Safeguarding the Noosphere: The Need for Cognitive Security

CLSAC 2023

William Regli
Professor of Computer Science
College of Computing, Mathematical & Natural Sciences
Affiliate Faculty, Clark School of Engineering
Founding Director, Applied Research Laboratory for Intelligence and Security
The University of Maryland at College Park
regli@umd.edu

UNIVERSITY OF MARYLAND

"On the Internet, nobody knows you're a dog."

The New Yorker on July 5, 1993

ARLIS: The Human-Domain UARC for the Intelligence Community & Security Enterprise

- Sponsored by OUSD(IS) as the trusted agent for RDT&E for 17+ IC & Security agencies.
- ARLIS is the only UARC with Social Science and Human Behavior in the core competencies.

Mission Areas

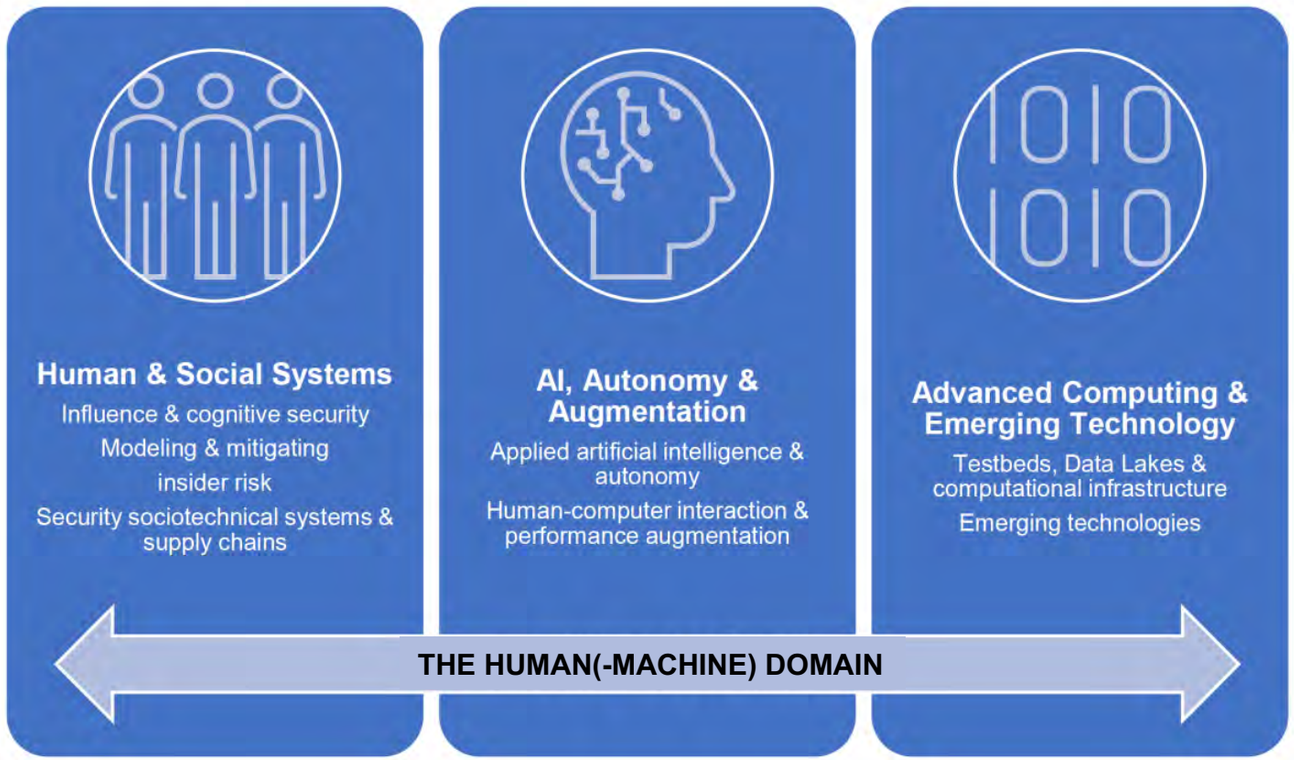
Cognitive Security/ Information & Influence
– IO/IW, influence, cyber, global competition, non-kinetic conflict

Insider Threat, Trust & Risk
– counterintelligence, personnel vetting, countering influence, counter radicalization, credibility

Acquisition & Industrial Security
– technology protection, supply chain assurance, 5G/Future G, analysis

Autonomy, Augmentation and AI
– human-in-the-loop, T&E/IV&V, HCI, collective intelligence

Core Competencies



The Hamming Question

What are the important problems in your field ... and why aren't you working on them?

