I2O Office Wide Proposers Day

November 7, 2024

DARPA

DISTRIBUTION STATEMENT A: Approved for public release, distribution unlimited

I2O Office Wide Proposers Day Agenda

10:00	*11:00*	*Check-in and Networking Coffee*
11:00	11:05	Security Overview
11:05	11:15	Opening Remarks – Rob McHenry
11:15	11:35	How to Work with DARPA
		Commercial Strategy – Jen Thabet
		Small Business – Aaron Sparks
		DARPA Connect – Sana Sood
11:35	12:35	I2O Strategy – Kathleen Fisher
12:35	*1:35*	*Break for Lunch*
1:35	2:05	PM Presentations – (Dewhurst, Bernsen, Sweet, Kuhn, Cook)
2:05	2:15	Delivering on the DARPA Mission – Matt Turek
2:20	3:55	Sidebars

Opening Remarks

Defense Advanced Research Projects Agency

Rob McHenry

Deputy Director

November 2024

Distribution Statement A: Approved for public release; distribution unlimited

Prevent U.S. technological surprise

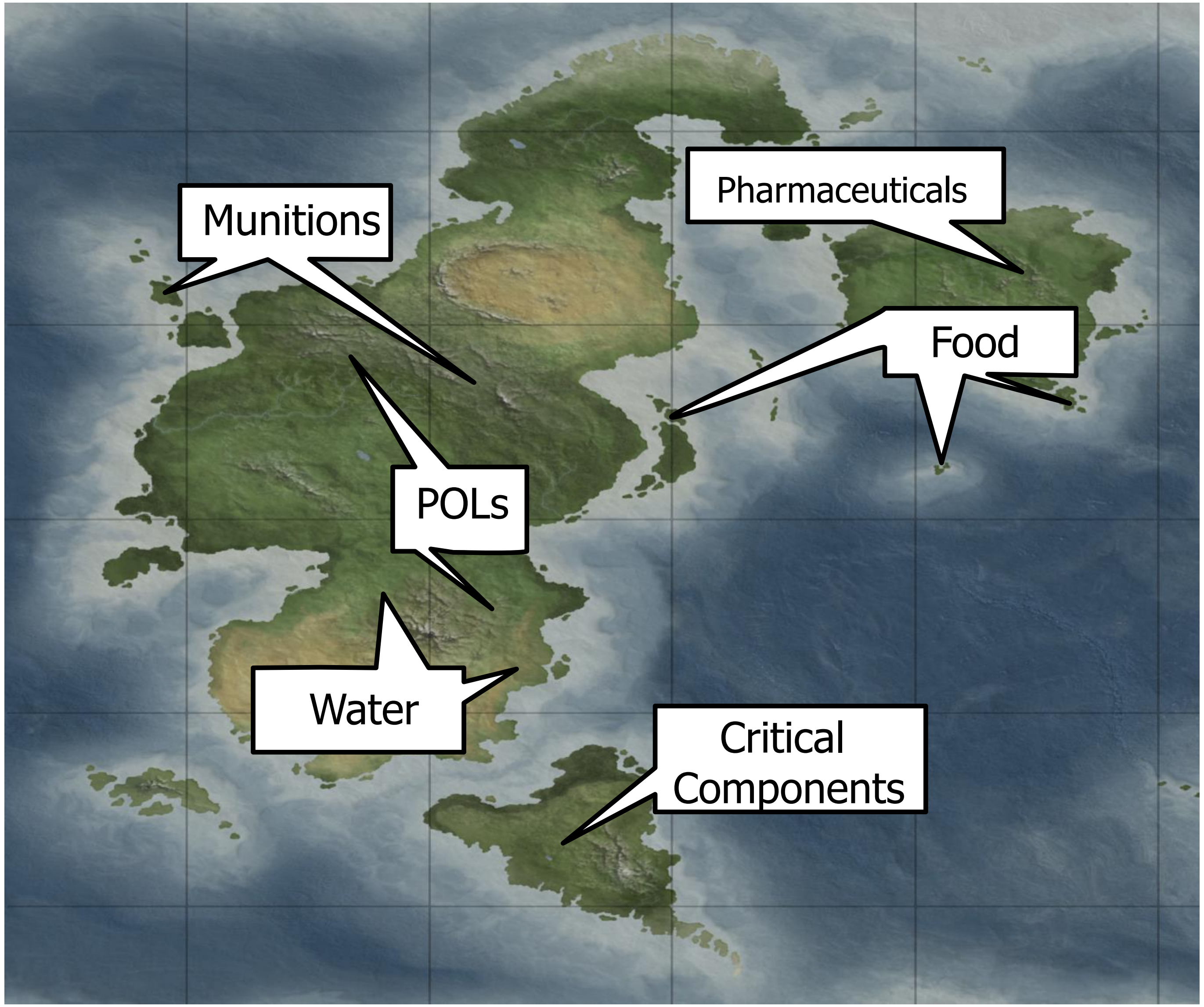
Maintain U.S.
technological superiority

DARPA

Created in 1958
as “never again” response to Sputnik

The independent science and
technology agency of the DoD

Disruption: Making at the Point of Need



DARPA Insight: make it, don't take it – foraging in the 21st century

In a prolonged conflict, we want to enable the fighting force with targeted production, while foraging appropriately and flexibly in/near the battlespace

Sustainment:

- Support the warfighter and warfighting activities to extend combat effectiveness post-30 days

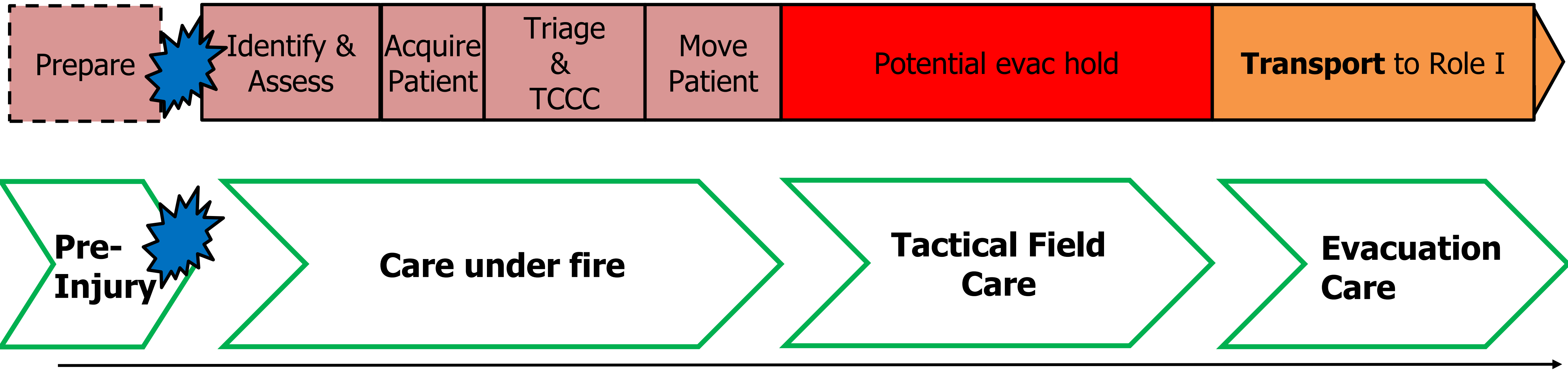
Flexibility:

- Make use of partner/ally/ battlespace infrastructure.
- Assess and adapt available inputs for mission-appropriate performance outputs.

POL: petroleum, oil, and lubricants.

