, and development of its AI technologies in a manner that is informed by those who will benefit or be affected by the technologies

- One purpose is democratizing AI as a measurable objective

Democratizing AI



Participation is key to the process of democracy

Seven participation levels (Pretty 1994)

- Passive / Coercive
- Extractive
- Consultative
- Incentivized
- Functional
- Interactive
- Self mobilized

plale@indiana.edu

Conceptual framework

Democratization of use: Make software, models, data, information and KGs more accessible to a wider range of potential users.

Democratization of AI development. Where a wider range of people contribute to its design and development.

Democratization of AI benefits. All people benefit from AI, not just big and medium-sized tech companies.

Democratization of AI governance There should be a process to facilitate the representation of diverse and conflicting beliefs and values about how people and their actions are governed

Cyberinfrastructure Knowledge Network (CKN)

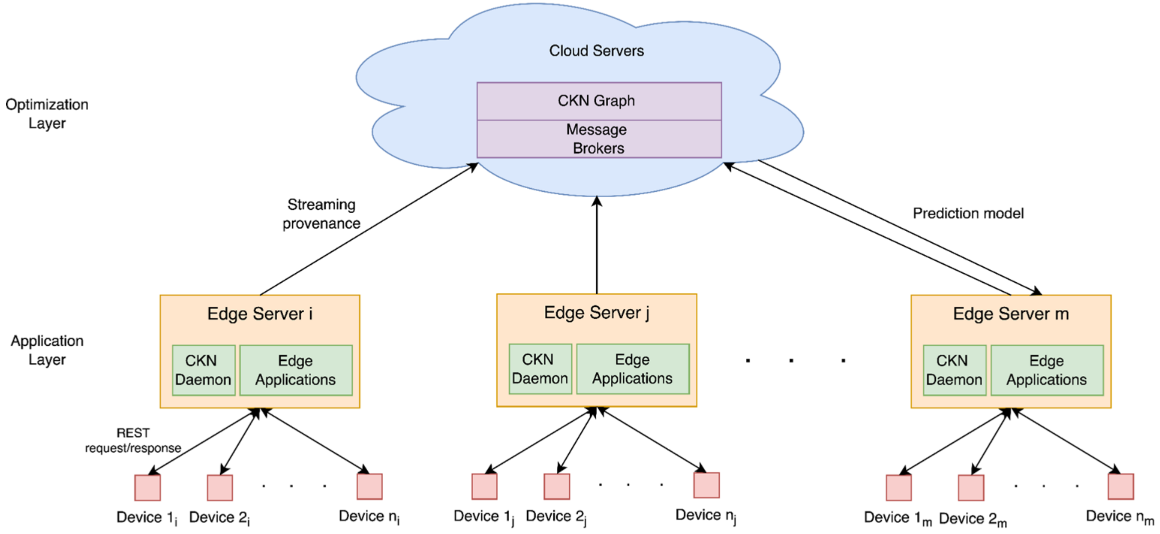
Using Quality of Experience(QoE) for the users in complex Edge AI systems:

- Use case:

  Animal Identification in the field is carried out via deploying Camera Traps in remote edge systems. Camera traps produce images that needs identified via AI models with certain accuracy and latency requirements. These constraint requirements vary over time.

- Problem:

  How can the AI models be deployed over the Edge System to optimize the Quality of Experience observed by the users (camera traps) while addressing the time-series requirement variability?



