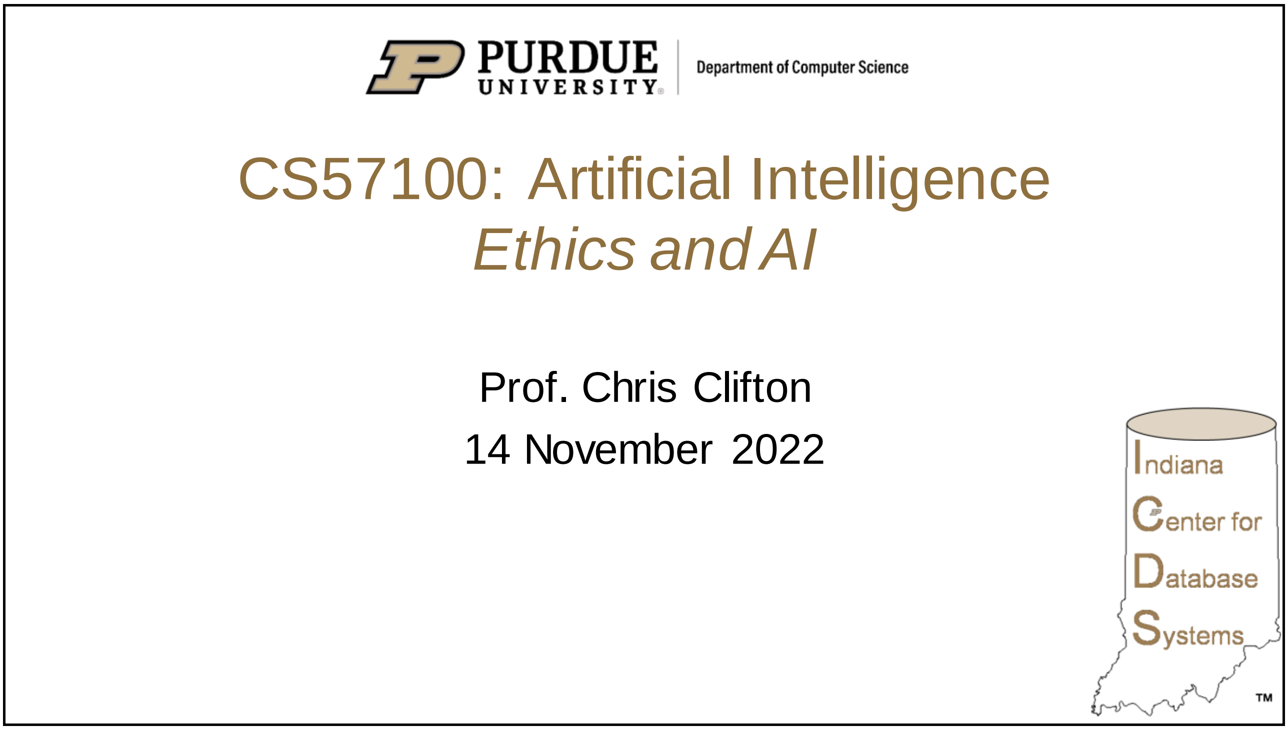


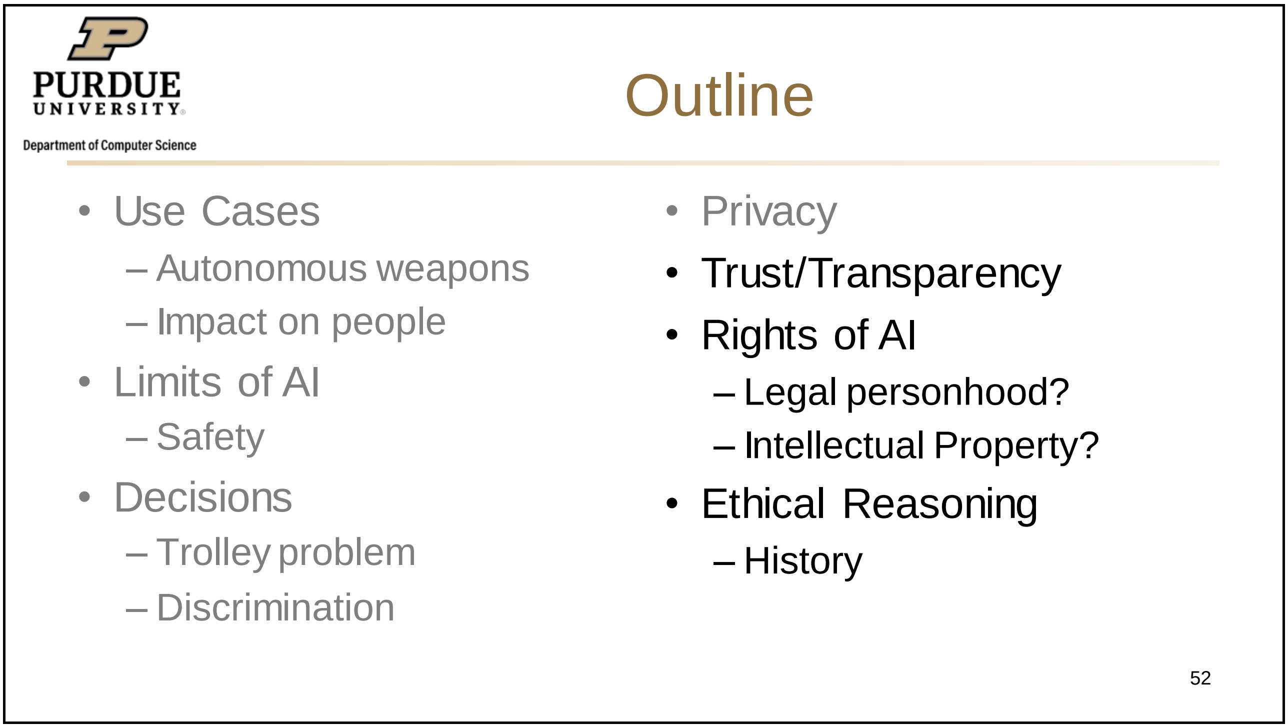
PURDUE UNIVERSITY | Department of Computer Science