Disruption: Combat Casualty Care

Live Chain



Prepare

Identify & Assess

Acquire Patient

Triage & TCCC

Move Patient

Potential evac hold

Transport to Role I

Pre-Injury

Care under fire

Tactical Field Care

Evacuation Care

POI

Self-care

Buddy-care

Medic-care

Field Stabilization (CCP)

Role 1 stabilization

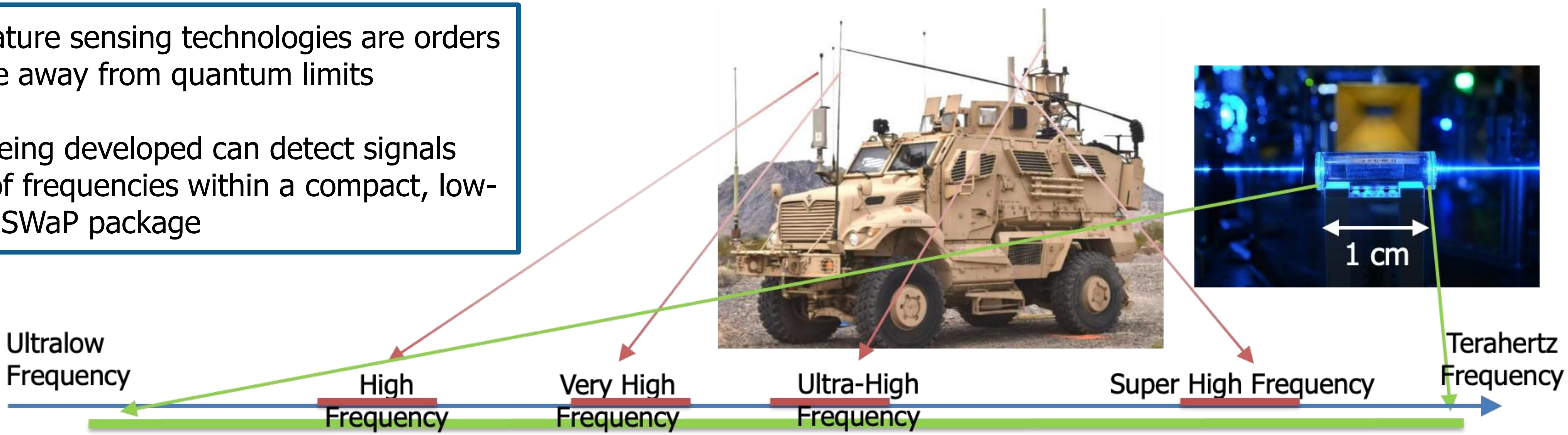
DARPA seeks to enhance treatment options at and near the point of injury (POI) by providing an easy to administer self-help/first aid

CCP: Casualty Collection Point

Disruption: Quantum Sensing

Current room temperature sensing technologies are orders of magnitude away from quantum limits

Rydberg sensors being developed can detect signals across a wide range of frequencies within a compact, low-SWaP package



Ultralow Frequency

High Frequency

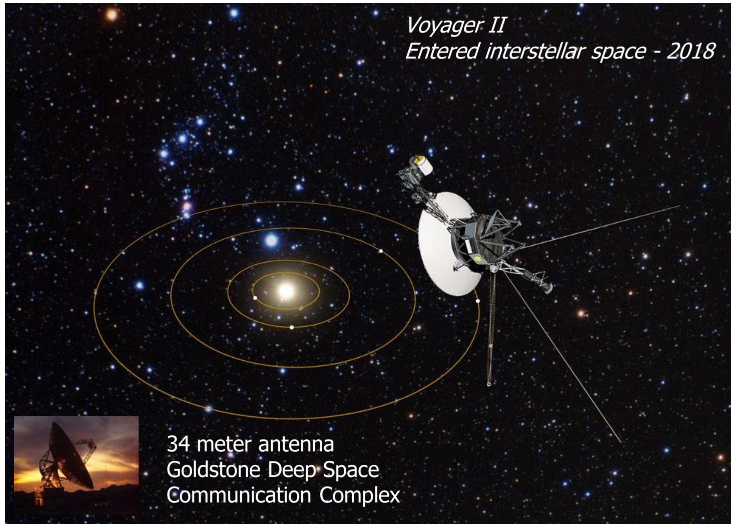
Very High Frequency

Ultra-High Frequency

Super High Frequency

Terahertz Frequency

Quantum receivers are improving in sensitivity faster than the reduction of Voyager II signal strength



Voyager II
Entered interstellar space - 2018

34 meter antenna
Goldstone Deep Space
Communication Complex

SWaP: Size, weight, and power

Disruption: Quantum Computing

Quantum Benchmarking (QB)

Would a very powerful
quantum computer be industrially useful?

