
CSET CENTER for SECURITY and EMERGING TECHNOLOGY
GEORGETOWN UNIVERSITY
Providing decision-makers with data-driven analysis on the security implications of emerging technologies
13 July 2023 | NIST Advisory Board
Dewey Murdick | Executive Director
cset@georgetown.edu | cset.georgetown.edu

Policy Analysis via an Interdisciplinary, Cross-Matrixed Team

Over 60 dedicated staff examining emerging technology, public policy, and security issues

- Analysis
- Data Science
- Translation
- External Affairs
- Operations

Convergence
between AI and other sectors from medical imaging to manufacturing

Map of Science: medicine, 3-yr growth forecast

Emerging Technology Observatory

MAP VIEW | LIST VIEW | SUMMARY VIEW

Applied filters
Cluster size: 100 to 13737
AI percentage: 10 to 100
% industry-affiliated: 1 to 67
% articles with patent citation (beta): 2 to 35
Subjects: General: medicine

CLEAR FILTERS | APPLY FILTERS

- Vital signs
- Countries and languages
- Hot topics

AI percentage: 10 – 100

Robotics percentage (English only): 0 – 89

Computer vision percentage (English only): 0 – 100

Natural language processing percentage (English only): 0 – 93

- Patents and industry (beta)

Top clusters

ADD/REMOVE COLUMNS

Cluster ID	Most common subject category	CSET phrases	AI percentage	Robotics percentage (English only)	Computer vision percentage (English only)	Natural language processing percentage (English only)	% articles with patent citation (beta)	% industry-affiliated
29289	biology	Blood-brain barrier, drug discovery	12.33	0.00	0.46	0.00	7.80	18.30
34269	biology	Drug-induced liver injury	10.37	0.00	0.00	0.00	6.07	14.04
34928	computer science	X-rays during surgery	58.33	1.50	52.50	0.00	10.29	9.67
24833	physics	Respiratory motion correct.	12.85	0.00	13.36	0.00	7.70	8.75
74168	medicine	Interval walking training	11.27	1.45	7.25	0.00	4.57	7.08
35774	medicine	Parkinson disease testing	25.84	2.30	2.30	0.00	2.50	5.83
8647	social science	Diagnostic pathology	13.90	0.00	14.90	0.00	2.39	5.61
55101	biology	Tissue imaging (tumors)	10.08	0.00	7.29	0.00	5.17	5.54
57294	computer science	Skin cancer detection	21.56	0.00	27.27	0.00	6.22	5.36
47166	computer science	Cough detection	53.25	0.20	2.76	2.17	3.74	5.21

1–10 of 159

Map of Science: business, 3-yr growth forecast

Emerging Technology Observatory

MAP VIEW | LIST VIEW | SUMMARY VIEW

Applied filters

Cluster size: 100 to 13737
AI percentage: 10 to 100
% industry-affiliated: 1 to 67
% articles with patent citation (beta): 2 to 35
Subjects: General: business

CLEAR FILTERS | APPLY FILTERS

Vital signs

Or And

Subjects: GENERAL: BUSINESS

Cluster size: 100 – 13737

Growth rating: 0 – 100

Citation rating: 0 – 91

Average paper age: 0 – 112

Extreme growth predicted

Countries and languages

Hot topics

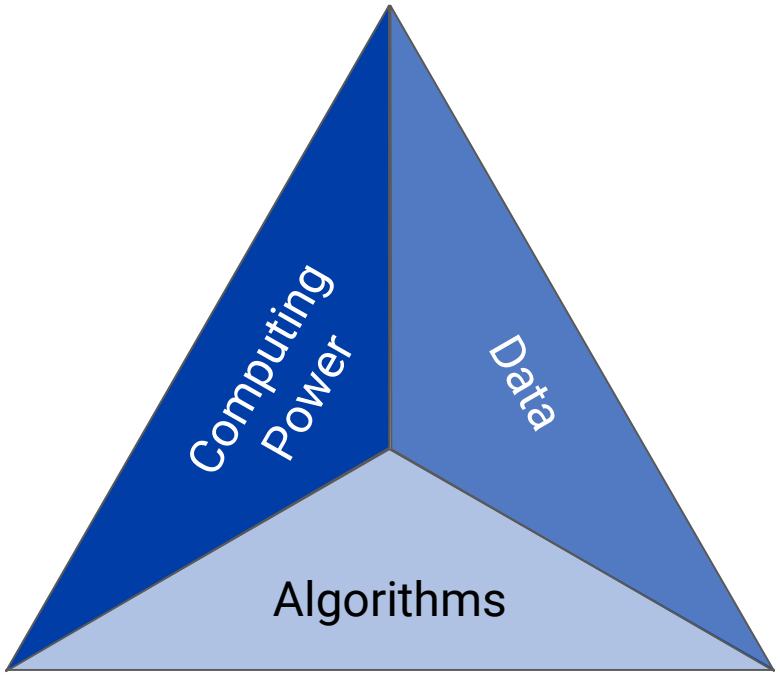
Patents and industry (beta)