CS57100: Artificial Intelligence
Ethics and AI

Prof. Chris Clifton

14 November 2022

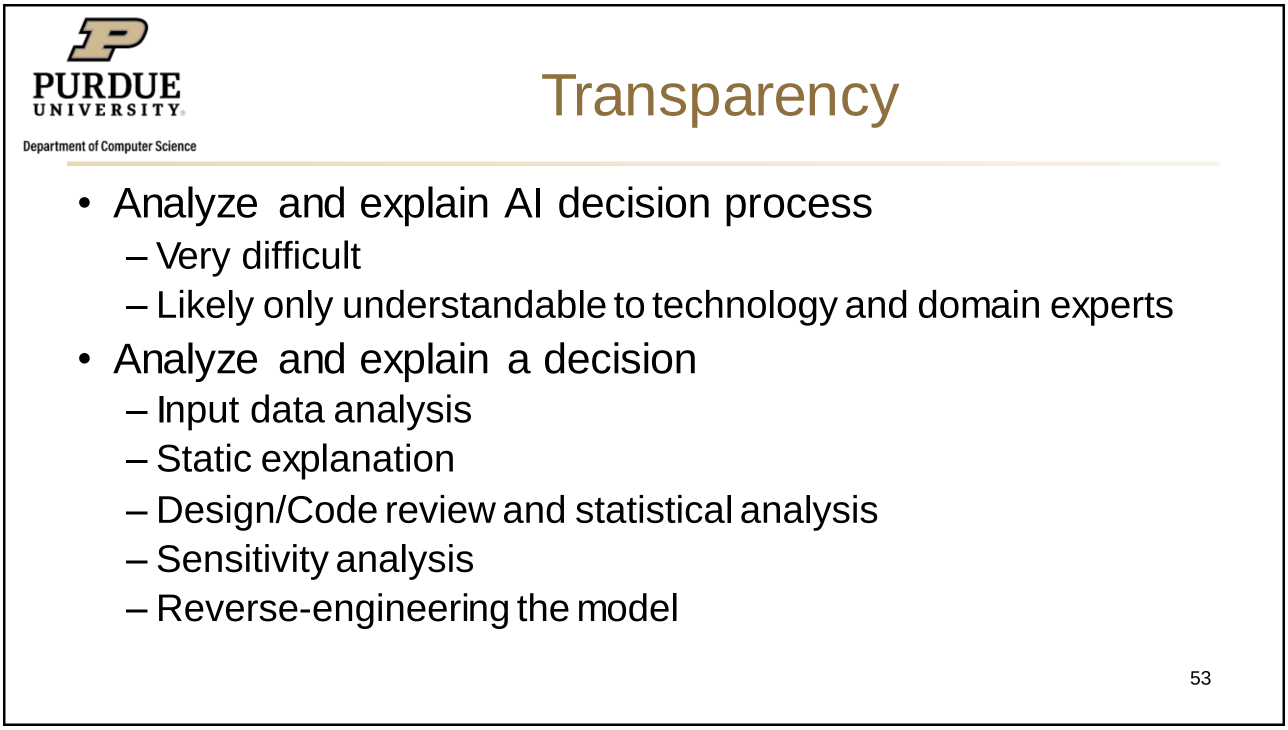
Indiana Center for Database Systems ™



PURDUE UNIVERSITY Department of Computer Science

Outline

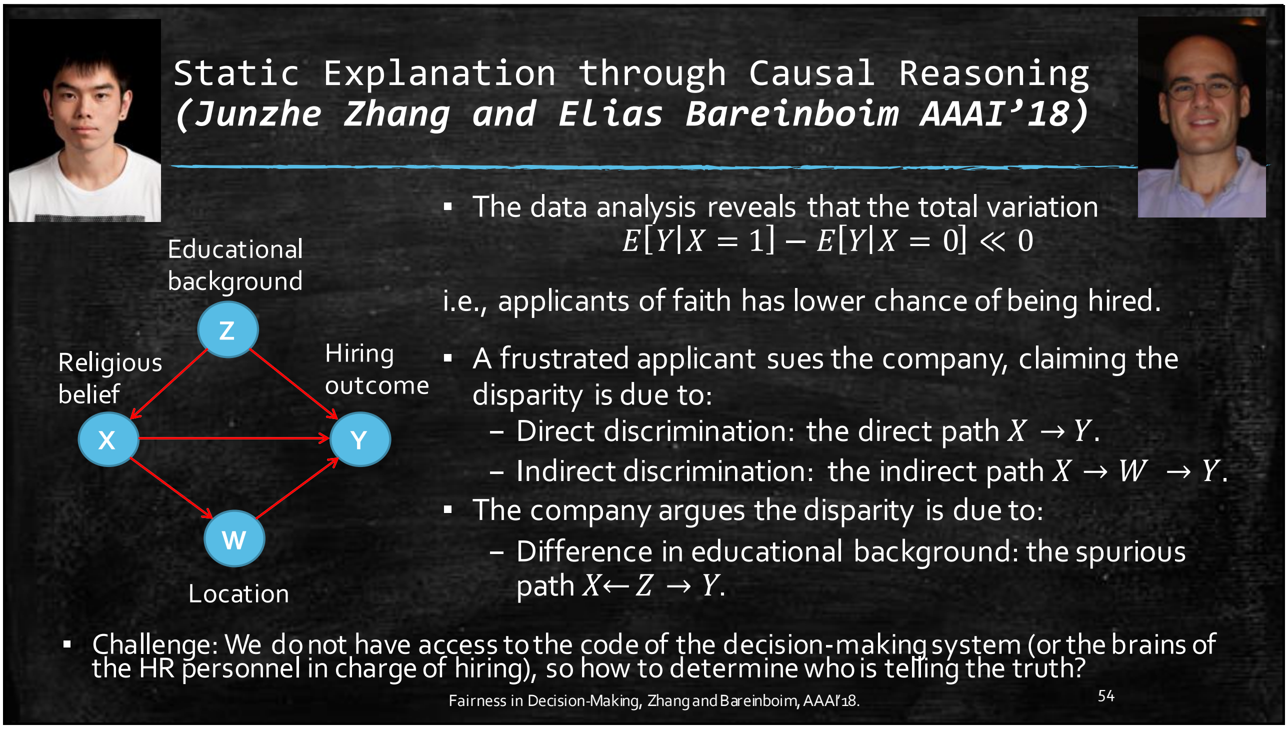
- Use Cases
  - Autonomous weapons
  - Impact on people
- Limits of AI
  - Safety
- Decisions
  - Trolley problem
  - Discrimination
- Privacy
- Trust/Transparency
- Rights of AI
  - Legal personhood?
  - Intellectual Property?
- Ethical Reasoning
  - History



PURDUE UNIVERSITY
Department of Computer Science

Transparency

- Analyze and explain AI decision process
  - Very difficult
  - Likely only understandable to technology and domain experts
- Analyze and explain a decision
  - Input data analysis
  - Static explanation
  - Design/Code review and statistical analysis
  - Sensitivity analysis
  - Reverse-engineering the model



Static Explanation through Causal Reasoning *(Junzhe Zhang and Elias Bareinboim AAAI'18)*