Underexplored Systems for Utility-Scale Quantum Computing

Can efforts to build a very powerful quantum
computer in the near term succeed?

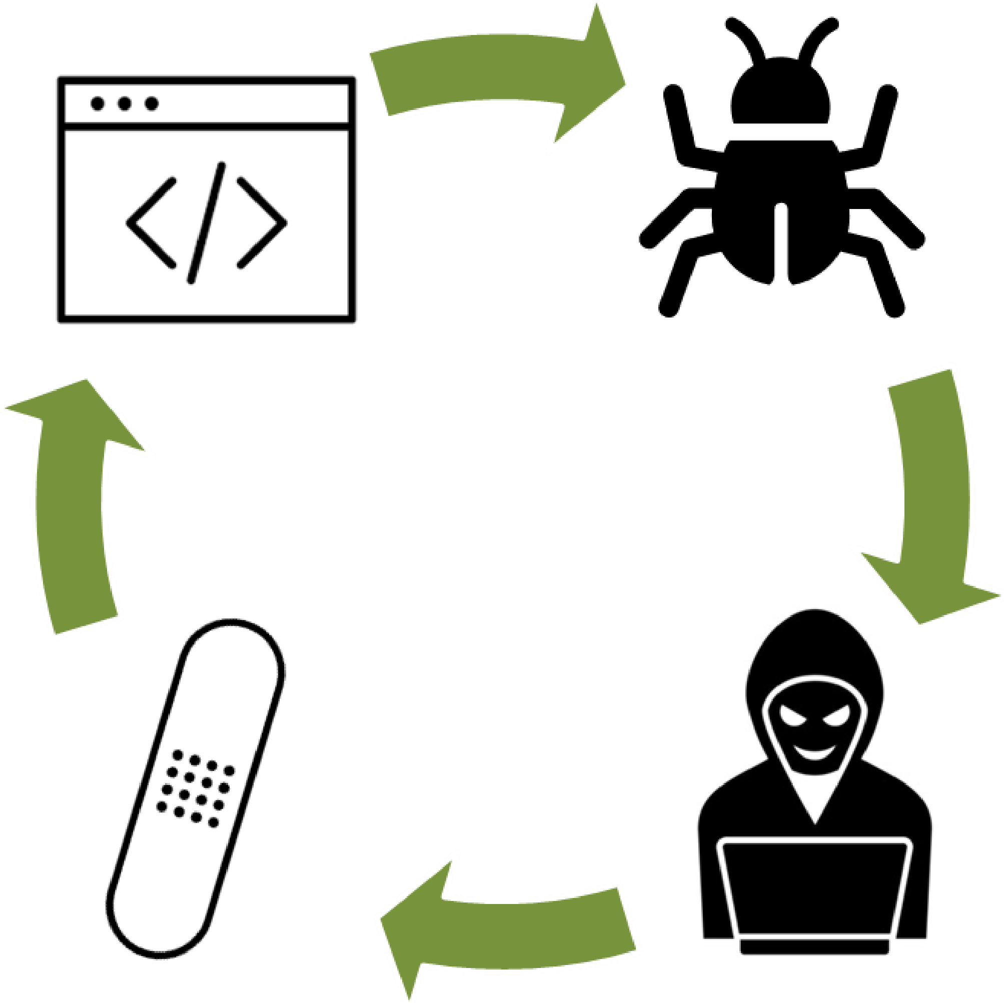
US2QC

Quantum Benchmarking Initiative (QBI)

Verify and validate if any quantum computing
approach can achieve utility-scale operation by
the year 2033

