ACM Intelligent User Interfaces Conference, April 13, 2021
https://iui.acm.org/2021/hcai_tutorial.html

Human-Centered AI: Reliable, Safe & Trustworthy Part 2

Ben Shneiderman **@benbendc**

Founding Director (1983-2000), Human-Computer Interaction Lab
Professor, Department of Computer Science

Member, National Academy of Engineering

UNIVERSITY OF MARYLAND

PETER WALL INSTITUTE FOR ADVANCED STUDIES
THE UNIVERSITY OF BRITISH COLUMBIA VANCOUVER

Photo: BK Adams

Interdisciplinary research community
- Computer Science & Info Studies
- Psych, Socio, Educ, Jour & MITH

hcil.umd.edu
vimeo.com/72440805

Annual Symposium: FREE Virtual

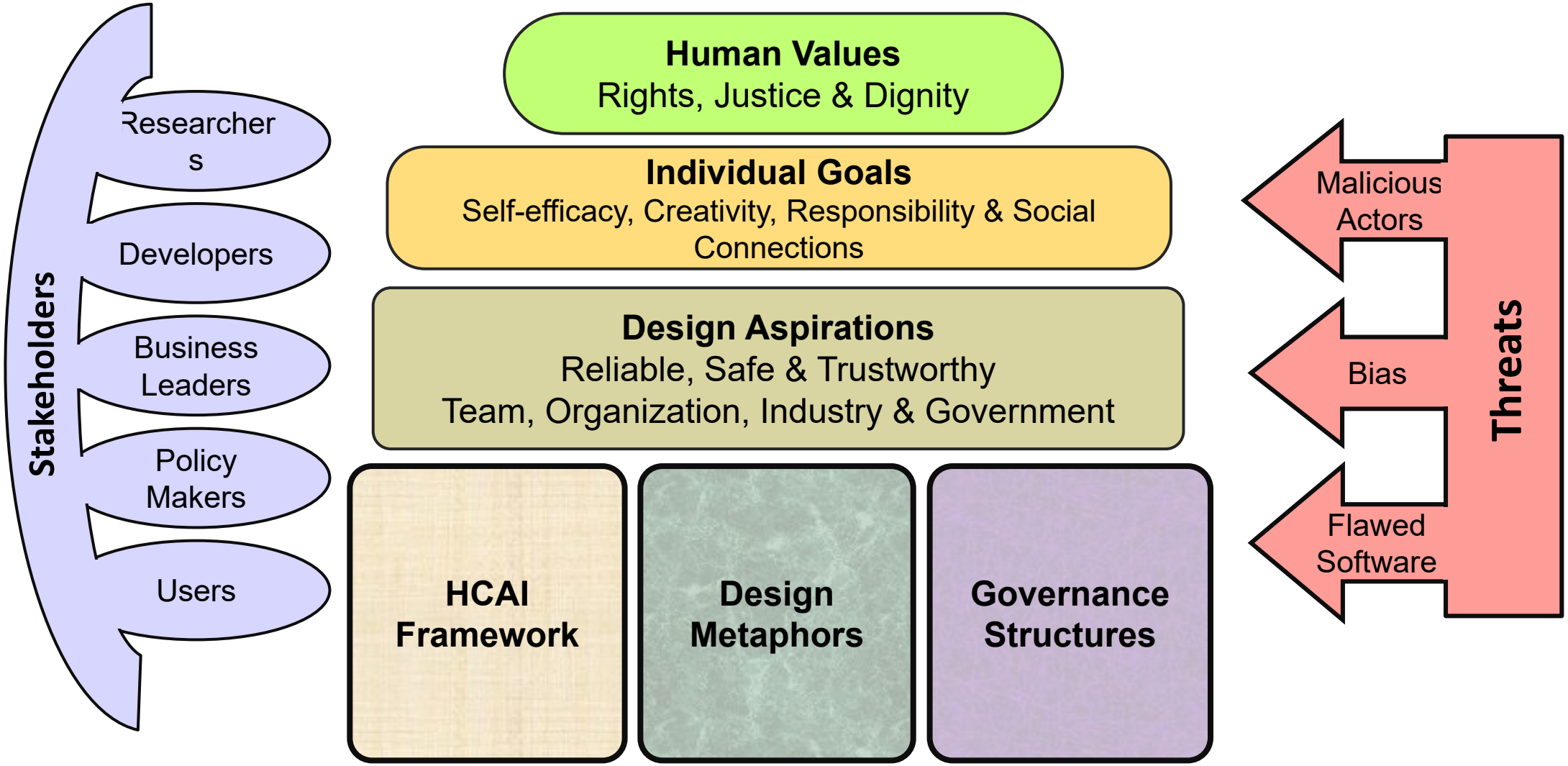
May 27, 2021, Thursday

https://hcil.umd.edu/2021-symposium/

What is Human-Centered AI?

Amplify, Augment, Empower & Enhance People

Human-Centered AI



Oxford University Press (Early 2022) **https://hcil.umd.edu/human-centered-ai/**

Governance Structures

ARTIFICIAL INTELLIGENCE AND LIFE IN 2030

Stanford University

SEPTEMBER 2016

NATIONAL STRATEGY FOR ARTIFICIAL INTELLIGENCE #AIFORALL (INDIA)

Pew Research Center

Artificial Intelligence and the Future of Humans

House of Commons

Science and Technology Committee

Algorithms in decision-making

European Group on Ethics in Science and New Technologies

Artificial Intelligence, Robotics and 'Autonomous' Systems

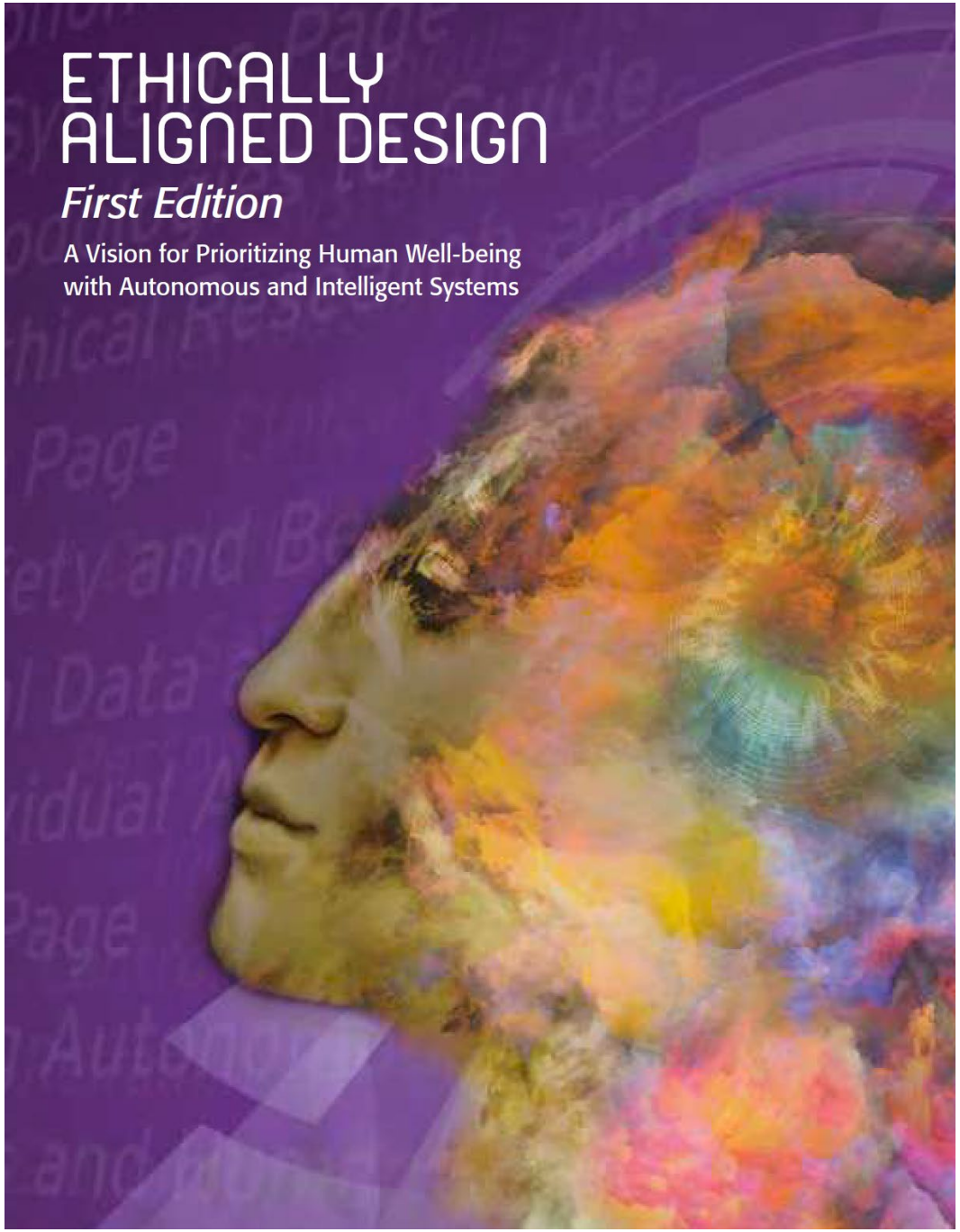
HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE