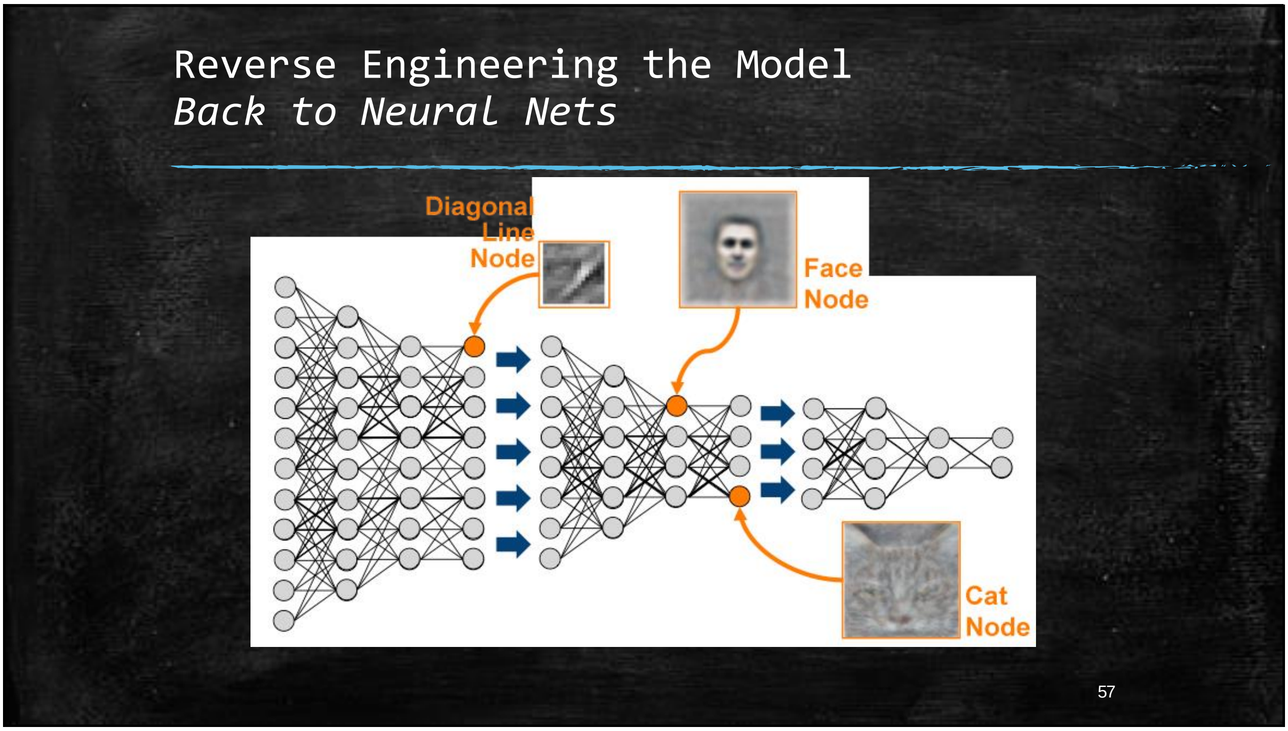
- The data analysis reveals that the total variation
$$E[Y|X=1]-E[Y|X=0]\ll 0$$

i.e., applicants of faith has lower chance of being hired.

- A frustrated applicant sues the company, claiming the disparity is due to:
  - Direct discrimination: the direct path $X\rightarrow Y$.
  - Indirect discrimination: the indirect path $X\rightarrow W\ \rightarrow Y$.
- The company argues the disparity is due to:
  - Difference in educational background: the spurious path $X\leftarrow Z\ \rightarrow Y$.

- Challenge: We do not have access to the code of the decision-making system (or the brains of the HR personnel in charge of hiring), so how to determine who is telling the truth?