QBI

Disruption: Resilient Software Systems



$6.6B

Annual direct cost of DoD cybersecurity

- High-Assurance Cyber Military Systems (HACMS)
- Safe Documents (SafeDocs)
- Pipelined Reasoning of Verifiers Enabling Robust Systems (PROVERS)
- Verified Security and Performance Enhancement of Large Legacy Software (V-SPELLS)
- Assured Micropatching (AMP)
- Hardening Development Toolchains Against Emergent Execution Engines (HARDEN)
- Cyber Assured Systems Engineering (CASE)
- Automated Rapid Certification Of Software (ARCOS)

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

Agency Strategic Focus Areas

Disruption

Velocity

Execution

Impact

DARPA

www.darpa.mil

How to Work with DARPA

Commercial Strategy

DARPA Commercial Strategy Overview

November 2024

Mr. Aaron Kofford
Senior Advisor, Commercial Strategy
Director’s Office

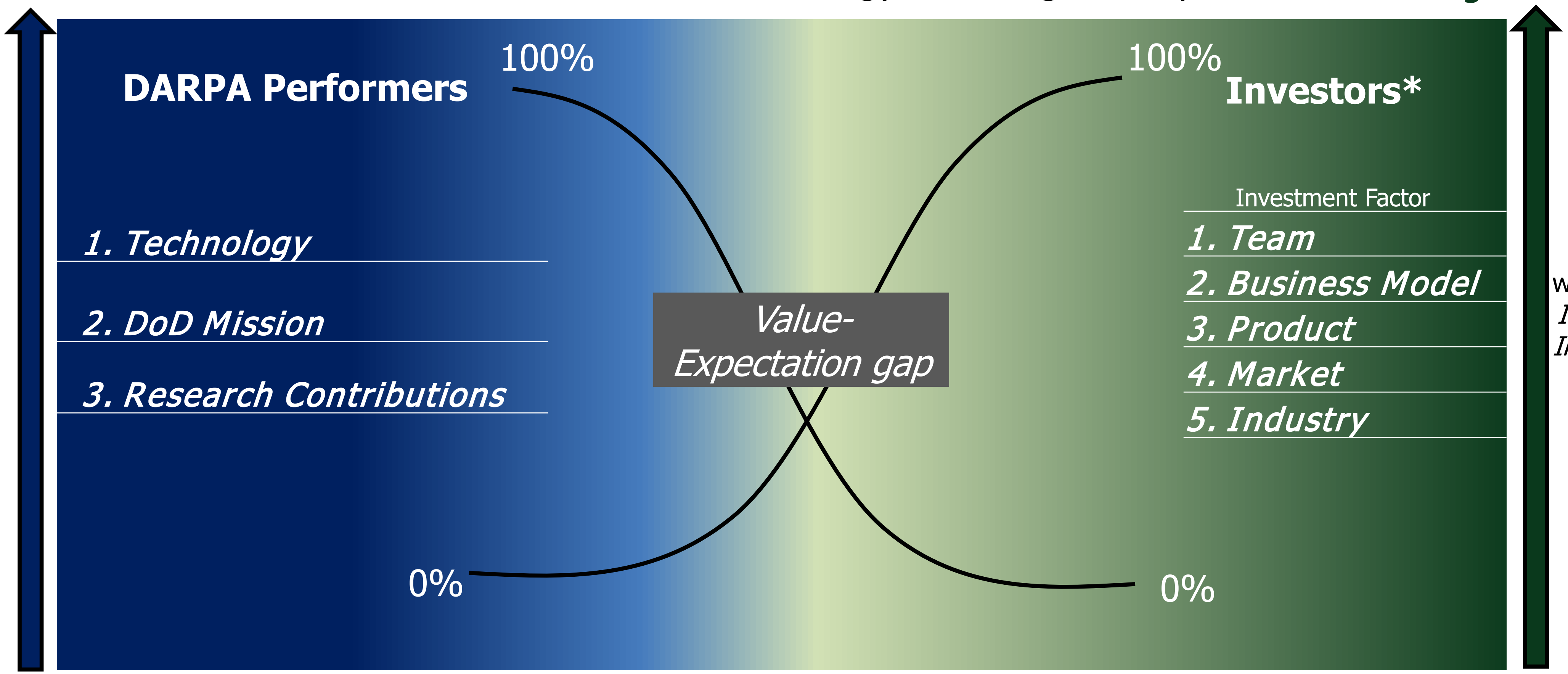
Distribution Statement “A”
(Approved for Public Release, Distribution Unlimited)