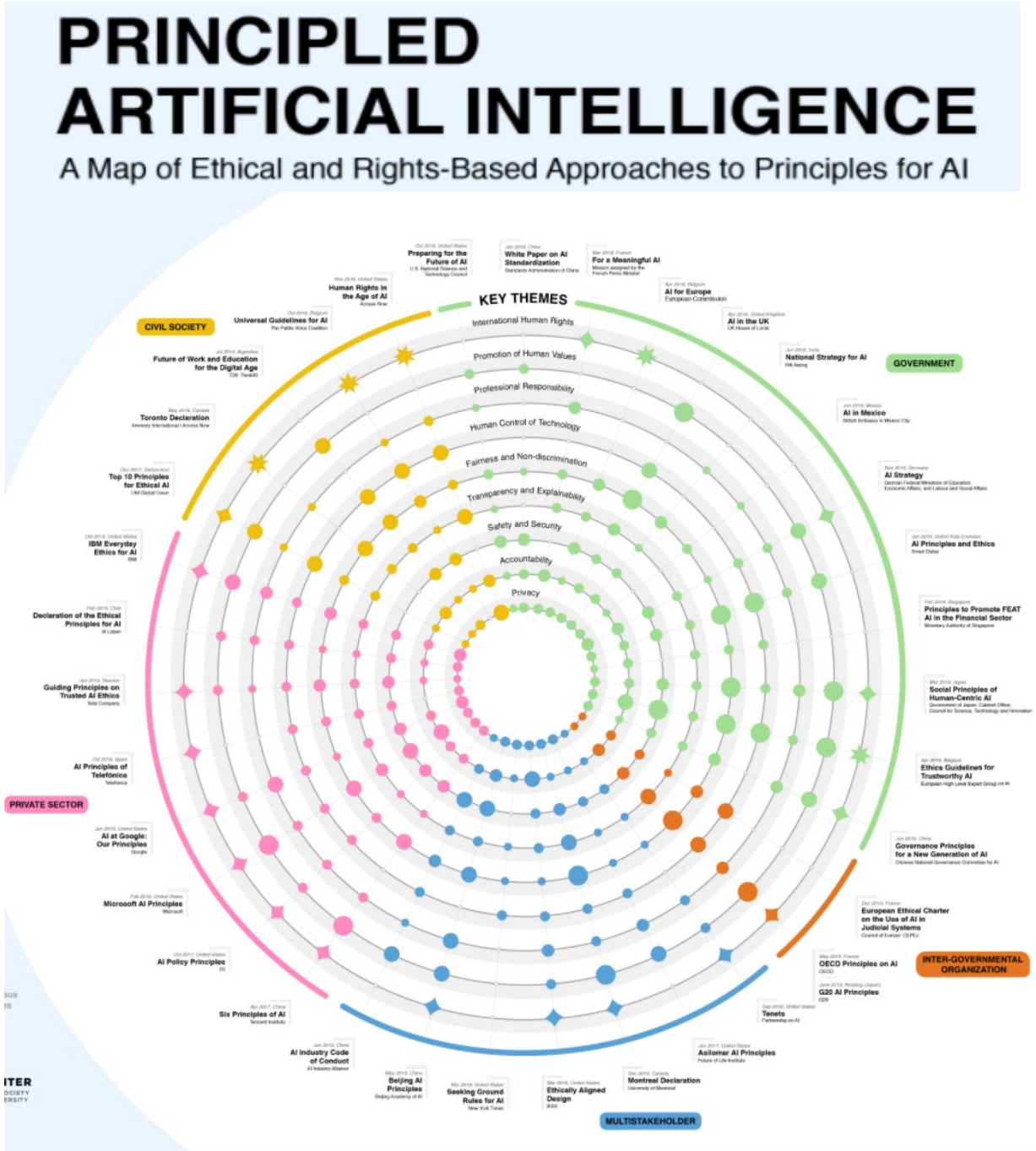
accessnow

Governing Artificial Intelligence: UPHOLDING HUMAN RIGHTS & DIGNITY

Mark Latonero

Data&Society


ETHICALLY
ALIGNED DESIGN
First Edition
A Vision for Prioritizing Human Well-being
with Autonomous and Intelligent Systems

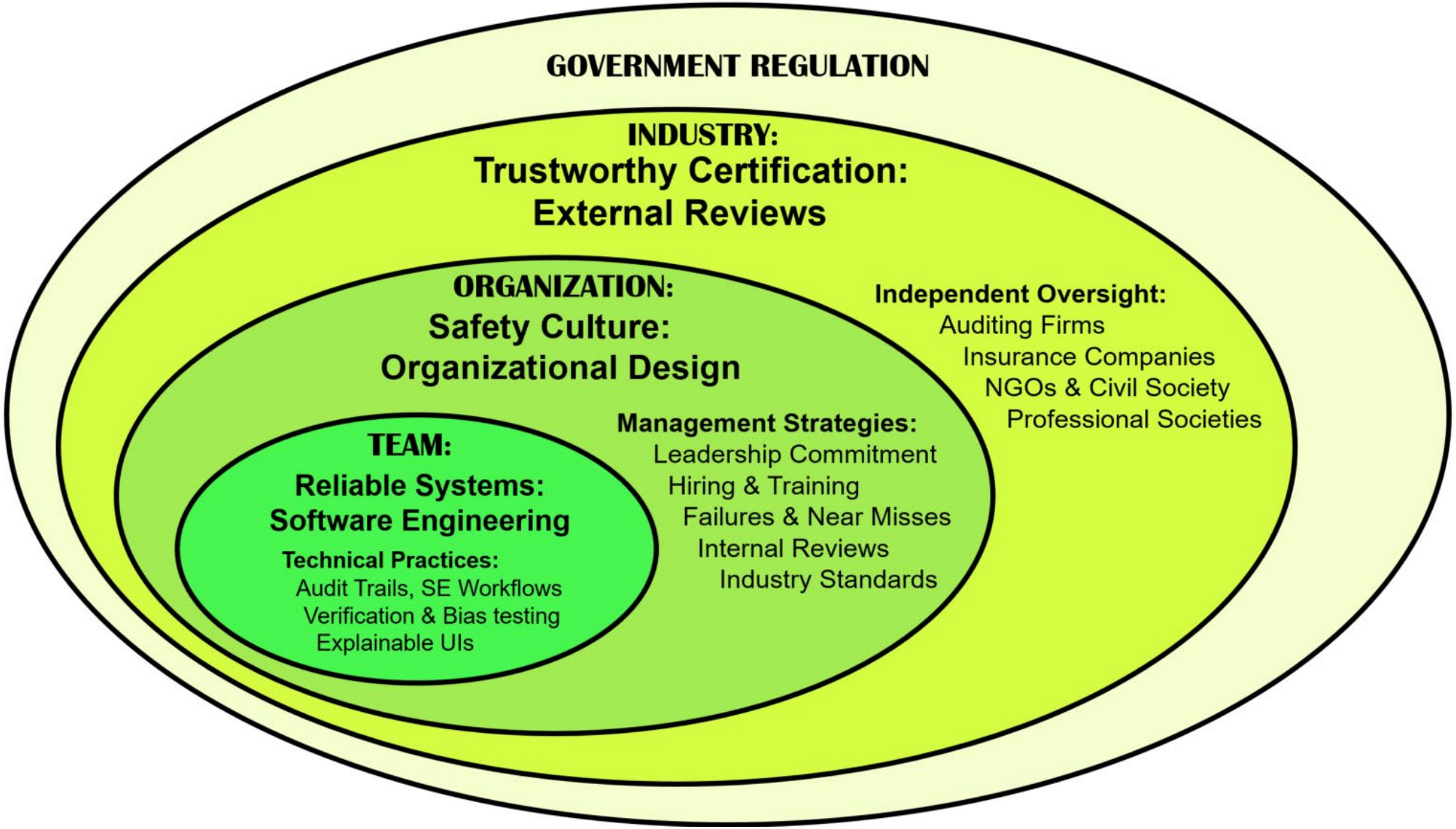

PRINCIPLED
ARTIFICIAL INTELLIGENCE
A Map of Ethical and Rights-Based Approaches to Principles for AI
KEY THEMES
International Human Rights
Promotion of Human Values
Professional Responsibility
Human Control of Technology
Fairness and Non-discrimination
Transparency and Explainability
Safety and Security
Accountability
Privacy
CIVIL SOCIETY
GOVERNMENT
PRIVATE SECTOR
INTER-GOVERNMENTAL ORGANIZATION
MULTISTAKEHOLDER
Preparing for the Future of AI
White Paper on AI Standardization
For a Meaningful AI
Human Rights in the Age of AI
AI for Europe
Universal Guidelines for AI
AI in the UK
Future of Work and Education for the Digital Age
National Strategy for AI
Toronto Declaration
AI in Mexico
Top 10 Principles for Ethical AI
AI Strategy
IBM Everyday Ethics for AI
AI Principles and Ethics
Declaration of the Ethical Principles for AI
Principles to Promote FEAT AI in the Financial Sector
Guiding Principles on Trusted AI Ethics
Social Principles of Human-Centric AI
AI Principles of Telefónica
Ethics Guidelines for Trustworthy AI
AI at Google: Our Principles
Governance Principles for a New Generation of AI
Microsoft AI Principles
European Ethical Charter on the Use of AI in Judicial Systems
AI Policy Principles
OECD Principles on AI
G20 AI Principles
Six Principles of AI
Tenets
AI Industry Code of Conduct
Asilomar AI Principles
Beijing AI Principles
Seeking Ground Rules for AI
Ethically Aligned Design
Montreal Declaration