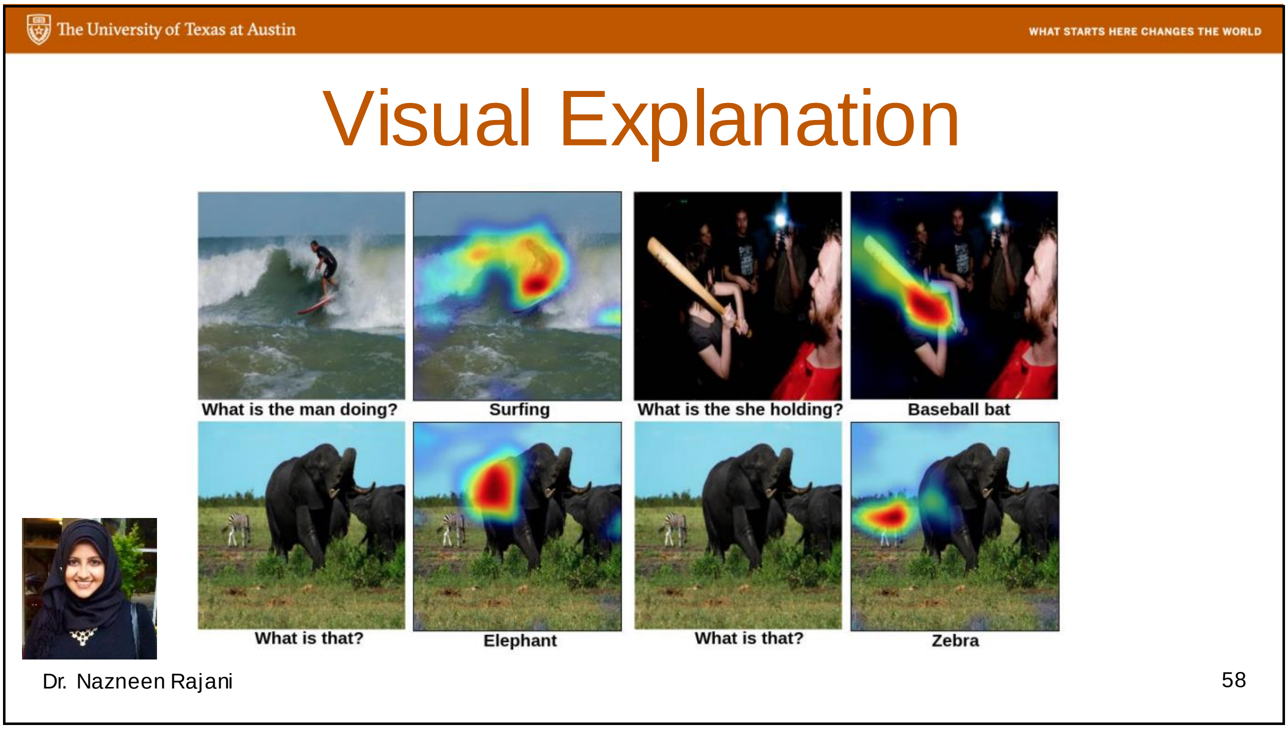
Fairness in Decision-Making, Zhang and Bareinboim, AAAI'18.



Reverse Engineering the Model

Back to Neural Nets

57



The University of Texas at Austin

WHAT STARTS HERE CHANGES THE WORLD

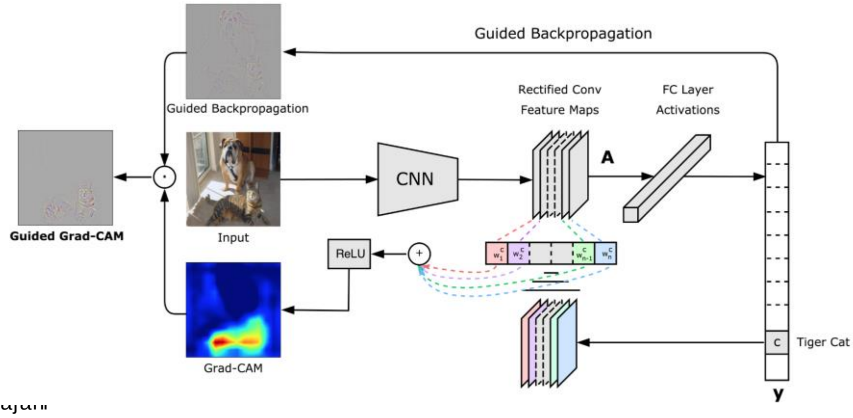
Visual Explanation

Dr. Nazneen Rajani

58

Generating Visual Explanation

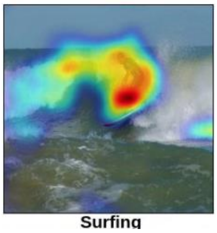
- *GradCAM* (Selvaraju et al., 2017) is used to generate heat-map explanations.



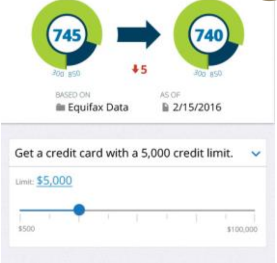
PURDUE UNIVERSITY
Department of Computer Science

Are Explanations Accurate?

- Do these explanations really capture how decisions are made?
  - Sensitivity Analysis, Causal Reasoning
    - Explain outcome, not process
  - Heat maps
    - maybe?
- But does it matter?



Surfing

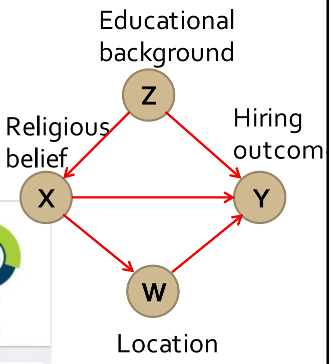


745 → 740
300 850 ↓5 300 850
BASED ON Equifax Data
AS OF 2/15/2016

Get a credit card with a 5,000 credit limit.