The Value-Expectation Gap between DARPA and Investors

DARPA Commercial Strategy is Closing the Gap

Highest Value

Highest Value



DARPA Performers

1. *Technology*
2. *DoD Mission*
3. *Research Contributions*

Value-Expectation gap

Investors*

Investment Factor

1. *Team*
2. *Business Model*
3. *Product*
4. *Market*
5. *Industry*

Lowest Value

Lowest Value

*1,000 Private Investors surveyed were asked their *Most Important Factors for Investment Decisions.*

Harvard Business Review, 2023[1]

1. https://academic.oup.com/jleo/article-abstract/35/3/513/5530735
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8888446/

Overcoming the DARPA-Commercial Value Expectation Gap

DARPA COMMERCIAL STRATEGY'S COUNTERMEASURES

Embedded Entrepreneur Initiative (EEI)

Connecting Brilliant Technical Minds to Brilliant Business Minds

- Program Manager (PM) nominated for SCA review; technology validated
- Entrepreneur joins a DARPA Performer R&D Team for 12+ months
- Access to techno-economic market mapping and curated capital

Tiger Teams

Tailored Commercial Strategies for Programs

- Nominated by DARPA Office Director (OD) or Deputy Director (DD), PM, or Liaison Officer (LNO)
- Less frequent, often classified
- Commercialization opportunities with high-value national and economic security considerations

Minimum Viable Products (MVPs)

Prove Commercial Viability, Service, and/or Capability

- PM nominated for SCA review
- Performer brings technology to market with the goal of generating sales and productizing

Venture Horizons

Bring Top Investment to DARPA

- RFI Released Fall 2024
- Top-tier investors who meet rigorous standards connected to DARPA
- PMs, OD, DDs, LNOs can update investors on DARPA programs

COMMERCIAL MECHANISMS TO SCALE

Commercial Solutions Opening (CSO)

Drive Commercial Solutions Derived from DARPA R&D

The contract vehicle by which Awardees can provide commercial solutions. Awards support the EEI as it looks to reduce business risk(s).

DARPA Commercial Accelerators

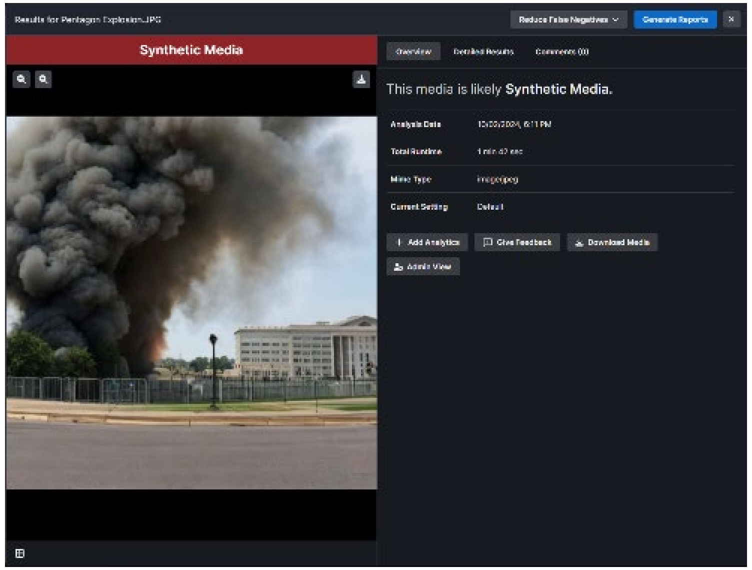
Scale the EEI/MVP

Accelerators will facilitate rapid commercialization, talent attraction, and commercial ecosystem development.