Ethical AI Principles

	Berkman Klein Center	IEEE Ethically Aligned Design
Close Match	Accountability	Accountability
	Transparency & explainability	Transparency
	Promotion of human values	Human rights
	Safety & security	Well-being
Similar	Human control of technology	Effectiveness
	Fairness & non-discrimination	Awareness of misuse
	Professional responsibility	Competence
	Privacy	Data agency

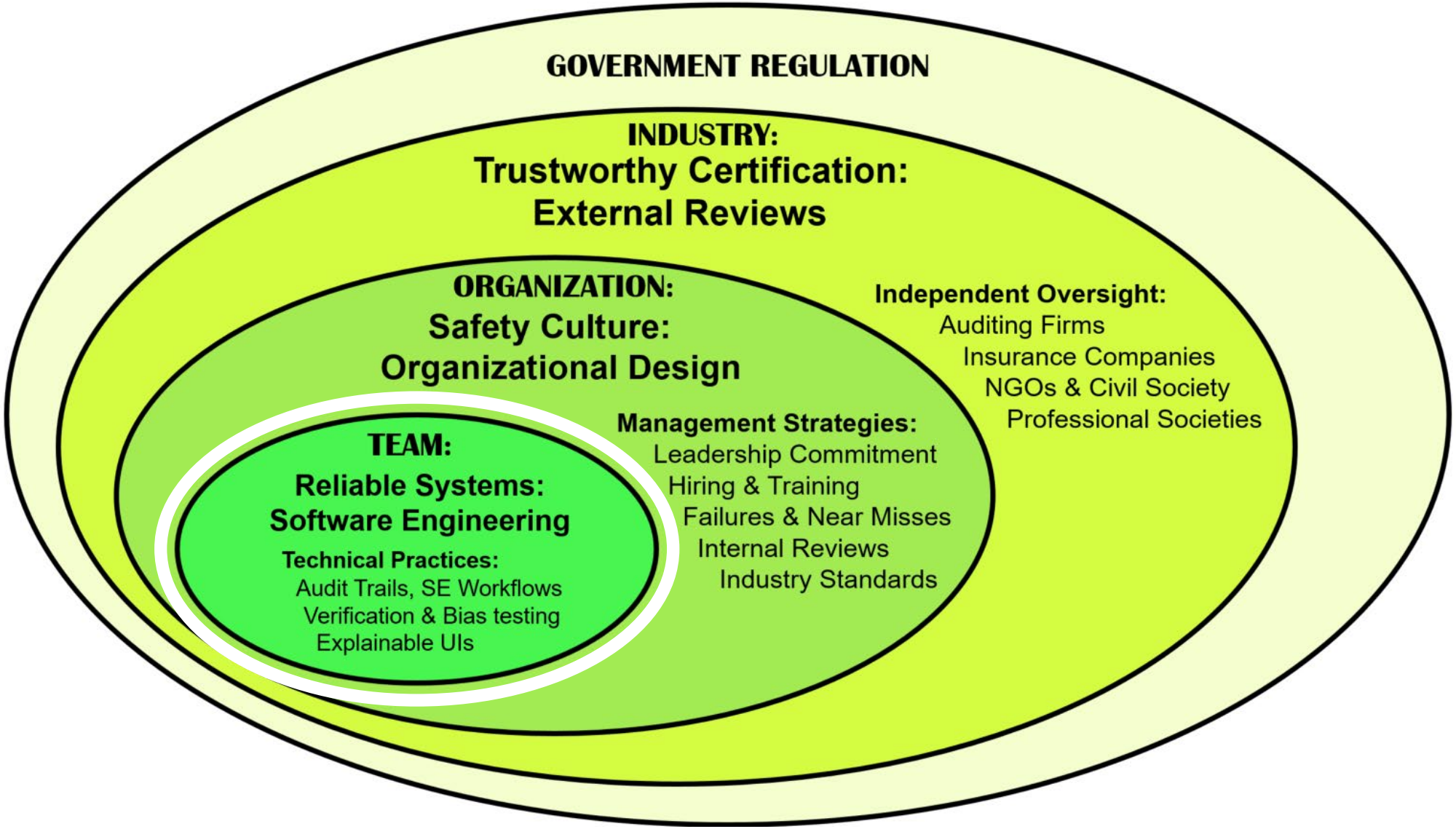
https://cyber.harvard.edu/publication/2020/principled-ai
https://ethicsinaction.ieee.org/

Governance Structures for Human-Centered AI



ACM TIIS (Oct 2020) **https://dl.acm.org/doi/10.1145/3419764**

Governance Structures for Human-Centered AI



ACM TIIS (Oct 2020) **https://dl.acm.org/doi/10.1145/3419764**

TEAM

Reliable systems based on software engineering practices

1) Audit trails and analysis tools
2) Software engineering workflows
3) Verification & validation testing
4) Bias testing to improve fairness
5) Explainable user interfaces

Reliable Systems

Software engineering practices for a TEAM

1) Audit trails and analysis tools

"Flight Data Recorder for Every Robot"

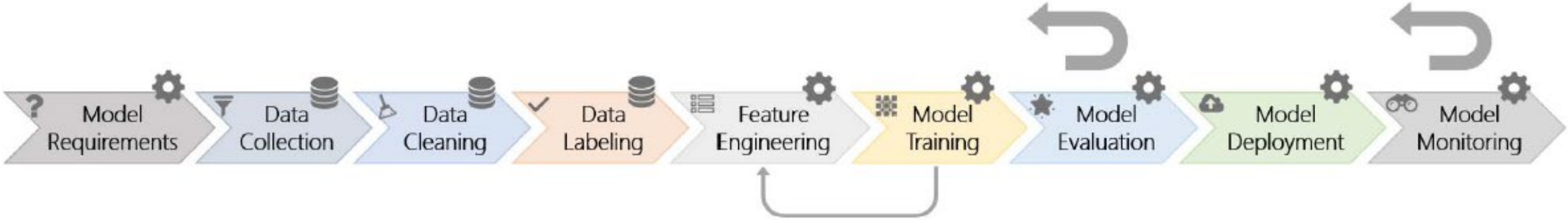
- Retrospective analysis of failures
- Understanding near misses
- Analysis to support preventive maintenance

Reliable Systems

Software engineering practices for a TEAM

2) Software engineering workflows

Microsoft's Workflow for Machine Learning



(Amershi, et al., IEEE/ACM ICSE 2019)

Reliable Systems

Software engineering practices for a TEAM

2) Software engineering workflows

Google Workflow for Algorithmic Auditing

1. Scoping: identify project scope & audit, raise questions of risk
2. Mapping: create stakeholder map & collaborator contact list, conduct interviews & select metrics
3. Artifact Collection: document design process, datasets & machine learning models
4. Testing: conduct adversarial testing to probe edge cases & failure possibilities
5. Reflection: consider risk analysis, failure remediation & record design history

(Raji et al., ACM FAT* 2020)

Reliable Systems

Software engineering practices for a TEAM

3) Verification & validation testing

- Traditional case-based
- Metamorphic
- Differential
- User Experience
- Red Teams

+ Microsoft's Datasheets for Datasets
+ Google's Model Cards
+ IBM FactSheets
+ Track history of bugs, problems, concerns

Reliable Systems

Software engineering practices for a TEAM

4) Bias testing to improve fairness

- Pre-existing, Technical, Emergent Bias
  (Friedman & Nissenbaum, 1996; Baeza-Yates, 2018)

- Facial Recognition Intersectional Bias
  (Buolamwini & Gebru, 2019; *Coded Bias*, 2021)

- IBM’s Fairness 360 toolkit

Reliable Systems

Software engineering practices for a TEAM

5) Explainable user interfaces

- Retrospective explanations (local & global)

New Goal: **Prevent** confusion and surprise

- Prospective user interfaces
- Interactive, visual, exploratory

Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

Mortgage amount requested	375000
Household monthly income	7000
Liquid assets	48000

Submit

Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

Mortgage amount requested	375000
Household monthly income	7000
Liquid assets	48000

Submit

Enter amounts to request mortgage:

Mortgage amount requested	375000
Household monthly income	7000
Liquid assets	48000

Submit

We're sorry, your mortgage loan was not approved. You might be approved if you reduce the Mortgage amount requested, increase your Household monthly income, or increase your Liquid assets.