--- Richard Hamming (1915-1998)
AT&T Bell Laboratories
ACM Turing Award 1968

Framing Hamming Questions

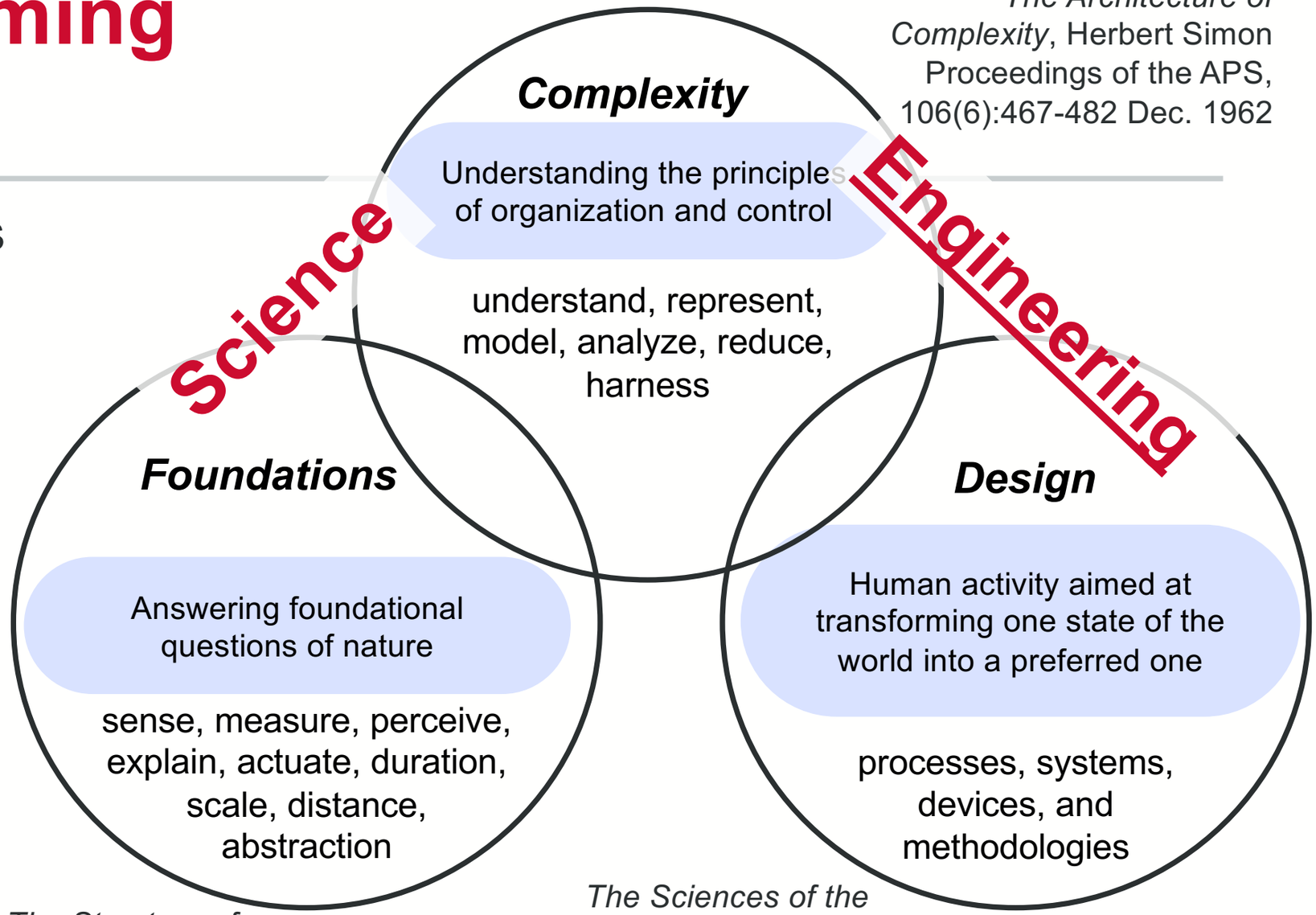
- Quantum sciences
- Health & Medicine
- Material Science
- Energy
- Climate
- Cyber Security
- AI/autonomy

  .

  .

  .

- **Sociotechnical Systems**



The Architecture of Complexity, Herbert Simon Proceedings of the APS, 106(6):467-482 Dec. 1962

The Structure of Scientific Revolutions, Thomas Kuhn, 1962

The Sciences of the Artificial, Herbert Simon, 1969

From Social+Technical to Sociotechnical

"These new problems, and the future of the world depends on many of them, **requires science to make a third great advance,** an advance that must be even greater than the nineteenth-century conquest of problems of simplicity or the twentieth-century victory over problems of disorganized complexity."

- Warren Weaver, "Science and Complexity" (1947)

https://www.pinterest.com/pin/60657926213583248

Sociotechnical Culture

Facts are constructed
Heuristics and meaning
Strategic interactions
Multi-agent systems and networks
Counterfactuals and causality
Embracing uncertainty

https://libguides.usask.ca/anthropology

Technical Culture

Facts *about* social interactions
Etic methods
Find *meaning in* data
Identify patterns
Prediction is goal
Uncertainty is a bug

Social Culture

Facts *from* social interactions
Emic methods
Find *meaningful* data
Identify constructs
Explanation is goal
Uncertainty is a feature

From Adam Russell (University of Maryland/USC ISI)

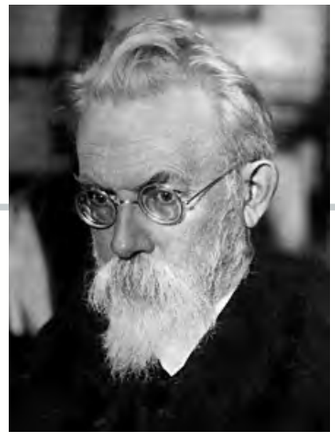
Relating to the themes of CLSAC

The socio-technical requires unique scientific and engineering instruments

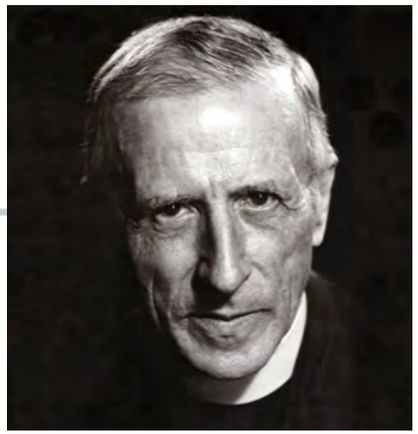
- High-fidelity models and simulations
  - Computational Social Science, agent-based modeling
- Measurement analytics for social systems
  - All media, real-time, multi-language, etc
- Experimental infrastructure
  - Support for repeatable, open science, data sharing, etc

Noosphere: A History

Def'n: the planetary "sphere of reason"; highest stage of biospheric development, that of humankind's rational activities. (summary via Wikipedia)