Limit: $5,000

$500 $100,000

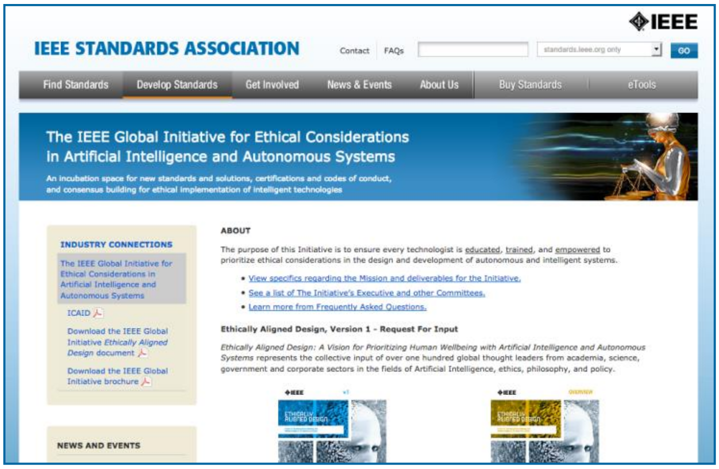


PURDUE
UNIVERSITY
Department of Computer Science

Emotional vs. Rational Decision-Making

- Humans have been shown to be emotional in their decision making
  - fMRI analysis of how decisions are made *(De Martino, Kumaran, Seymour, Dolan, Science 2006)*
- We rationalize our decisions
  - Explanations justify why we the decisions are good, not how we make them
- Is this good enough for explaining AI?
  - *Does this qualify as making ethical decisions?*

What do we do about it?
Standards and Best Practices



The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

An incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies

INDUSTRY CONNECTIONS

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

ICAID

Download the IEEE Global Initiative *Ethically Aligned Design* document

Download the IEEE Global Initiative brochure

NEWS AND EVENTS

ABOUT

The purpose of this Initiative is to ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems.

- View specifics regarding the Mission and deliverables for the Initiative.
- See a list of The Initiative's Executive and other Committees.
- Learn more from Frequently Asked Questions.

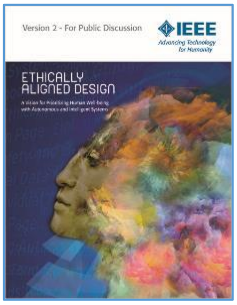
Ethically Aligned Design, Version 1 - Request For Input

Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems represents the collective input of over one hundred global thought leaders from academia, science, government and corporate sectors in the fields of Artificial Intelligence, ethics, philosophy, and policy.

IEEE STANDARDS ASSOCIATION IEEE

Ethically Aligned Design

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems


Version 2 - For Public Discussion
IEEE
Advancing Technology for Humanity
ETHICALLY ALIGNED DESIGN
A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

Version 2

- Launched December 2017 as a Request for Input
- Created by over 250 Global A/IS & Ethics professionals, in a bottom up, transparent, open and increasingly globally inclusive process
- Incorporates over 200 pages of feedback from public RFI and new Working Groups from China, Japan, Korea and more
- Thirteen Committees / Sections
- Contains **over one hundred twenty** key Issues and Candidate Recommendations