Done

Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

Mortgage amount requested	375000
Household monthly income	7000
Liquid assets	48000

Submit

Enter amounts to request mortgage:

Mortgage amount requested	375000
Household monthly income	7000
Liquid assets	48000

Submit

We're sorry, your mortgage loan was not approved. You might be approved if you reduce the Mortgage amount requested, increase your Household monthly income, or increase your Liquid assets.

Done

Prospective User Interface

Adjust sliders to report your situation:

Mortgage amount requested

375000

Household monthly income

7000

Liquid assets

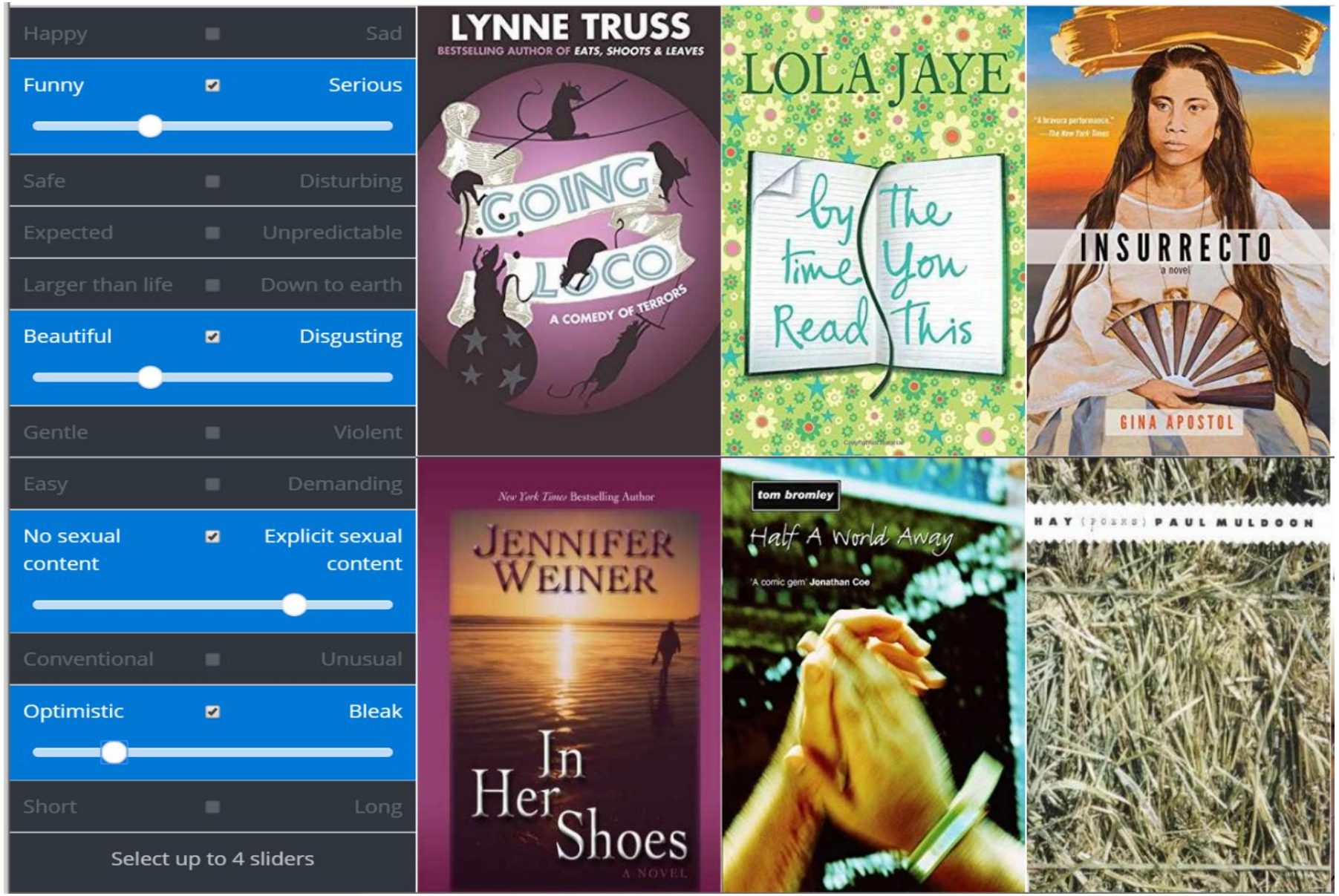
48000

Score needed for approval

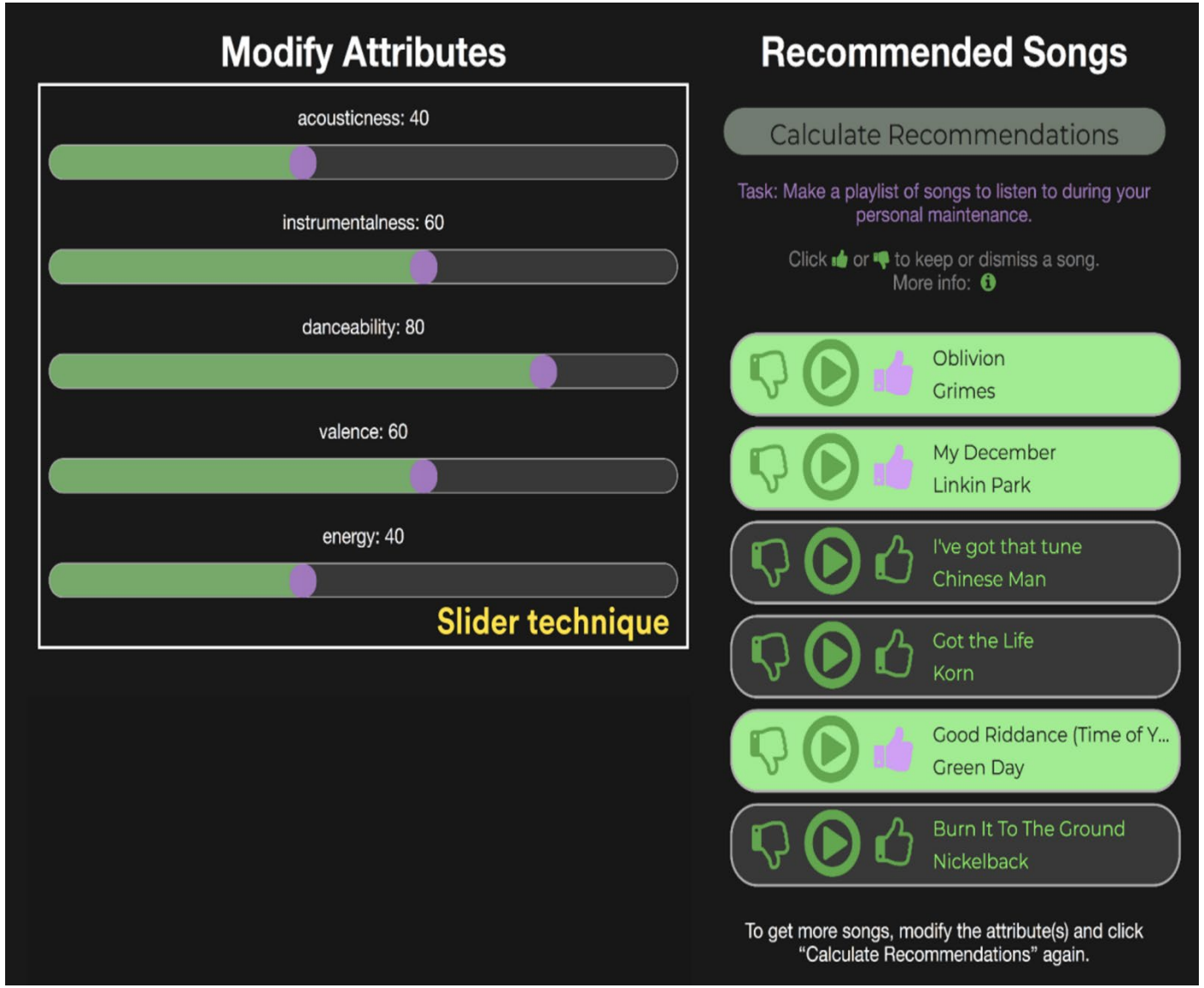
Your score

Done

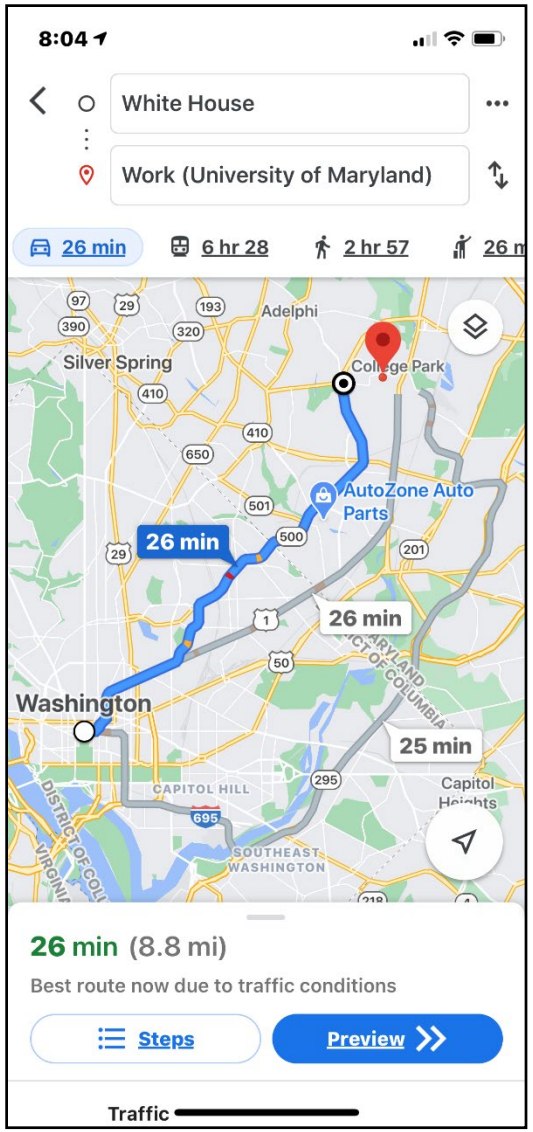
Recommenders: Whichbook.net



Recommender Control Panels

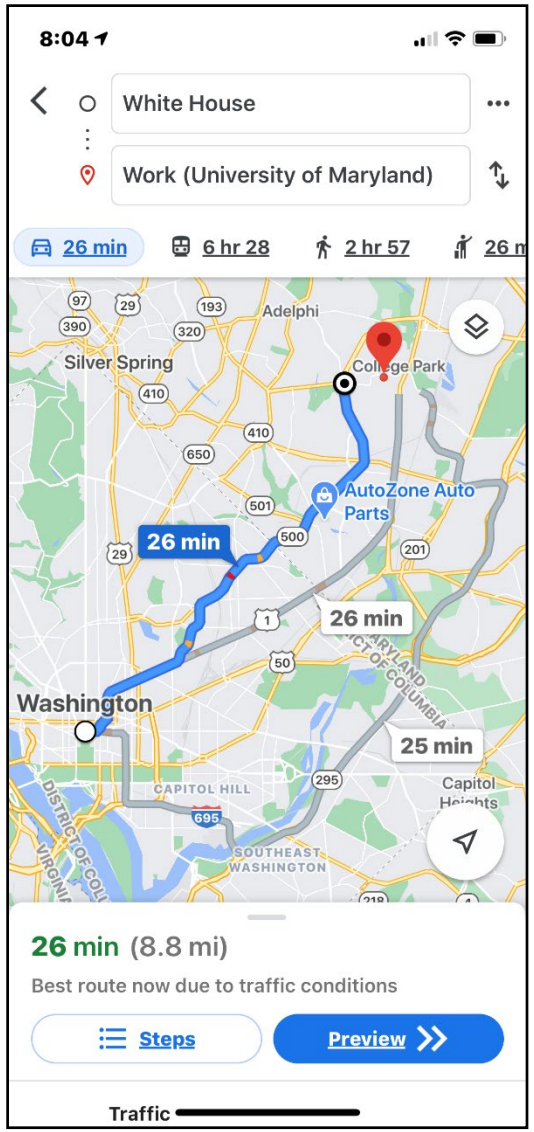


Visual Previews



- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand

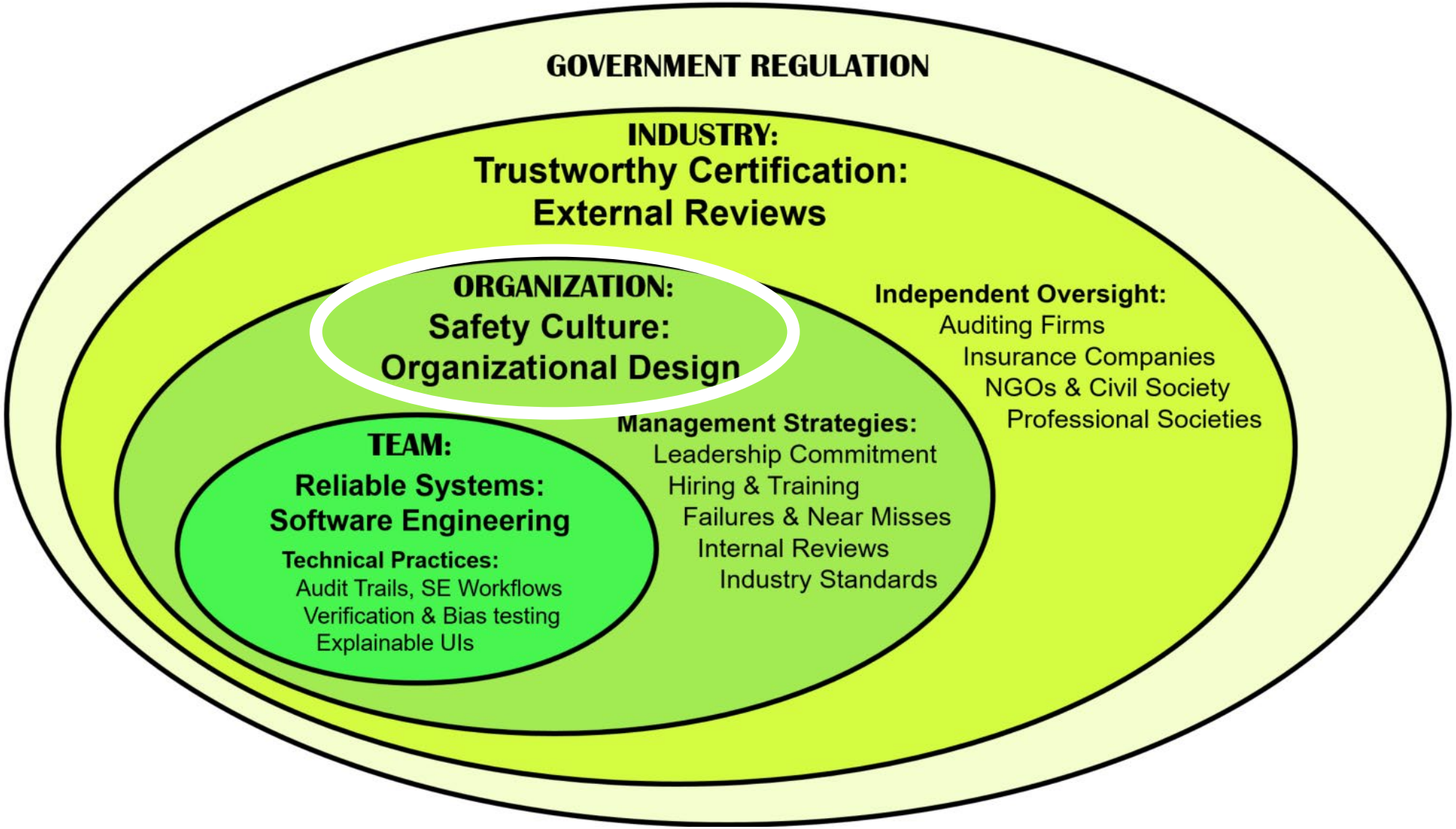
Visual Previews



- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand

- Preview first, select & initiate, then show execution
- Preview first, select & initiate, then show execution
- Preview first, select & initiate, then show execution
- Preview first, select & initiate, then show execution

Governance Structures for Human-Centered AI



ACM TIIS (Oct 2020) **https://dl.acm.org/doi/10.1145/3419764**

ORGANIZATION

Safety culture through business management strategies

6) Leadership commitment to safety
7) Hiring and training oriented to safety
8) Extensive reporting of failures and near misses
9) Internal review boards for problems and future plans
10) Alignment with industry standard practices

Safety Culture

Business management strategies for an ORGANIZATION

6) Leadership commitment to safety

- Normal Accident Theory
- High Reliability Organizations
- Resilience Engineering
- Safety Culture: Repeated public statements,
  Vision statements, Budget, Openness about failures,
  Annual reports on safety, Competitive advantage

Safety Culture

Business management strategies for an ORGANIZATION

7) Hiring and training oriented to safety

- Job notices consistently emphasize safety
- Interviews focus on safety
- Training for safety
- Retraining & simulated disasters

Safety Culture

Business management strategies for an ORGANIZATION

8) Extensive reporting of failures and near misses

- Public & internal reporting
- Data analysis & reviews
- US FAA, FDA Adverse event reporting
- Bug & bias bounties
- AI Incident Database (https://incidentdatabase.ai/)
- Tesladeaths.com

Safety Culture

Business management strategies for an ORGANIZATION

9) Internal review boards for problems & future plans

- Regular reviews & public reporting
- Oversight group gains experience
- Audit committees