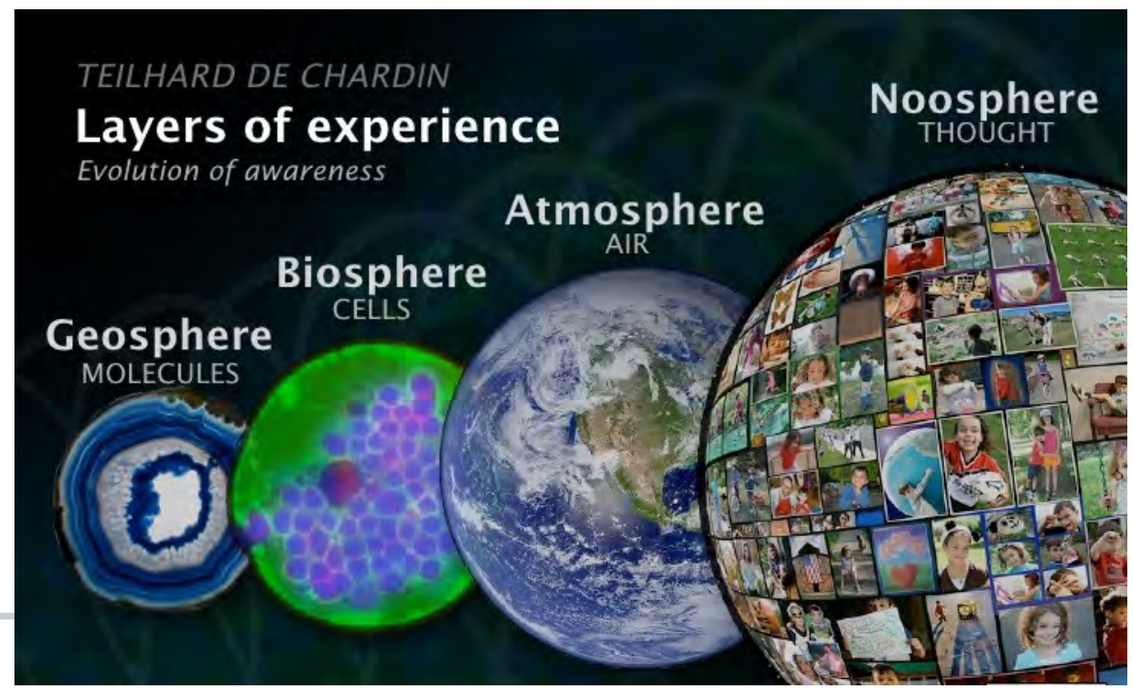
Vladimir Vernadsky
1863 – 1945



Pierre Teilhard de Chardin, S.J.
1881 – 1955

Related thinkers

- Marshall McLuhan
  - Understanding Media: The Extensions of Man (1964)
- Richard Dawkins
  - The Selfish Gene (1976)
- Eric Beinhocker
  - The Origin of Wealth (2007)
- Yuval Noah Harari
  - Sapiens (2015) & Homo Deus (2017)



https://mcluhangalaxy.wordpress.com/2017/06/17/the-mystical-side-of-marshall-mcluhan/noosphere-2/

UNIVERSITY OF MARYLAND

Sociotechnical Systems ➔ The Noosphere

The “physical, cultural and social environments as they relate to the sphere of human activity that exists within an area of interest, conflict, or military operations other than war”

https://smallwarsjournal.com/jrnl/art/conceptualizing-human-domain-management#_edn2

“…the totality of the human sphere of activity or knowledge.”

https://smallwarsjournal.com/jrnl/art/joint-force-2020-and-the-human-domain-time-for-a-new-conceptual-framework

“…the moral, cognitive, social, and physical aspects of human populations in conflict.”

https://smallwarsjournal.com/jrnl/art/joint-force-2020-and-the-human-domain-time-for-a-new-conceptual-framework

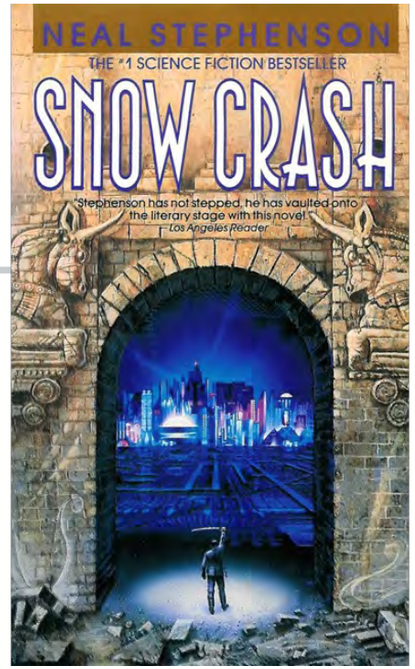
“Cultural: which affects how people interpret and orient themselves includes ideological concepts, language, religion, and psychological issues including fears, attitudes, etc. Institutional: which embody cultural ideas and norms as practices and beliefs in political, military, economic, or legal systems. Technological: includes how communities shape the environment and create infrastructure with technology and media. Physical: where people live and where resources are found.”

https://smallwarsjournal.com/jrnl/art/does-the-human-domain-matter

The Noosphere: where having better understanding of human cognitive, social, and behavioral patterns, systems, and differences means having "advantage".

From Adam Russell (University of Maryland/USC ISI)

Today's Noosphere, hacked by AI!



"AI has gained some remarkable abilities to manipulate and generate language, whether with words, sounds or images. AI has thereby hacked the operating system of our civilisation."

Yuval Noah Harari in The Economist, Apr 2023

But we are also hacked by

- The social-media-driven "attention economy"

  "the so-called perfect Entertainment… that danger of Entertainment so fine that it will kill the viewer… The Entertainment exists."

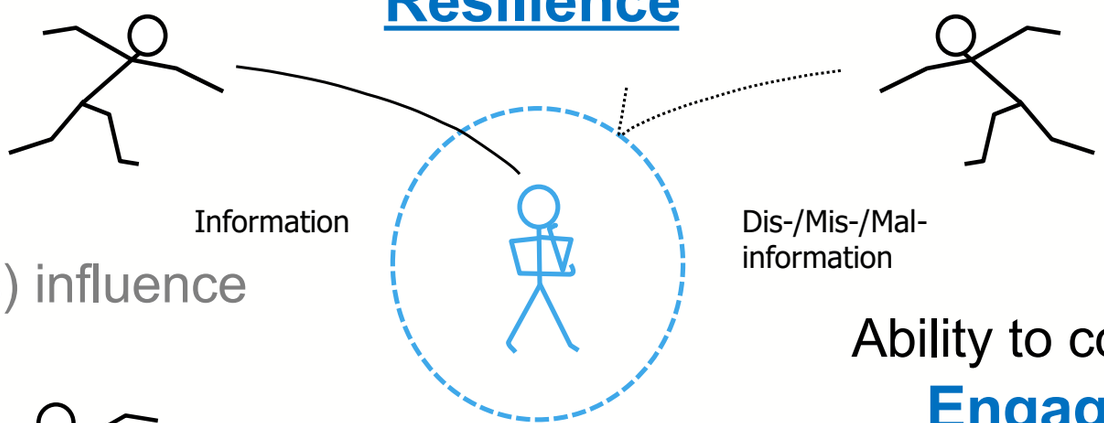
  — Infinite Jest (1996), David Foster Wallace, pp 318-319
- Malinformation & disinformation
- State and non-state actors exploiting social media
  - Internet Rule 66: Everything has a fandom, everything.