Top clusters

ADD/REMOVE COLUMNS

Cluster ID	Most common subject category	CSET phrases	AI percentage	Robotics percentage (English only)	Computer vision percentage (English only)	Natural language processing percentage (English only)	% articles with patent citation (beta)	% industry-affiliated
39447	computer science	Custom network segment...	14.30	0.26	0.00	0.00	2.55	20.62
24979	computer science	Virtual network security	10.17	0.00	0.00	0.00	2.70	19.88
58210	social science	Controlled experiments for software development	15.63	0.00	0.35	0.69	2.39	19.06
64854	engineering	Simulated engr'd-system testing & preparation	13.56	10.43	0.00	0.00	2.30	14.32
47366	social science	Contextual ad platform	43.80	0.00	7.63	9.16	4.63	14.16
45690	computer science	Cyber-physical industrial automation & manufact. sys	22.32	14.21	0.55	0.00	3.45	12.29
47583	computer science	Online DB query cost prediction	18.98	0.00	0.00	0.00	5.40	9.48
5760	social science	PII management	18.13	0.00	0.00	11.38	3.33	9.35
88747	computer science	Teleoperated vehicles	42.70	19.66	8.43	0.00	2.13	9.30
80171	computer science	Industrial automation control system design	10.00	1.79	0.45	0.00	3.34	9.11

The AI Triad — *Powered by Human Talent*

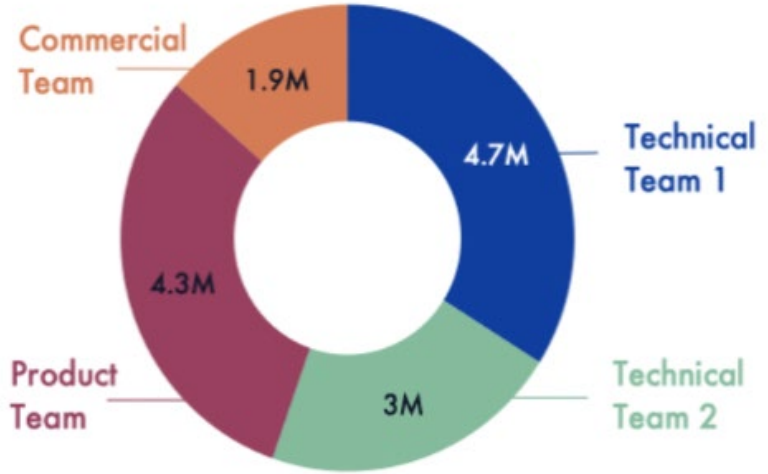


The AI Triad

"Machine learning systems use computing power to execute algorithms that learn from data."

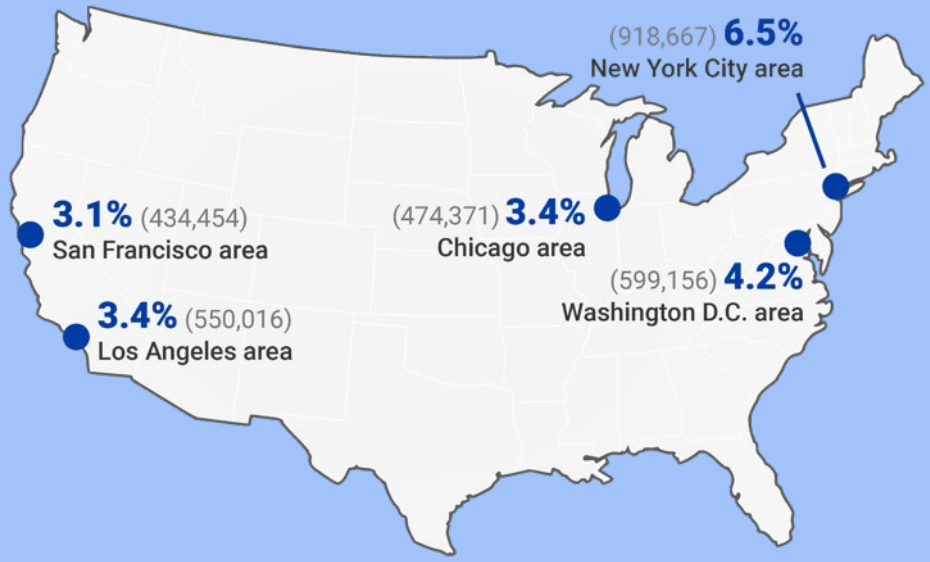
– Ben Buchanan, "The AI Triad and What It Means for National Security Strategy" (2020)