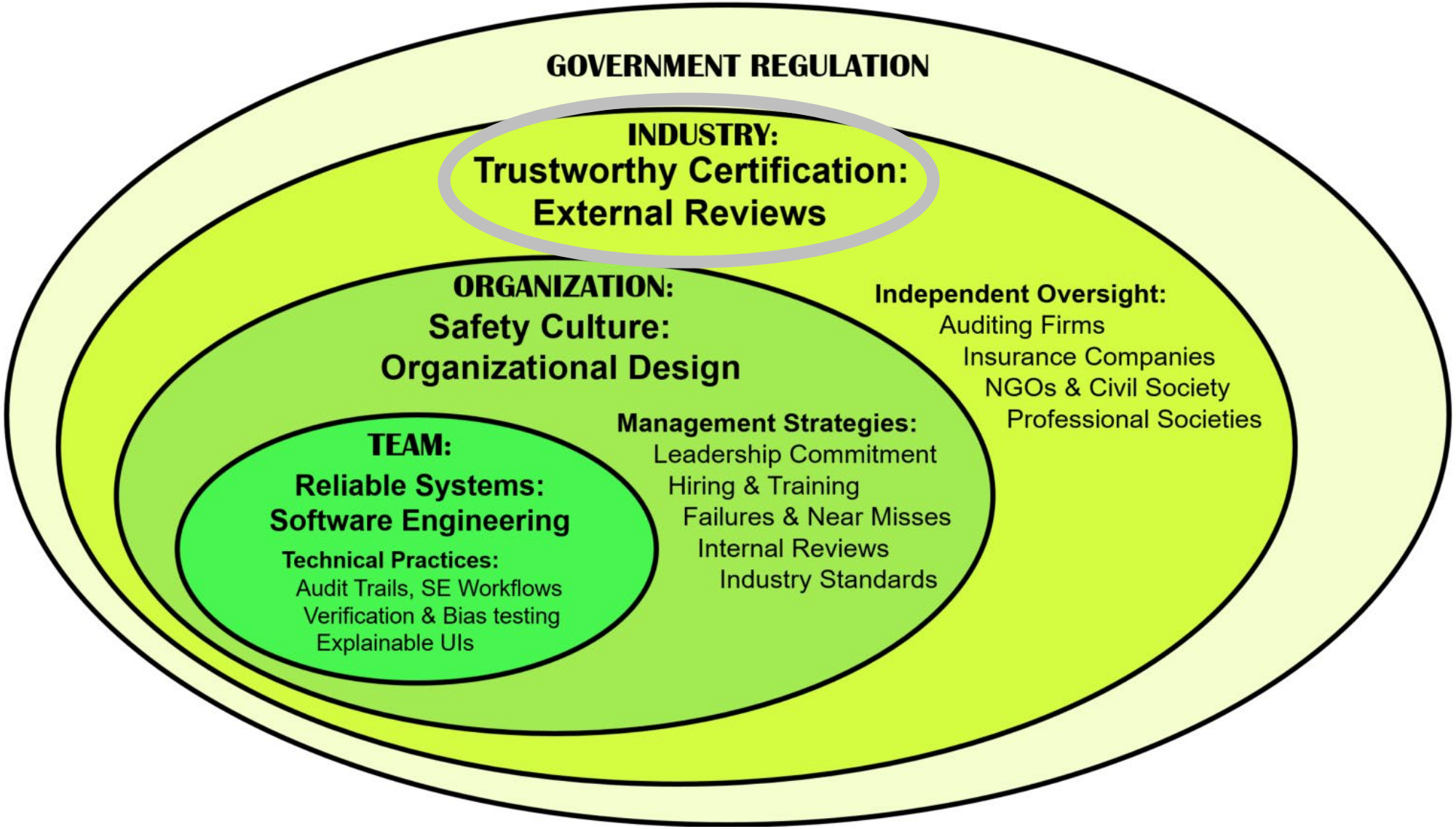
Safety Culture

Business management strategies for an ORGANIZATION

10) Alignment with industry standard practices

- ISO Technical Committee on Robotics
- Robotics Industry Association
- Underwriters Laboratory, Consumer Reports
- IEEE P7000 Series: Ethics, Wellbeing, Transparency
- Capability Maturity Models

Governance Structures for Human-Centered AI



ACM TIIS (Oct 2020) **https://dl.acm.org/doi/10.1145/3419764**

INDUSTRY

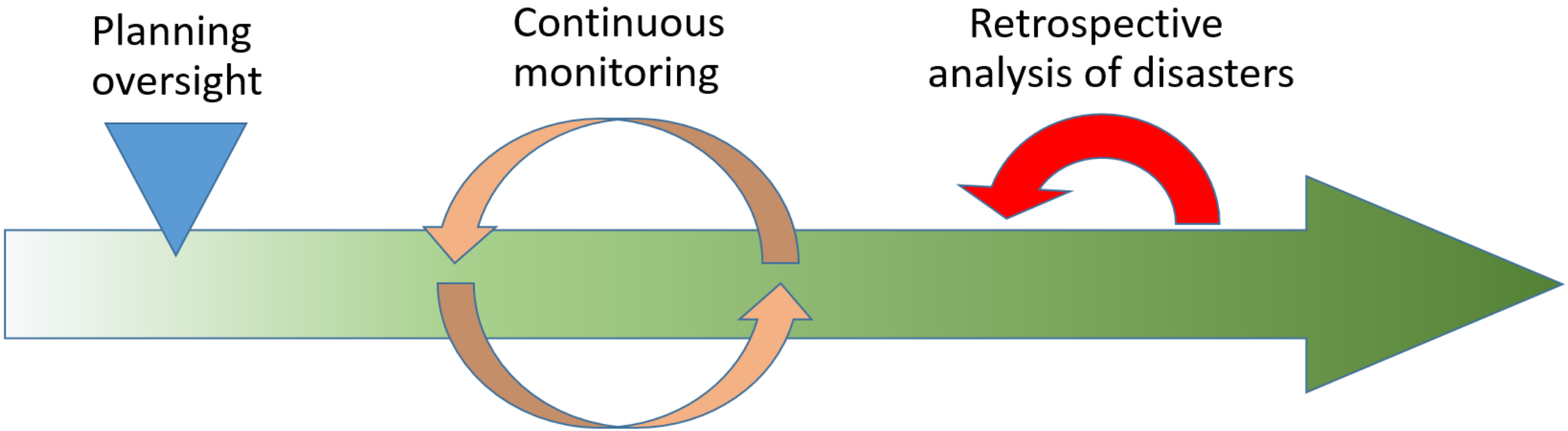
Trustworthy certification by independent oversight

11) Accounting firms conduct external audits
12) Insurance companies compensate for failures
13) Non-governmental and civil society organizations
14) Professional organizations and research institutes

15) Government interventions and regulation

Trustworthy certification

Independent oversight for an INDUSTRY



- Degree of Independence, subpoena power
- Powers to enforce recommendations

PNAS Opinion: (November 29. 2016) http://www.pnas.org/content/113/48/13538.full
To mitigate the dangers of faulty, biased, or malicious algorithms requires independent oversight

Planning oversight

Continuous monitoring

Retrospective analysis

NTSB Tweet: 6:25 PM - 7 Jul 2013

Trustworthy Systems

Certification by independent oversight for an INDUSTRY

11) Accounting firms conduct external audits

- Accounting firms can now include AI Audits
  Deloitte, Ernst & Young, KPMG, PwC
- Consulting companies may play a role
  Accenture, Boston Consulting, McKinsey & Co
- Experience across companies adds value

Trustworthy Systems

Certification by independent oversight for an INDUSTRY

12) Insurance companies compensate for failures

- Success in construction, healthcare, transportation,…
- Building codes for building code (Carl Landwehr)
- Self-driving car insurance premiums guide safety
- Skeptics don’t trust insurance companies

Trustworthy Systems

Certification by independent oversight for an INDUSTRY

13) Non-governmental and civil society organizations

- Many NGOs already active
- Algorithmic Justice League success: facial recognition