"Mobile Lovers", Attr. Bansky (2014)

The Need for Cognitive Security

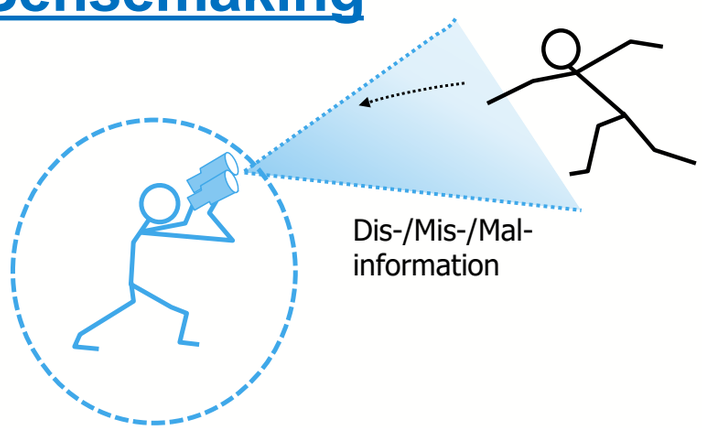
Susceptibility to (malicious) influence

Resilience



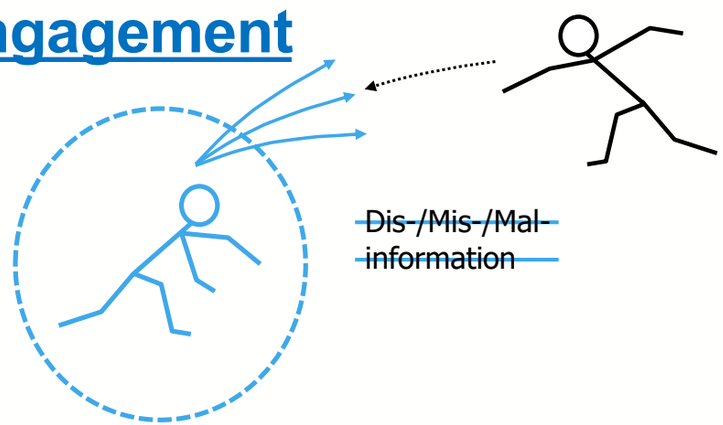
Awareness of (malicious) influence

Sensemaking

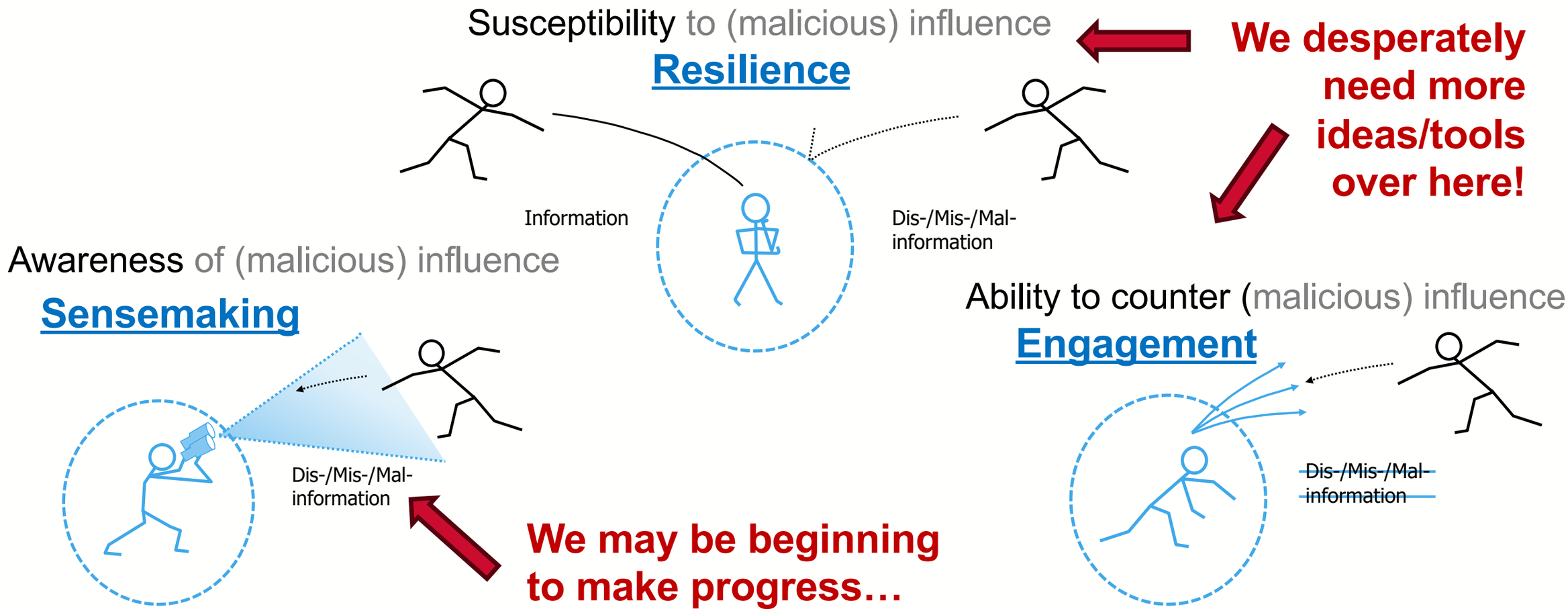


Ability to counter (malicious) influence

Engagement



The Need for Cognitive Security



(Example) **Sensemaking:** **Understanding Emotion**

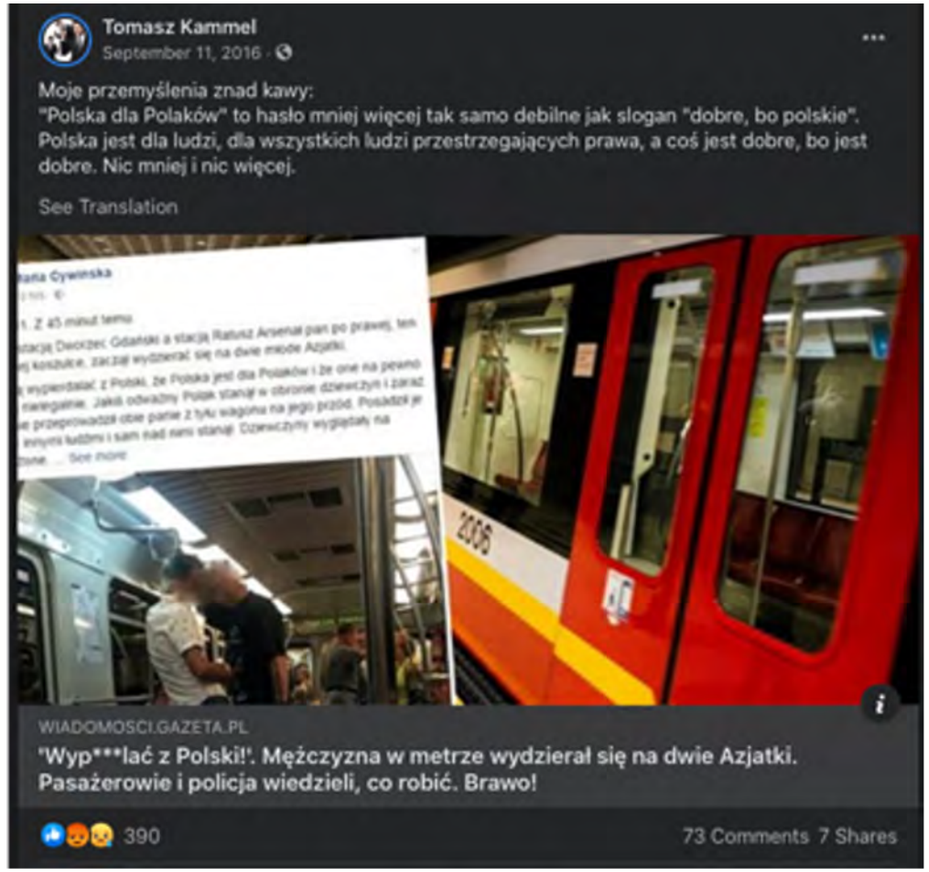
Translation

"My thoughts over coffee: "Poland for Poles" is a slogan more or less as dumb as the slogan "good because it's Polish". Poland is for people, for all law-abiding people, and something is good because it is good. Nothing less and nothing more."