In 2019, the U.S. AI workforce consisted of 14 million workers, or about 9% of total U.S. employed.



Source Data: Diana Gehlhaus and Santiago Mutis, "The U.S. AI Workforce: Understanding the Supply of AI Talent" (CSET, January 2021).

AI Workforce is Geographically Concentrated



Source Data: Diana Gehlhaus and Ilya Rahkovsky, "U.S. AI Workforce: Labor Market Dynamics" (CSET, April 2021)

- Talent shortages vary by occupation
- Diversity can be greatly improved (*Tech 1 & 2 → 6-7% black, 0.3% native american, 4-5% other; 73% male*)
- AI/AI-related certifications have limited value currently → community and technical colleges

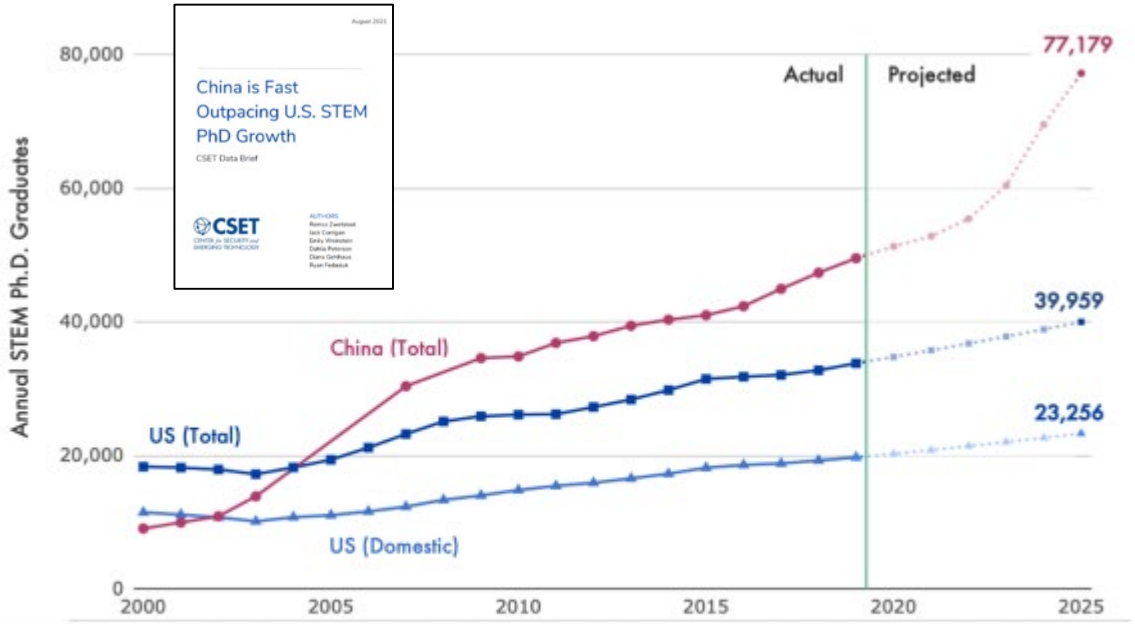
Peer innovators, competition, alliances, and a long-term wrestling match

U.S. AI Education Risks Lagging China

China exceeds the US in both scale and scope of AI education with curricula for all levels