Trustworthy Systems

Certification by independent oversight for an INDUSTRY

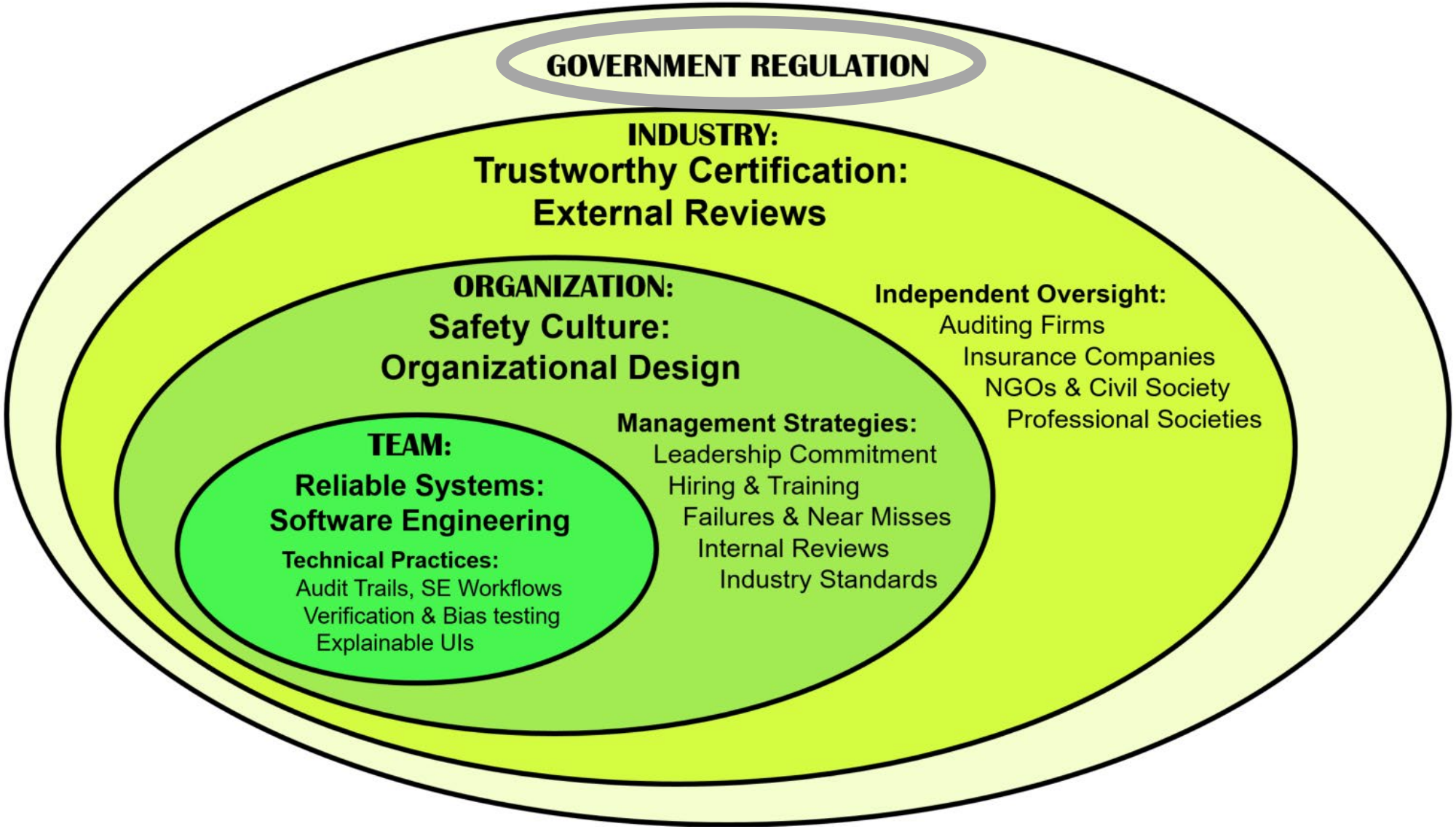
14) Professional organizations and research institutes

- IEEE P7000 series of standards
- IEEE Ethics of Autonomous & Intelligent Systems
- Montreal AI Ethics Institute
- OECD AI Policy Observatory
- ACM Technology Policy Committee
- Partnership on AI

University Research Groups

- Brown Univ	Humanity Centered Robotics Initiative
- Columbia Univ	Data Science Institute
- Harvard Univ	Berkman Klein Center for Internet and Society
- Johns Hopkins Univ	Institute for Assured Autonomy
- Monash Univ, Australia	Human Centered AI
- New York Univ	Center for Responsible AI
- Northwestern Univ	Center for Human-Computer Interaction + Design
- Stanford Univ	Human-centered AI (HAI) Institute
- Univ of British Columbia, Canada	Human-AI Interaction
- Univ of California-Berkeley	Center for Human-Compatible AI
- Univ of Cambridge, UK	Leverhulme Centre for the Future of Intelligence
- Univ of Canberra, Australia	Human Centred Technology Research Centre
- Univ of Chicago	Chicago Human+AI Lab (CHAI)
- Univ of Oxford, UK	Internet Institute, Future of Humanity Institute
- Univ of Toronto, Canada	Ethics of AI Lab
- Utrecht Univ, Netherlands	Human-centered AI

Governance Structures for Human-Centered AI



ACM TIIS (Oct 2020) **https://dl.acm.org/doi/10.1145/3419764**

Government Regulation

Certification by independent oversight for an INDUSTRY

15) Government interventions and regulation

- Regulations: Good or Bad?
- U.S. FAA, FDA, FTC, NIST
- EU General Data Protection Regulation
- OECD Principles of HCAI
- UN AI for Good Global Summit

Government Regulation

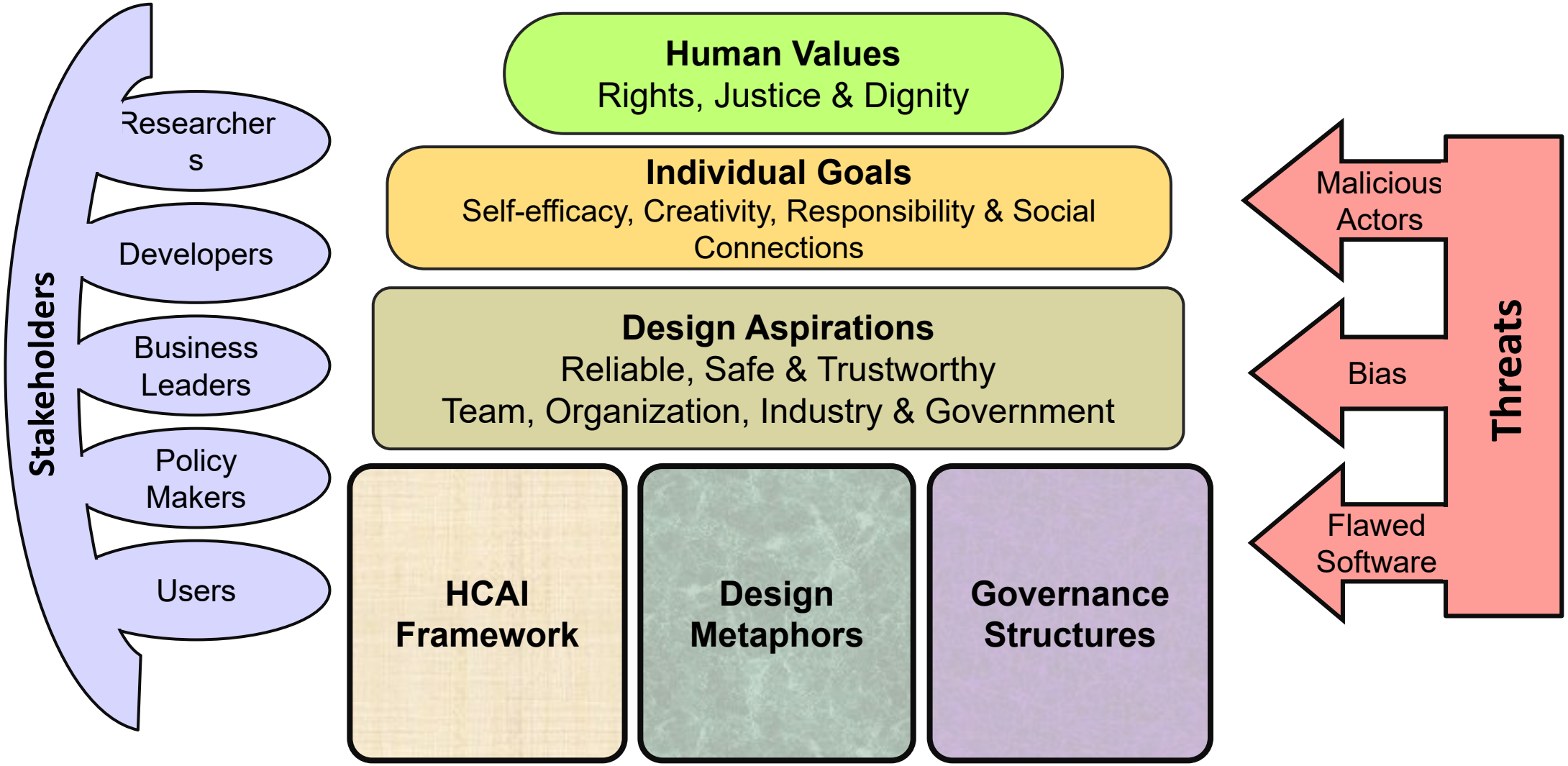
Certification by independent oversight for an INDUSTRY

15) Government interventions and regulation

- Regulations: Good or Bad?
- U.S. FAA, FDA, FTC, NIST
- EU General Data Protection Regulation
- OECD Principles of HCAI
- UN AI for Good Global Summit