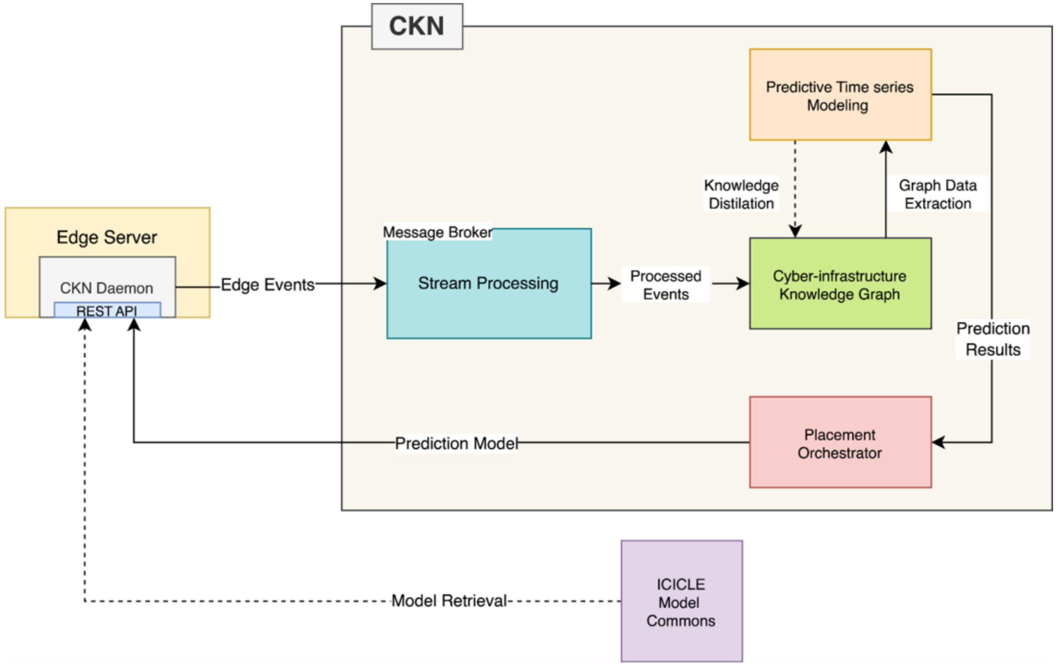
High-level overview of the Edge System

Sachith Withana and Beth Plale, Edge AI Distributed Framework, 19th Int’l Conference on eScience, Oct 2023

Solution

Model the incoming device constraints through time and predictively place models to optimize the Quality of Experience.

- Data Collection
  - Incoming events from the edge devices are captured and sent to a distributed event streaming system.
  - Real-time stream processing aggregates the event streams using tumbling windows and ingests the historical data into the Knowledge Graph
- Decision making
  - Stored historical data is used to model the device constraints through time via training a time-series Deep Learning Model
  - Trained DNN model is placed at the Edge Servers to predictively place the required AI models optimizing the QoE of requests.



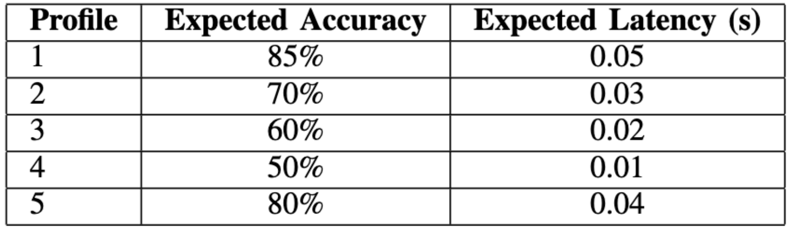
Architectural overview of the Edge-cloud continuum

plale@indiana.edu

Results

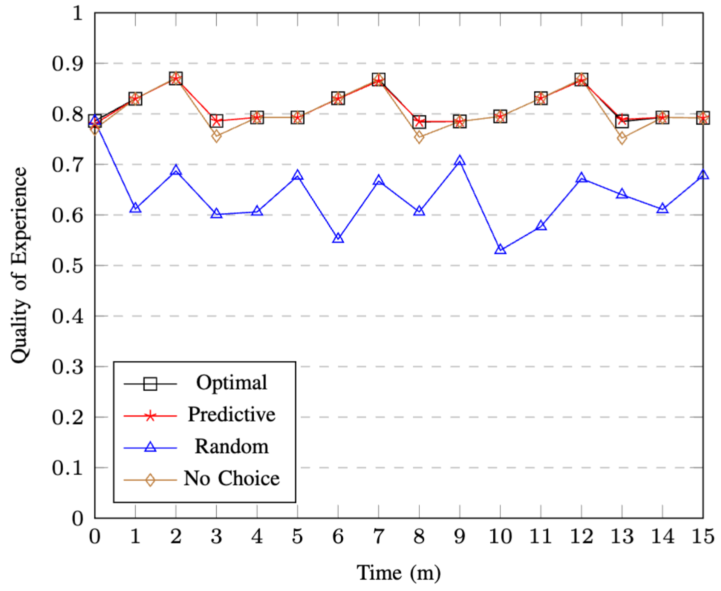
Evaluated the proposed solution on Jetstream 2 infrastructure using the image classification use case with seven readily-available CNN models for image inference trained on ImageNet data.