Commercial Strategy + I2O Case Study

Program: SemaFor SEMANTIC FORENSICS

- Partnering with a major news outlet and company working with Hollywood
- Workshops in progress to assess commercial use cases of semantic technologies for analyzing media



Results for Pentagon Explosion.JPG

This media is likely **Synthetic Media.**

Analysis Date	10/30/2024, 6:11 PM
Total Runtime	1 min 47 sec
Mime Type	image/jpeg
Current Setting	Default

Senior Commercial Advisor Conducts:

- Ongoing Coaching & Mentoring
- Go to Market Strategy
- Business Structuring
- Business Modeling
- Financial Modeling
- Financial Structuring
- Intellectual Property Strategy
- Techno-Economic Market Mapping
- Introduction to Private Investors
- Introduction to Customers
- Recruitment to Embedded Entrepreneur

Gain Commercial Insights

(top-tier connections, enhanced program support)

Reach Out to Commercial Strategy

The Embedded Entrepreneur Initiative

Connecting Brilliant Technical Minds to Brilliant Business Minds

Learn More at:

https://eei.darpa.mil/

DARPA

www.darpa.mil

Small Business Programs Office (SBPO)

Small Business Opportunities with DARPA

Small Business Programs Office

703-526-4170 | sbir@darpa.mil
http://www.darpa.mil/work-with-us/for-small-businesses

Distribution Statement "A"
(Approved for Public Release, Distribution Unlimited)

DARPA SBIR/STTR Program Details

SBIR·STTR
POWERED BY INNOVATORS
MOVING TECHNOLOGY FORWARD

DARPA's mission is to make pivotal investments in breakthrough technologies for national security.

Uniqueness

- Program manager-centric
- Just-in-time topic development
- Topics tied to DARPA programs
- SBIR XL
- Transition & Commercialization Support Program; No Technical and Business Assistance (TABA) Funding

Funding

SBIR Program

3.2% of all extramural RDT&E

FY24 – $113.4M

STTR Program

.45% of all extramural RDT&E

FY24 – $15.9M

Program Structure

Phase I

- $250,000-$275,000
- ~ 6 months
- Feasibility Study

Phase II

- $1,800,000
- 24-36 months
- Continued Research and Prototype
- Adoptions/Co-funds