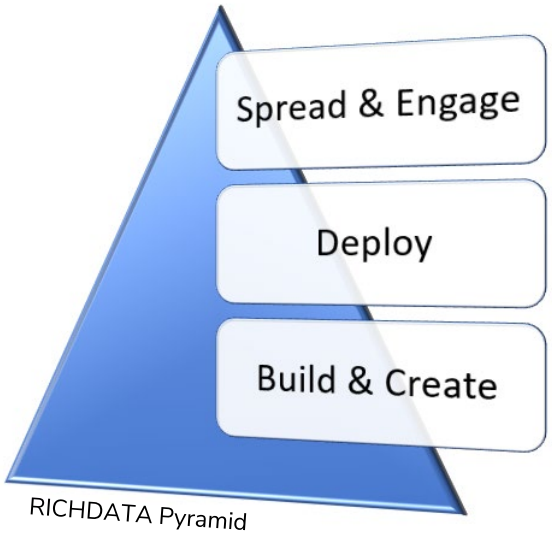
China is Fast Outpacing U.S. STEM PhD Growth →

China's STEM doctorates production is projected to greatly expand



Source: National Center for Education Statistics' Integrated Postsecondary Education Data System (IPEDS) for U.S. data, Ministry of Education for Chinese data (see Appendix A).

AI-enabled disinformation campaigns



RICHDATA Pyramid

- Precise targeting of intended audiences that identify societal fissures, determining sentiment, etc.
- Development of more realistic (harder to trace) personas & pages with ML-generated images, text generation, etc.

- Accelerated content generation at scale with large language models (see CSET's work with GPT-3)
- Boost trolling and online chaos with more capable chatbots

Source: Katerina Sedova, et al., "AI and the Future of Disinformation Campaigns: Part 2: A Threat Model" (Center for Security and Emerging Technology, December 2021).

Four considerations and a few actions to discuss

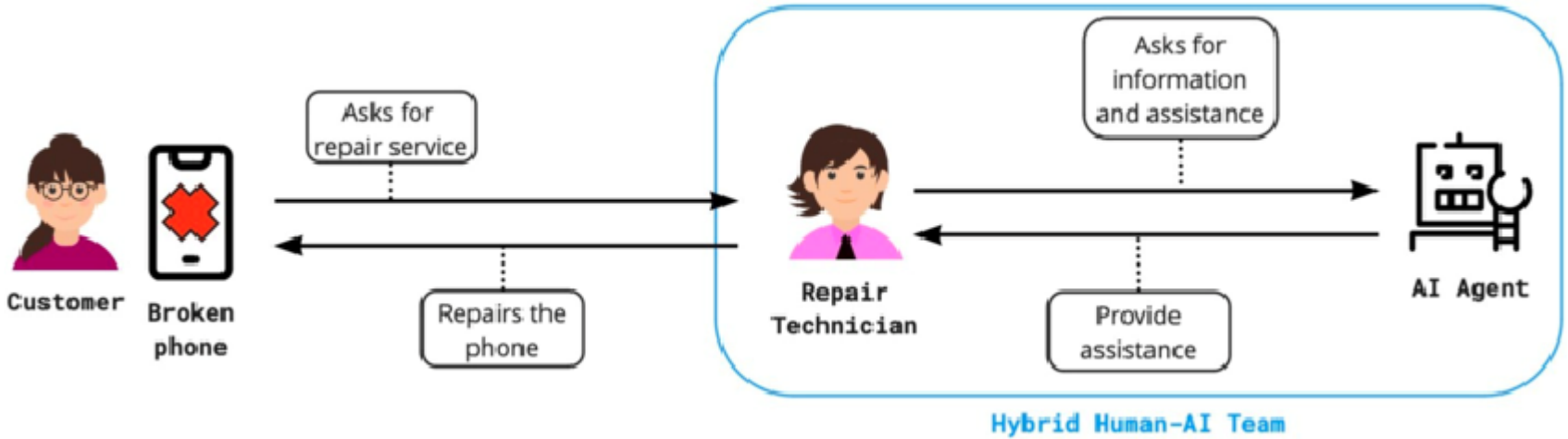
Importance of human-machine teaming

Human-AI Teaming

Can the AI be consistently **relied upon**? Is this confidence **calibrated**?
Does AI **perform consistently** across similar tasks in varied situations?
Does AI provide **usable explanations** and uncertainty estimates?
Has the AI system **enhanced productivity** or work quality?

AI-Assisted Human Learning

How effectively is AI **facilitating learning**?
Are **outcomes improved** or **faster** due to AI assistance?