Text on the bottom: " ‘Get the f**k out of Poland!’ A man on the subway yelled at two Asian women. Passengers and police knew what to do. Bravo!"

Emotion Annotation

Anger: 50	Hate: 20
Contempt: 30	Admiration: 43
Pride: 5	Sadness: 10
Fear: 4	Excitement: 18



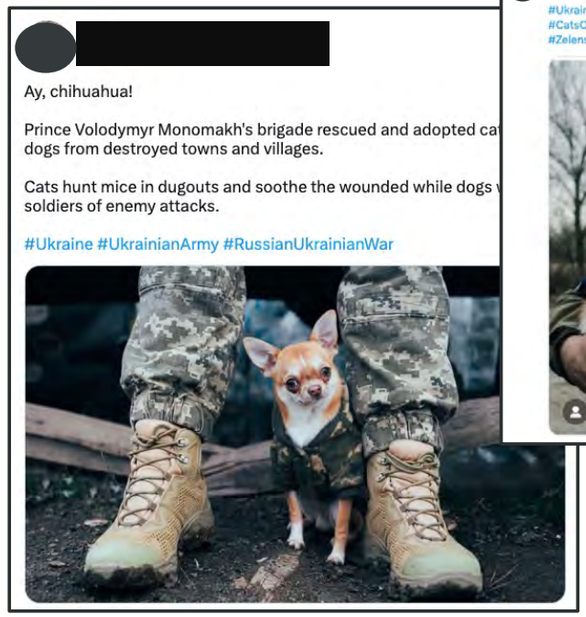
(Example) **Sensemaking: Understanding Manipulation**

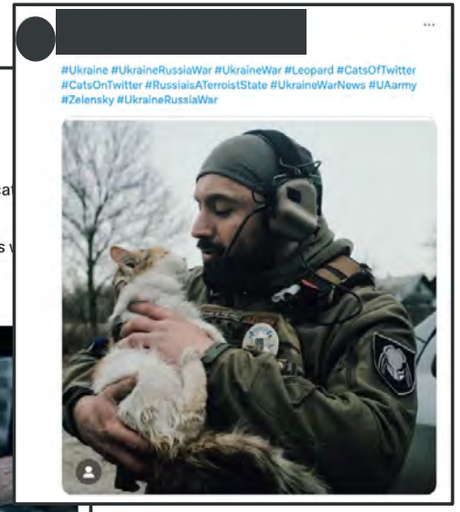
From Dabiq propaganda magazine, https://www.independent.co.uk/news/world/middle-east/isis-kittens-honey-bees-dabiq-propaganda-recruits-photo-soften-image-terror-a7168586.html

1. What are the dimensions that make up "cute" content and reactions?

   Does "cute" social media content evoke kama muta*?

2. Does the amount (intensity and frequency) of "cute" content in a social media post predict sharing behaviors on social media?