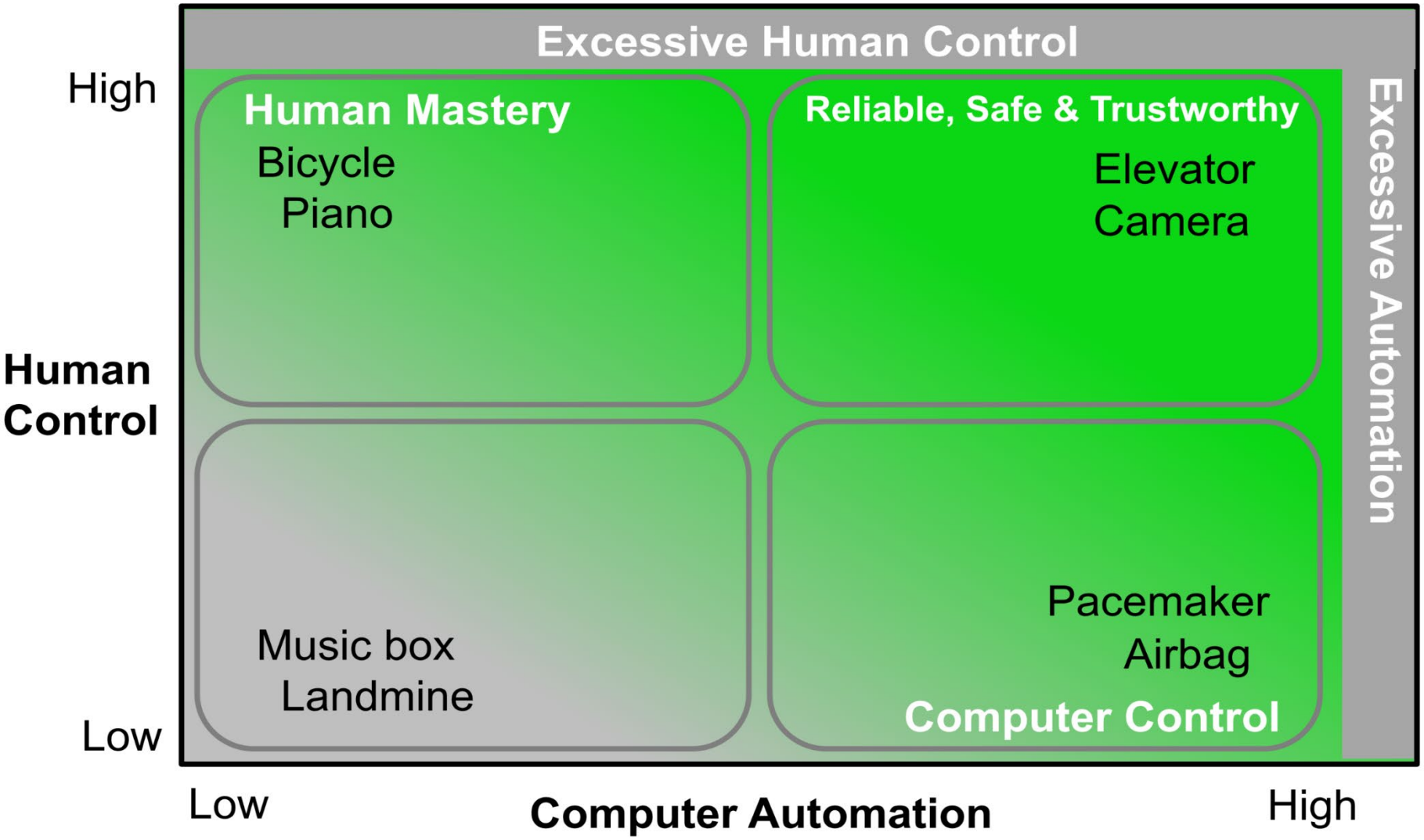
Summary

Human-Centered AI



Oxford University Press (Early 2022) **https://hcil.umd.edu/human-centered-ai/**

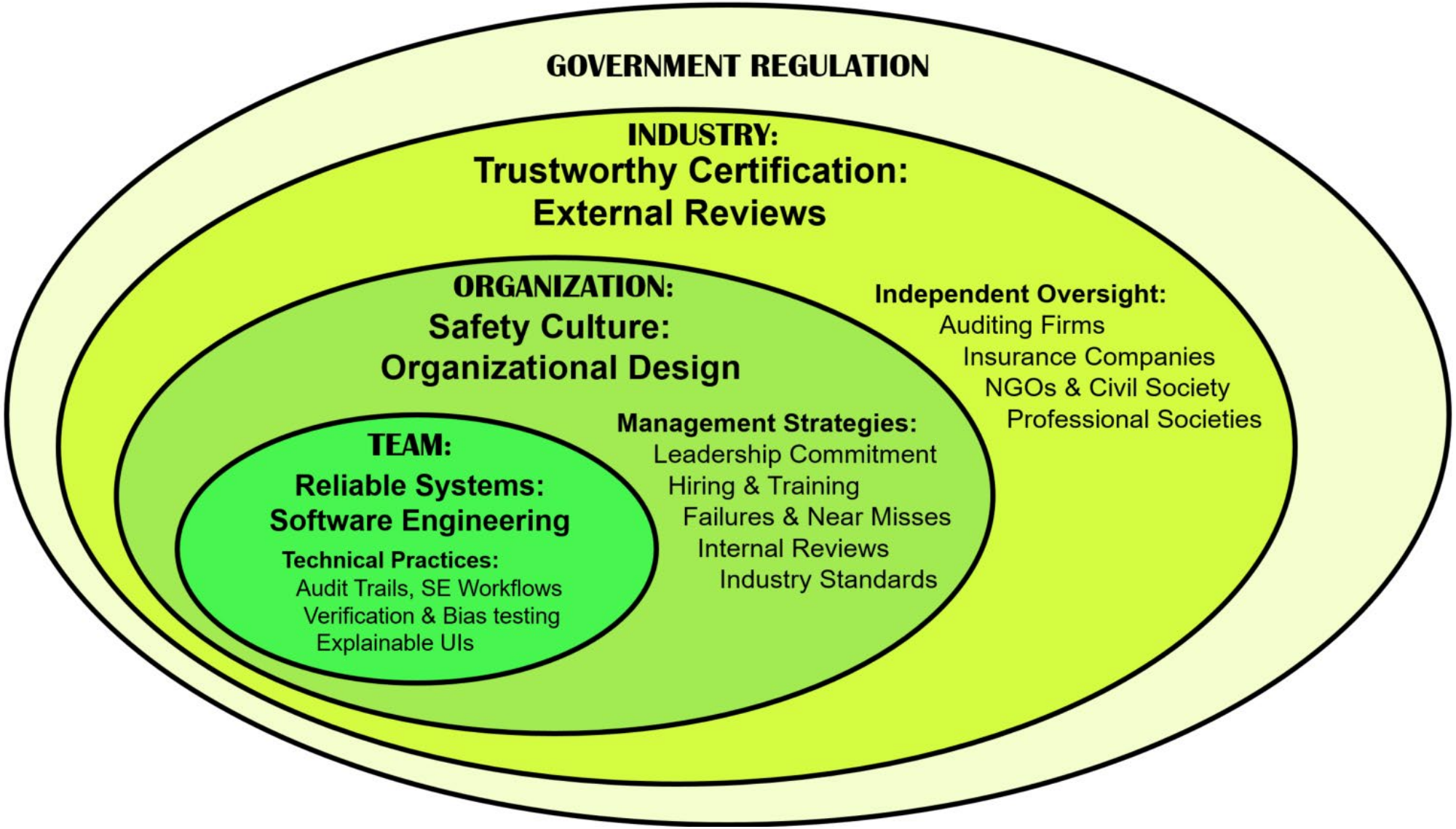
HCAI Framework



Design Metaphors

Science Goal	Combined Designs	Innovation Goal
Intelligent Agents Thinking Machine, Cognitive Actor, Artificial Intelligence, Knowledgeable		**Supertools** Extend Abilities, Empower Users, Enhance Human Performance
Teammates Co-active Collaborator, Colleague, Helpful Partner, Smart Co-worker		**Tele-operated Devices** Steerable Instrument, Powerful Prosthetic, Boost Human Perceptual & Motor Skills
Assured Autonomy Independent, Self-directed, Goal-setting, Self-monitored		**Supervised Autonomy** Human Control & Oversight, Situation Awareness, Predictable Actions
Social Robots Anthropomorphic, Humanoid, Android, Bionic, Bio-inspired		**Active Appliances** Consumer-oriented, Wide Use, Low Cost Comprehensible Control Panels

Governance Structures for Human-Centered AI



ACM TIIS (Oct 2020) **https://dl.acm.org/doi/10.1145/3419764**

Human-Centered Artificial Intelligence: Reliable, safe & trustworthy, *International Journal of Human-Computer Interaction 36,* 6 (March 2020). https://doi.org/10.1080/10447318.2020.1741118
Design lessons from AI's two grand goals: Human emulation and useful applications, *IEEE Transactions on Technology & Society 1,* 2 (June 2020). https://ieeexplore.ieee.org/document/9088114

Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems, *ACM Trans. on Interactive Intelligent Systems 10,* 4 (Oct 2020). https://dl.acm.org/doi/10.1145/3419764

Human-Centered Artificial Intelligence: Three fresh ideas, *AIS Trans. on Human-Computer Interaction 12,* 3 (Oct 2020). https://aisel.aisnet.org/thci/vol12/iss3/1/

Human-Centered AI. *NAS ISSUES 37, 2* (Winter 2021). https://issues.org/human-centered-ai/

Summary & resources: https://hcil.umd.edu/human-centered-ai/

Human-Centered AI: Google Group

https://groups.google.com/g/human-centered-ai

Annual Symposium: FREE Virtual

May 27, 2021, Thursday

https://hcil.umd.edu/2021-symposium/