E. Van Zoelen et al. *Developing Team Design Patterns for Hybrid Intelligence Systems. HHAI 2023: Augmenting Human Intellect, P. Lukowicz et al. (Eds.)* [CC BY-NC 4.0]

Data governance practice that

- *Contributes to justice and genuine respect for people in the way that makes **people and groups visible**, **represented**, and **empowered as beneficiaries in the collection and use of data** for the development of AI systems*
- *Widens the lens beyond a narrow view focused on **compliance and individualised privacy** and accounts for **collective identities** and **community-level decision making***
- *Seeks out and engages with a full set of **impacted stakeholders** in the design, development, and deployment of AI systems and **agilely adapts** when new stakeholder perspectives are discovered*

GPAI

Data Governance Working Group

"Data Justice Policy Brief: Putting data justice into practice" (GPAI, November 2022). (more)

Applying cybersecurity lessons to **Securing AI**

Defend AI systems against malicious actors seeking to subvert a system or steal the underlying model or its associated data.

Secure AI ≠ Safe (Trustworthy) AI, but reduces the chance AI systems will be deliberately misused for harm.

Relevant Lessons from Cybersecurity

- Begin now to protect data sets, trained models, and open-source libraries.
- Confidentiality and integrity of models and data will become increasingly important over time

- **Threats**
  - Recognize sensitivity of **model parameter** weights which can be accessed through open source sharing, stolen, or inferred
  - Understand potential of model stealing via **distillation attacks**
  - Protect against **adversarial attacks**
  - Prep for **downstream attacks** arising as developers build integrated plug-ins
- **Security measures**
  - **Limit access** to sensitive models through APIs and data confidentiality measures
  - Apply traditional network security measures (e.g., **firewalls**, **access controls**)

John Bansemer and Andrew Lohn. *Securing AI Makes for Safer AI* (CSET, 6 July 2023).

Balancing goals-in-tension across society

Innovate

Tech Protect

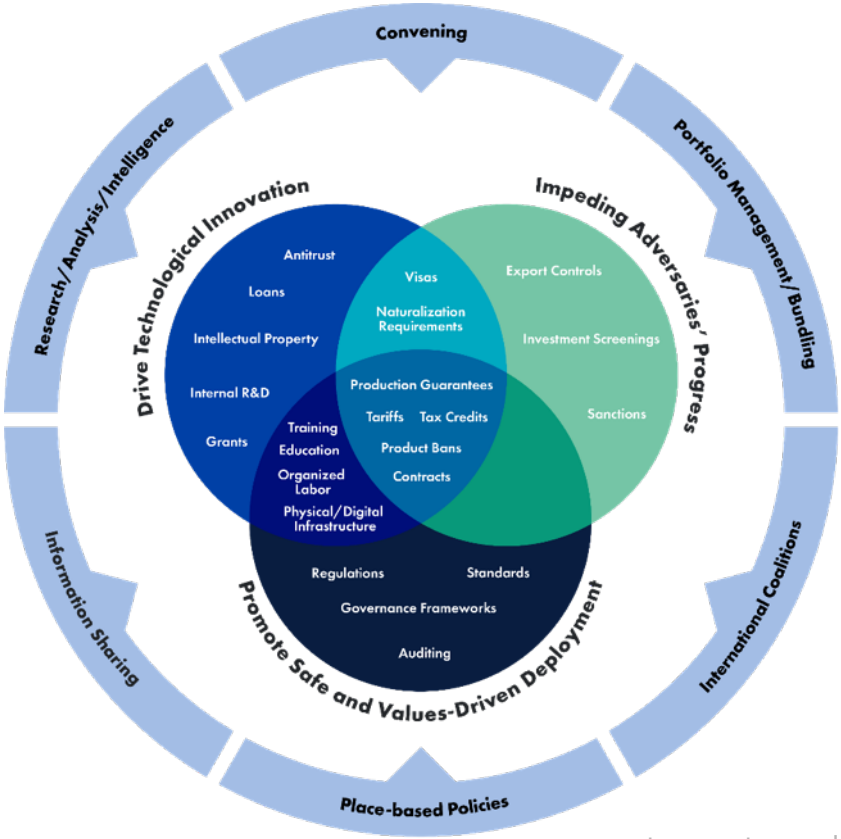
Safety

In the face of a technology-enabled soft/hard power and a clash of values, how do we:

- Enable the public & private tech innovation ecosystem to **“run faster”** and help solve pressing problems and create jobs?
- **Slow** unwanted technology transfer and protect key enabling technology?
- Ensure AI systems (etc.) are **safe** and operate in harmony with our best values?