https://ethicsinaction.ieee.org/

IEEE STANDARDS ASSOCIATION IEEE

IEEE P70xx Standards Projects

IEEE P7000: Model Process for Addressing Ethical Concerns During System Design

IEEE P7001: Transparency of Autonomous Systems

IEEE P7002: Data Privacy Process

IEEE P7003: Algorithmic Bias Considerations

IEEE P7004: Child and Student Data Governance

IEEE P7005: Employer Data Governance

IEEE P7006: Personal Data AI Agent Working Group

IEEE P7007: Ontological Standard for Ethically Driven Robotics and Automation

IEEE P7008: Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems

IEEE P7009: Fail-Safe Design of Autonomous and Semi-Autonomous Systems

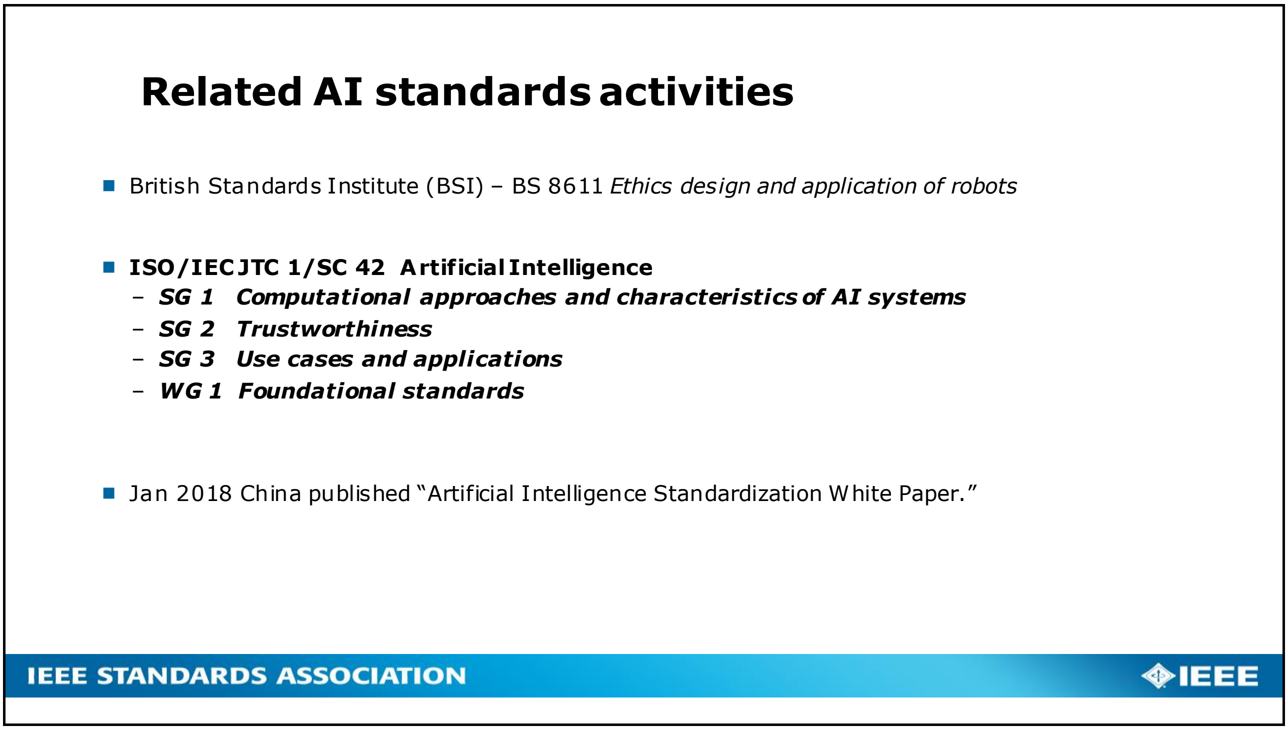
IEEE P7010: Wellbeing Metrics Standard for Ethical AI and Autonomous Systems

IEEE P7011: Process of Identifying and Rating the Trustworthiness of News Sources