- Dataset
  - Carefully designed synthetic dataset[1] to model time-variant behavior of incoming requests constraint via device profiling.

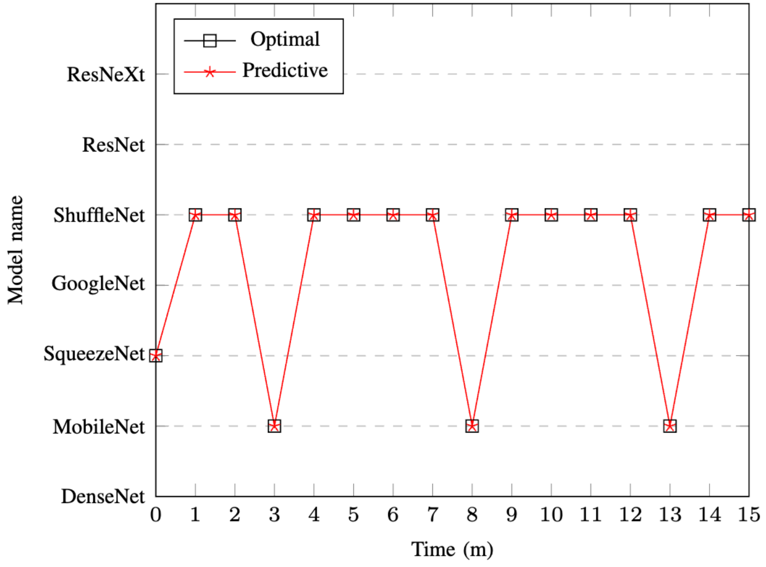


Profile	Expected Accuracy	Expected Latency (s)
1	85%	0.05
2	70%	0.03
3	60%	0.02
4	50%	0.01
5	80%	0.04

- Algorithms compared
  - Predictive placement
  - Optimal placement
  - No-choice placement
  - Random Placement



QoE evaluation of the algorithms over time.



Model placement decisions comparing optimal solution and proposed predictive modeling

[1] Sachith Withana and Beth Plale. CKN Edge AI Dataset for Image inference at the Edge (CEAD). (1.1) [Data set].Zenodo.https://doi.org/10.5281/zenodo.8023205, June 2023.

Takeaways

Evaluate against intent

Evaluate where conversational agents are missing the mark with respect to user intent. In other words, we need more user studies and as accepted part of our research methodologies

Capture true cost in our assessments of our innovations

True Cost Accounting is the balancing of all costs and consequential costs that arise in connection with the production of a product.

How much food would really have to cost if one also included the environmental follow-up costs that arise during production and the entire supply chain.

- 4% price premium on conventional apples,
- 30% on organic mozzarella and
- 173% on conventionally produced meat

Use metrics that capture true cost of products:

e.g., compare cost to build (train, retrain) and execute against (older) alternates

Use that cost to make conscious deployment choices

Except in rare cases, Britain will pay for new drugs only when their effectiveness is high relative to their prices

German regulators may decline to reimburse a new drug at rates higher than those paid for older therapies, if they find that it offers no additional benefit

Computing for the common good

NATIONAL AI RESEARCH INSTITUTES

ACCESS

NSF

Consider the privilege of society’s trust

Common good asks that we as cyberinfrastructure researchers care for the shared interest - that is, the innovations, the infrastructure

It is an obligation to 1) not add to misinformation, 2) not add AI innovations without assessment of potential harms, 3) not add to a resource depleted planet

plale@indiana.edu

NSF

PERVASIVE
TECHNOLOGY INSTITUTE

ICICLE
DEMOCRATIZING AI

Thank you

plale@Indiana.edu