Actionable levers of power for tech-policy

Identify relevant “levers” that are operated within a decentralized, international & locally-informed system to achieve meaningful goals (almost always in tension)



Jack Corrigan, Melissa Flagg and Dewey Murdick, *The Policy Playbook* (CSET, June 2023).

So, what do we do? [Short Term]

- Talent
  - Basic AI Literacy & K-12 Education
  - Community College & Certification
  - STEM Education
  - Diversity is a National Security Asset
  - High Skilled Immigration
- Information Gathering *(and engage the democratic process)*
  - Track AI harms via *incident reporting*
  - Learn how to request key model and training data used in important applications
  - Encourage the development of the third party auditing and red teaming ecosystem
- Development
  - Measure and metric development and implementation, including human-machine teaming
  - Improve the quality of shared resources, such as open source training and pretrained models that form the backbone of many of today's AI systems.

So, what do we do? [Longer Term]

- We need to keep a close eye on our policies, each implementation needs an effective monitoring system.
- Rethink software deployment model ("run fast break things" → clinical trials)
- If we license or register AI software (a common proposal) or make sure it's used safely, we will need to update the authorities for existing agencies — and perhaps create a new one that could
  - Check how AI is being used in and overseen by existing agencies;
  - Be the first to deal with problems, directing those in need to the right solutions;
  - Fill gaps that existing sector-specific agencies don't cover; and
  - Work closely with NIST and industry to identify and update evaluation standards.

Optional

Bonus: OECD & NIST-Relevant Info

Risk Management Framework and Friends

Expert Group on AI Risk & Accountability (OECD WP & ONE AI)

- Many AI-relevant risk management standards and frameworks exist, they need greater consistency and interoperability; working to align key versions
- Mapping of relevant actors, issues, and terminology for “Responsible Business Conduct in AI”

CSET’s related efforts to create

- A matrix for selecting responsible AI frameworks from an integrated set of 40
- Profile to guide the translation of RMF into management and deployment AI-system practice (important to tie each cat/sub-cat to specific roles in an org)
- AI Incidents and harms taxonomy (see next slide)