Twitter April 2023, Ukraine/Russia War

Kama muta: from Sanskrit "moved by love", a heartwarming feeling of being moved or touched

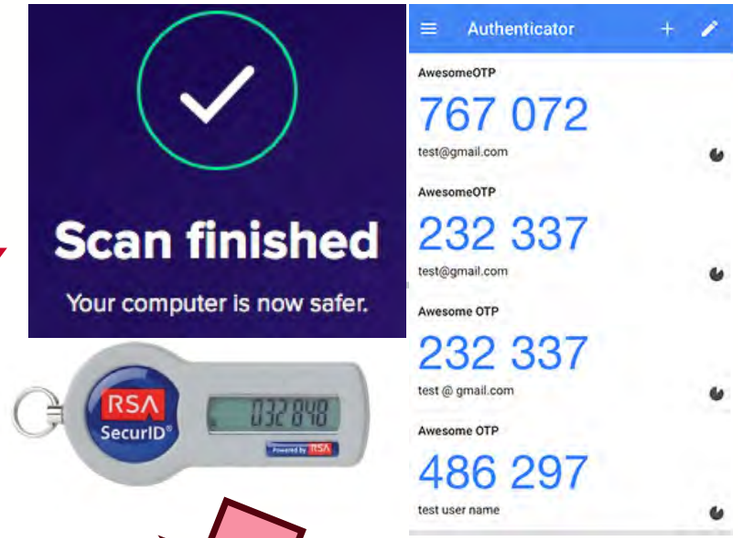
The University of Maryland, Golonka, Jones, et al., 2023, *Frontiers in Psychology*

UNIVERSITY OF MARYLAND

Advancing **Engagement:**
How should we respond?

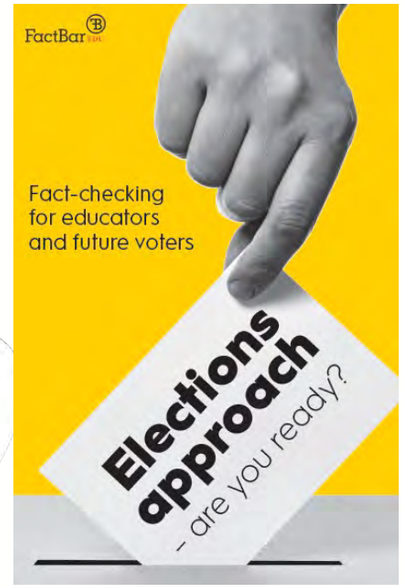
- Lessons from related areas, such as cyber, indicate we need new tools and technologies
  - Software for attribution, provenance, detection of AI-modified content, etc
- The problem itself is socio-cultural
  - New forms of education and training are needed
  - Trusted organizations and trusted governments are needed
  - Caution: The Observer Effect & The Hawthorne Effect

Key question: How to enable and protect new forms of collective action that improve and enhance humankind's resilience?

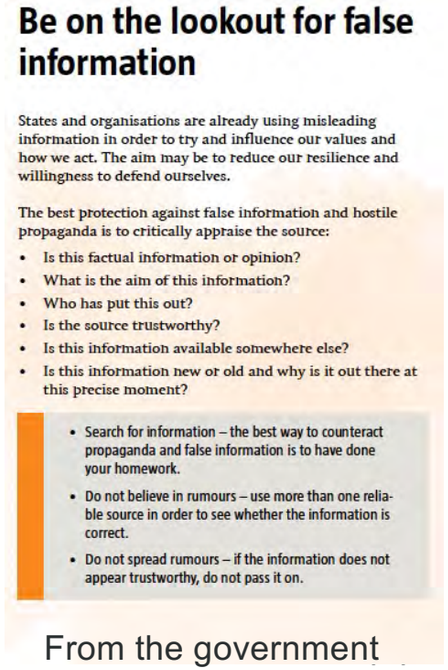


EU DISINFO LAB
We gather. We source. We map.

Stanford | Internet Observatory
Cyber Policy Center



From the EU & Finland's Faktabaari



From the government of Sweden

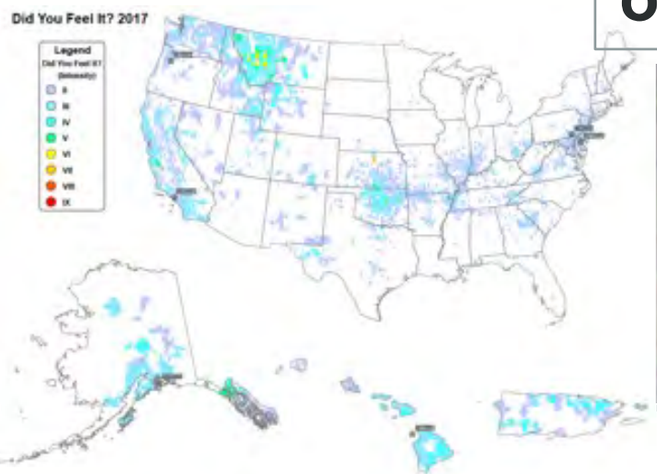
Enabling **Resilience**: **Can we build shared understanding our situation?**

How do we get from saying "its raining outside" to understanding cognitive "weather" and "climate"?

- Large, multi-disciplinary, models are required
- Rules, laws, treaties are needed to govern how such models are built & used
  - And note that our observations change the system

https://en.wikipedia.org/wiki/NOAA

Examples in other areas:



https://en.wikipedia.org/wiki/Bloomberg_Terminal

https://www.cdc.gov/nwss/wastewater-surveillance.html https://en.wikipedia.org/wiki/USGS

Historical Context:

Creation of SAGE and the DEW Line against cold war threats (1957)

https://en.wikipedia.org/wiki/Semi-Automatic_Ground_Environment

https://en.wikipedia.org/wiki/Distant_Early_Warning_Line

SAGE: Semi-Automatic Ground Environment
DEW: Distant Early Warning

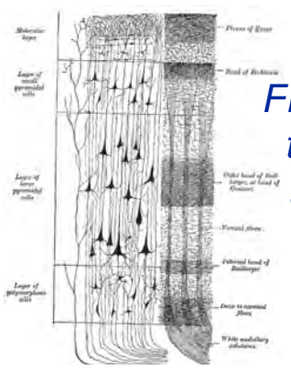
16

Hamming Questions for the Noosphere

How can we address the fundamentals, the deep scientific and engineering questions, that require answers in the Noosphere?