IEEE P7012: Standard for Machines Readable Personal Privacy Terms

IEEE STANDARDS ASSOCIATION IEEE

65

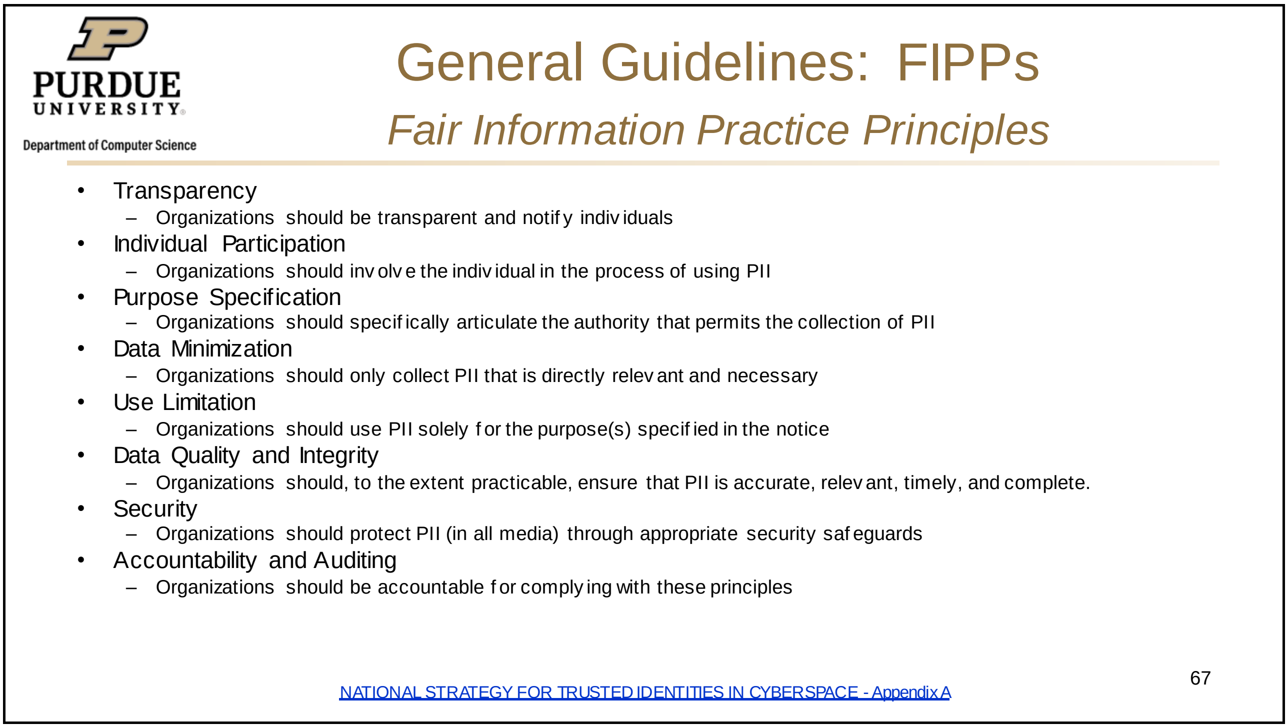


Related AI standards activities

- British Standards Institute (BSI) – BS 8611 *Ethics design and application of robots*
- **ISO/IEC JTC 1/SC 42 Artificial Intelligence**
  - ***SG 1 Computational approaches and characteristics of AI systems***
  - ***SG 2 Trustworthiness***
  - ***SG 3 Use cases and applications***
  - ***WG 1 Foundational standards***
- Jan 2018 China published "Artificial Intelligence Standardization White Paper."

IEEE STANDARDS ASSOCIATION

IEEE



PURDUE UNIVERSITY

Department of Computer Science

General Guidelines: FIPPs

Fair Information Practice Principles

- Transparency
  - Organizations should be transparent and notify individuals
- Individual Participation
  - Organizations should involve the individual in the process of using PII
- Purpose Specification
  - Organizations should specifically articulate the authority that permits the collection of PII
- Data Minimization
  - Organizations should only collect PII that is directly relevant and necessary
- Use Limitation
  - Organizations should use PII solely for the purpose(s) specified in the notice
- Data Quality and Integrity
  - Organizations should, to the extent practicable, ensure that PII is accurate, relevant, timely, and complete.
- Security
  - Organizations should protect PII (in all media) through appropriate security safeguards
- Accountability and Auditing
  - Organizations should be accountable for complying with these principles

67

NATIONAL STRATEGY FOR TRUSTED IDENTITIES IN CYBERSPACE - Appendix A

PURDUE
UNIVERSITY
Department of Computer Science

Outline

- Use Cases
  - Autonomous weapons
  - Impact on people
- Limits of AI
  - Safety
- Decisions
  - Trolley problem
  - Discrimination
- Privacy
- Trust/Transparency
- Rights of AI
  - Legal personhood?
  - Intellectual Property?
- Ethical Reasoning
  - History

PURDUE
UNIVERSITY
Department of Computer Science

Rights of AI

- Can a machine have legal rights?
  - Animals do
  - Corporations, too
- What sort of rights should a machine have?
  - Rights of corporations?
  - Existence / not be “unplugged”?

Rights of AI: Intellectual Property

- U.S.: Only people own patents
  - Ever seen IBM or Google as the inventor in a patent?
  - Australia, South Africa have listed AI systems as inventors on patents
- What about copyright?
  - Australia, U.S. – copyright can only be awarded to a person
  - Is AI-generated art then public domain (uncopyrightable?)
    - Entertainment industry exploring this…

Rights of AI: Intellectual Property
UK Intellectual Property Office

“Consultation” updated 28 June 2022

- Copyright for AI-Generated Works
  - Currently protected under UK law
  - Plan no changes, but envisions potential for future changes
- Text/Data Mining
  - Plan to introduce copyright exception to allow TDM for any purpose
  - Still have safeguards for copyright holders
- Patent for AI Inventions
  - Currently AI *cannot* be held to be an inventor
    - But neither can human who invented the AI (unless involved in the invention)
  - As with copyright, no changes, but continue to review to support UK economic interests

Ethical Reasoning

- Ethical: Of or relating to moral principles
- Moral (of an action): having the property of being right or wrong, voluntary or deliberate and therefore open to ethical appraisal
- Ethical Reasoning in the context of AI ([NSW Government]()):
  - A process of identifying ethical issues and weighing multiple perspectives to make informed decisions
  - Not about knowing right from wrong, but being able to think about and respond to a problem fairly, justly, and responsibly

Some suggestions

- Attend relevant talks
  - CS colloquium series (lists.purdue.edu – cs-colloq)
  - [www.purdue.edu/critical-data-studies]()
- Data Ethics courses (a few)
  - ILS 23000: Data Science and Society: Ethical, Legal, Social Issues
  - PHIL 20700: Ethics for Technology, Engineering, and Design
  - PHIL 20800: Ethics of Data Science