Developing a common incident reporting framework & a common harms definition

AI INCIDENT DATABASE

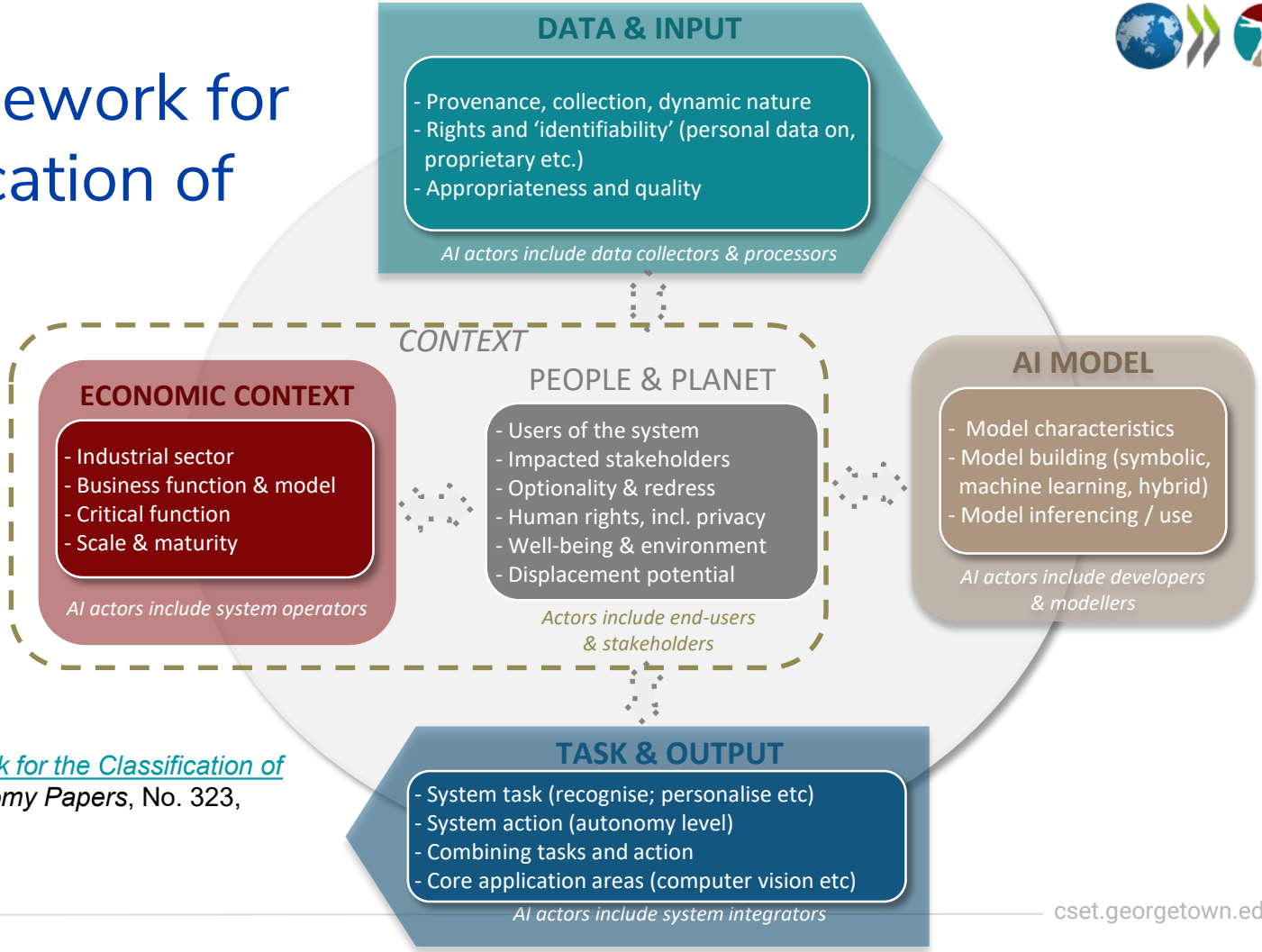
Recommender system reportedly promotes false or misleading claims on the Ukraine war (#185)

TikTok

The newly organized Responsible AI Collaborative designed an initial database (arXiv) and populates the db with found news incidents.

Observed issues with “smart summons” (#178) and “full self-driving” (#145) modes have been reported.

OECD framework for the classification of AI systems



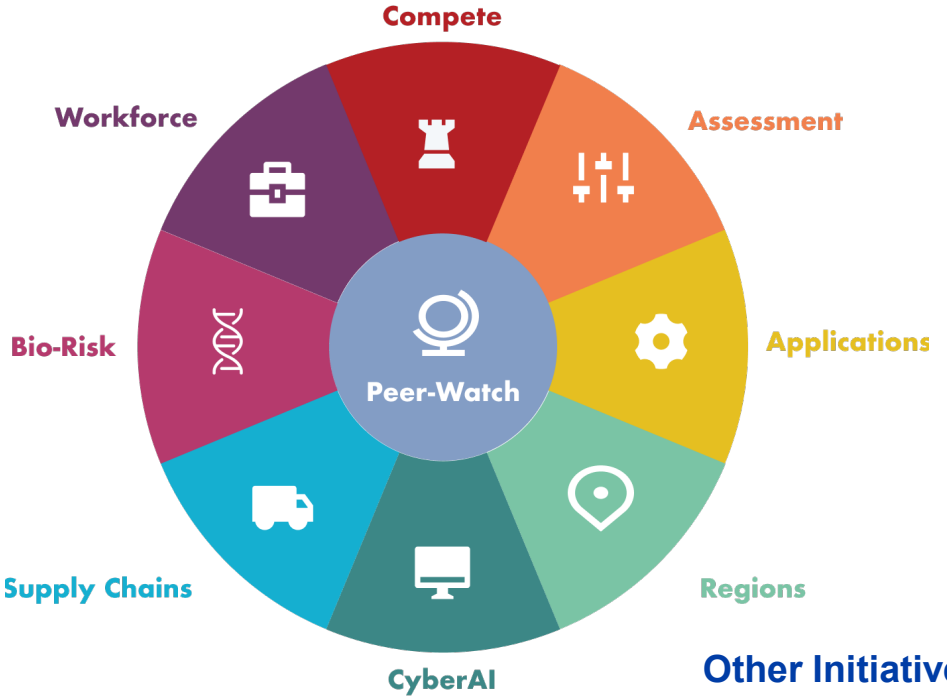
OECD (2022), *OECD Framework for the Classification of AI systems*, *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris.