Understand Foundations

- Use of Social Sciences to understand behavior of individuals or groups
- The physics & biology of nature's processes for computation

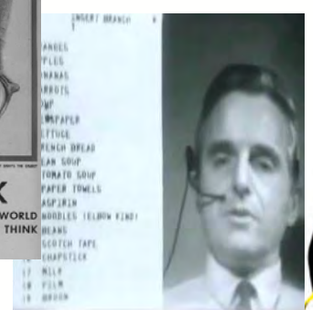


From neurons to people & societies…

From UMD/RPI/USC

Shift Complexity

- Enable integration of humans & machines (symbiosis)



… from human interfaces to integration…

Drexel Univ.
Prof. K. Izzetoglu

Systems Design

- Engineering control of Risk, Safety & Trust, Bias & Influence

…to whatever this is…

Hamming Questions for the Noosphere

Understand Foundations

- Find fundamental metrics
  - Physics has "SWAP"
- Develop new scientific methods and tools
  - Enable reproducibility, reduce bias, change rewards
- Create theories via hypothesis-driven inquiry
  - This is a new landscape requires a new language for interdisciplinary science
- Achieve scale
  - Is there a LIGO, CERN, JWST, or Human Genome?

Shift Complexity

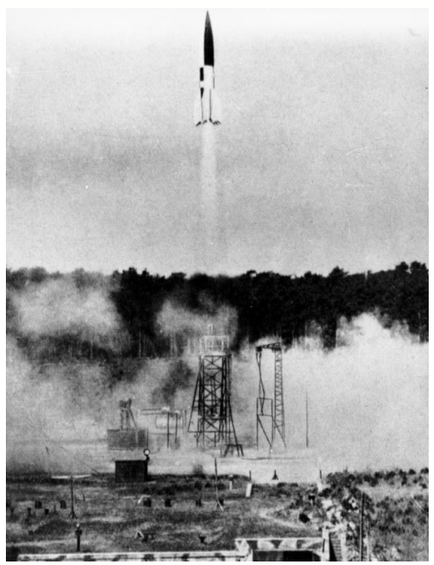
- Create Human-Machine approaches to problems…
  - Technology question: How will new human/machine interaction affect us? LLMs, AR/VR, BCIs, etc

… while retaining human agency and ensuring value alignment

- Scientifically address the questions from psychology, ethics, philosophy, and law
- Identify precise means of measuring and ensuring transparency and explainability

Systems Design

- Establish systems safety norms
  - What are the ethical and legal rules?
  - Every question now has human beings "in-the-loop"; ethical norms are in flux
- Design systems that actually improve decision-making and resilience
  - … and avoid algorithmic authoritarianism
- Test & Evaluate
  - Simulation? In-the-wild? Live-Virtual-Constructive?



An observation: Rocket Science in 1942 ≈ Noosphere 2023

(An anecdote from Von Braun: Dreamer of Space, Engineer of War by Michael Neufeld, 2008)

Pic: Bundesarchiv, Bild 141-1880 / CC-BY-SA 3.0, CC BY-SA 3.0 de, https://commons.wikimedia.org/w/index.php?curid=5338474

A fundamental noosphere metric

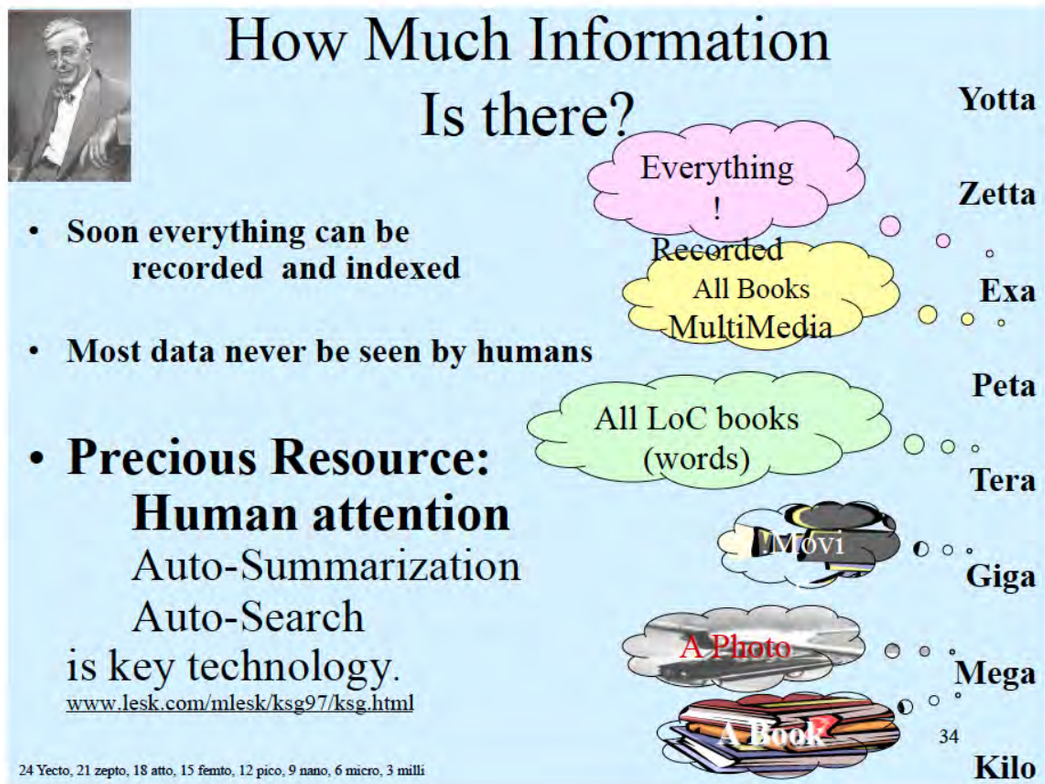
The fundamental metric: How well do we safeguard ***Human Attention?***

Key design issue: How to allocate attention while retaining human agency?

- What is responsible delegation to AI?

Safeguards are urgently needed:

- **For AI where "Machine-as-Tool":** Machines magnify our existing processes, existing processes get new scale/mass
- **For AI where "Machine-as-Partner"** Machines create new decision processes



From: Jim Gray, 1998 ACM Turing Award Lecture

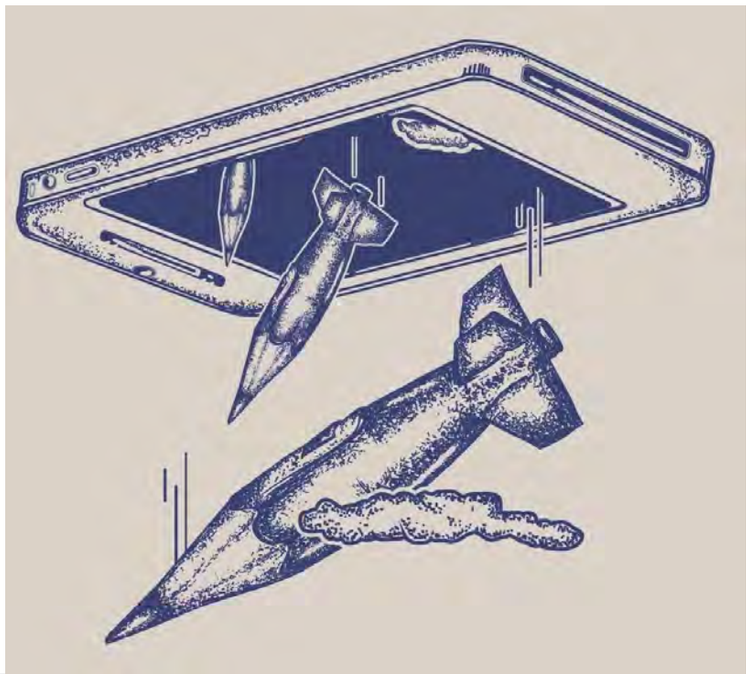
Connecting to CLSAC: Security of the Analytic, by the Analytic and for the Analytic

Situation Understanding

- Enhancing OSINT pipelines
- Auto-summarization and analysis ("explain this")
- Activity recognition
  - Complex activities, short time windows
- Developing corpora
  - De-identified data
- Modeling & Simulation
  - Realistic multi-agent models of human/social behaviors
- Real-time predictive analytics
  - Meme detection and emergence; campaign identification; etc
- Tools to help human analysts

A **new type of thinking** is essential if mankind is to survive and move toward higher levels.

--- Albert Einstein
writing in the *New York Times*
25 May 1946



UNIVERSITY OF MARYLAND

https://bjpeters.medium.com/big-tech-is-weaponized-media-or-why-social-media-even-after-deplatforming-still-backs-the-trump-4585cf0e3151

Some references for the sociotechnical

1. As We May Think, Vannevar Bush, The Atlantic, July 1945
2. Science and Complexity, Warren Weaver, 1947
3. Man-Computer Symbiosis, J.C.R. Licklider, IRE Trans. on Human Factors in Electronics, Vol HFE-1, pp 4-11, Mar 1960
4. Augmenting Human Intellect: A Conceptual Framework, SRI Report AFOSR-3223, Engelbart, Douglas C, October 1962
5. The Architecture of Complexity, Herbert Simon, Proceedings of the APS, 106(6):467-482 Dec. 1962
6. The Structure of Scientific Revolutions, Thomas Kuhn, 1962
7. Understanding Media: The Extensions of Man, Marshall McLuhan, 1964
8. The Sciences of the Artificial, Herbert Simon, 1969
9. The Selfish Gene, Richard Dawkins, 1976
10. Infinite Jest, David Foster Wallace, 1996
11. The Origin of Wealth, Eric Beinhocker, 2007
12. Defending against “The Entertainment”, Bulletin of the Atomic Scientists, William Regli, 2018
13. LikeWar: The Weaponization of Social Media, P. W. Singer and Emerson T. Brooking, 2018
14. Sapiens (2015) & Homo Deus (2017), Yuval Noah Harari
15. On Intelligence (2004) & A Thousand Brains (2021), Jeff Hawkins et al

The Detroit Industry Murals of Diego Rivera (1933)

William Regli
Professor of Computer Science
The University of Maryland at College Park
regli@umd.edu

UNIVERSITY OF
MARYLAND

CLSAC 2023

Security of the Analytic, by the Analytic and for the Analytic

October 30-- Nov 2, 2023
Westin Annapolis
Annapolis, Maryland

Theme:

Analytics are pervasive in our lives, ranging from healthcare, transportation, finance, energy, agriculture, weather, science, and government. Analytics and systems they run on can be decades old and extraordinarily slow to change (e.g. IRS, FAA traffic control). On the other hand, they often require low-latency computing on ephemeral, real-time data, running on cutting edge hardware that rapidly changes due to the constant race for competitive advantage. These two extremes make it very difficult to insure the correctness, resilience, and security of the analytic. Security is too often an underappreciated afterthought or cost to be avoided in the full system design and operation of the analytic. The high consequence impact of 'getting things wrong' can be life-changing.

Hardware, software, processes, supply chain, resilience at all levels, and the enormous interdependent complexities of systems all impact system vulnerability. This complexity greatly increases the attack surface and consequently, we are in a highly asymmetric arms race between defenders and attackers (primarily well-funded state actors and criminal enterprises) in favor of the attackers.

In CLSAC 2023, we explore how to 'get things right' when it comes to the role of security in the analytics world. In 'Of the Analytic' we explore how security is addressed by the analytic itself – its architecture, design, tools, and methodologies. In 'Security by the Analytic' we look at analytics in the role of providing security. Lastly, in 'For the analytic' we review environments/tools that form the ecosystem that provides security to analytics.