How can CSET help?

If time

CSET Lines of Research

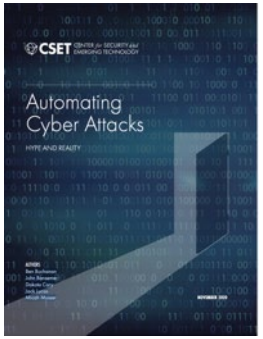
https://cset.georgetown.edu/research-topics

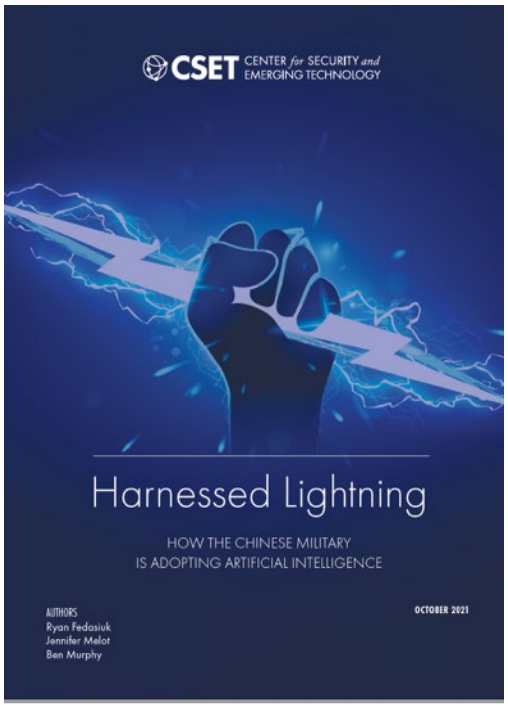


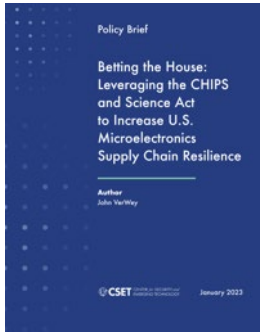
Other Initiatives

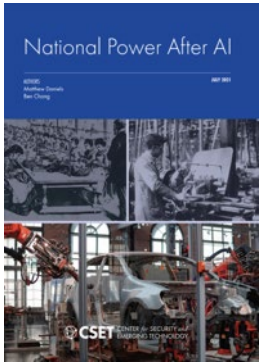
- Emerging Technology Observatory (ETO)
- Foreign-Language Translations
- Foundational Research Grants (FRG)

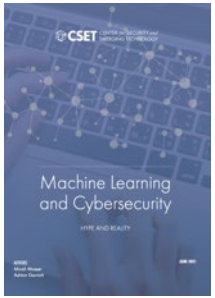
>260 products (since Sept 2019), >400 translations

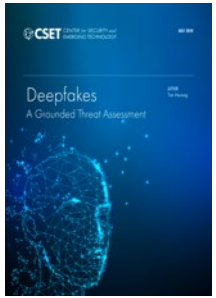








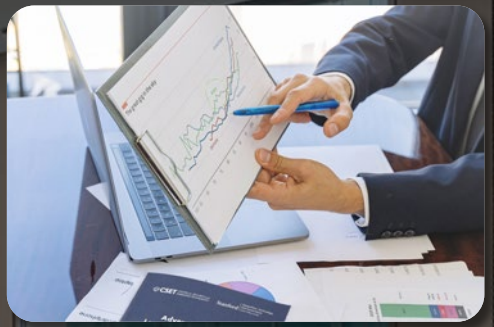




CSET's Unique Value Add

Proactive Research Agenda

Data Investment and Integration



Support Decisions

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

- Research at https://cset.georgetown.edu/publications
- Sign up to receive research the day it's issued, subscribe to our newsletter, and get invited to our events at https://cset.georgetown.edu/sign-up/
- To request briefings, contact Danny Hague (danny.hague@georgetown.edu)
- Share your questions and knowledge gaps

---

cset.georgetown.edu | policy.ai | @csetgeorgetown | cset@georgetown.edu