Phase II Enhancement

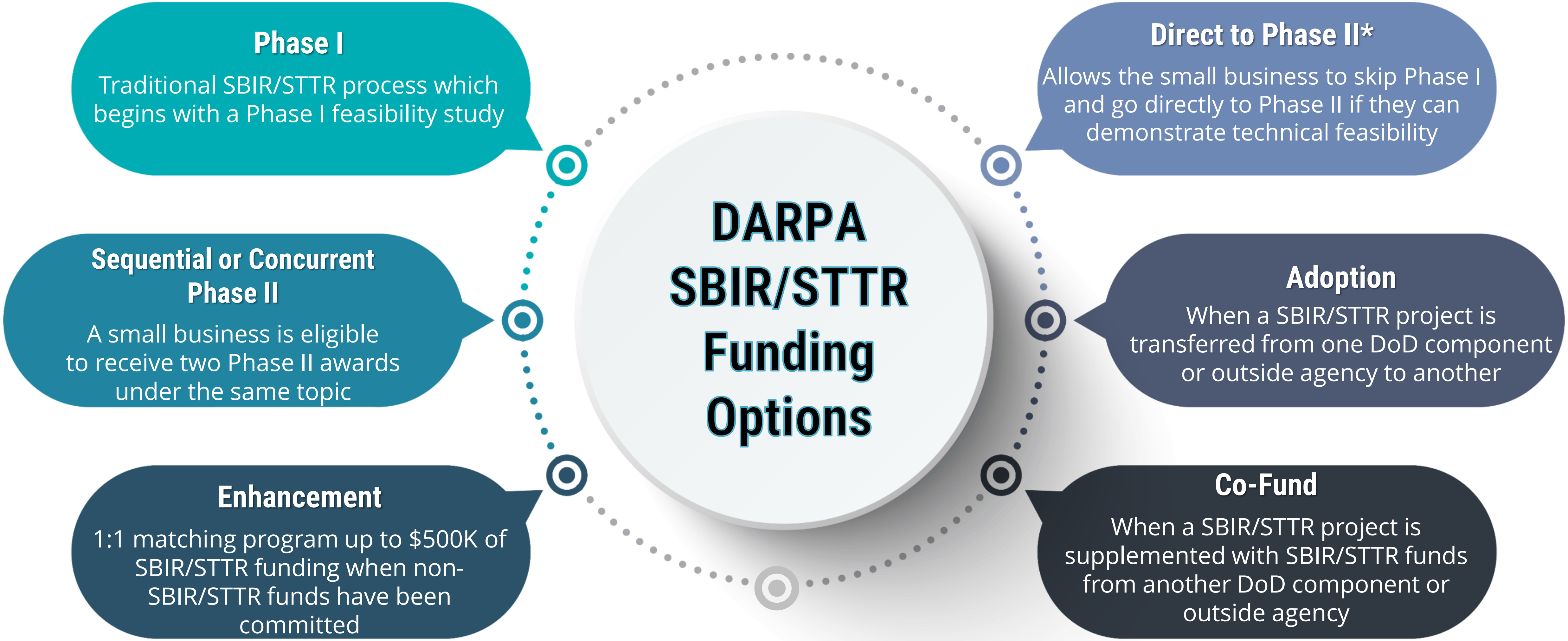
- $1:$1 Match
- Up to 12 months
- Up to $500K

Phase III

- No time limit
- No SBIR/STTR funds

DARPA is funded at $4.1B and executes unique SBIR/STTR programs focused on lower Technology Readiness Level (TRL) efforts, offering greater flexibility for DARPA Program Managers (PMs) with diverse transition opportunities for performers.

Creatively Enabling SBIR/STTR Program Utilization



DARPA SBIR/STTR Funding Options

Phase I

Traditional SBIR/STTR process which begins with a Phase I feasibility study

Direct to Phase II*

Allows the small business to skip Phase I and go directly to Phase II if they can demonstrate technical feasibility

Sequential or Concurrent Phase II

A small business is eligible to receive two Phase II awards under the same topic

Adoption

When a SBIR/STTR project is transferred from one DoD component or outside agency to another

Enhancement

1:1 matching program up to $500K of SBIR/STTR funding when non-SBIR/STTR funds have been committed

Co-Fund

When a SBIR/STTR project is supplemented with SBIR/STTR funds from another DoD component or outside agency

SBIR XL Pilot*

Increases opportunities for DARPA-funded technology by reimagining SBIRs to transform ideas into successful small businesses that scale

*SBIR only (no STTR)

Breaking Down Barriers to Entry for Nontraditional Performers

DARPAConnect is designed to broaden DARPA's reach and stimulate **growth and collaboration** between DARPA, businesses, and academia.

- Regional and Virtual Events
- Networking Opportunities
- Training and Development
- Customized Support and Mentoring

https://www.darpaconnect.us/
DARPAConnect@darpa.mil

Distribution Statement "A"
(Approved for Public Release, Distribution Unlimited)

Small Business Programs Office

(703) 526-4170

sbir@darpa.mil

https://www.darpa.mil/work-with-us/for-small-businesses

DARPAConnect

DARPA CONNECT

DISCOVER · COLLABORATE · CONTRIBUTE

www.DARPAConnect.us
DARPAConnect@darpa.mil

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Breaking Down Barriers to Entry for Nontraditional Performers

DARPA CONNECT

DISCOVER · COLLABORATE · CONTRIBUTE

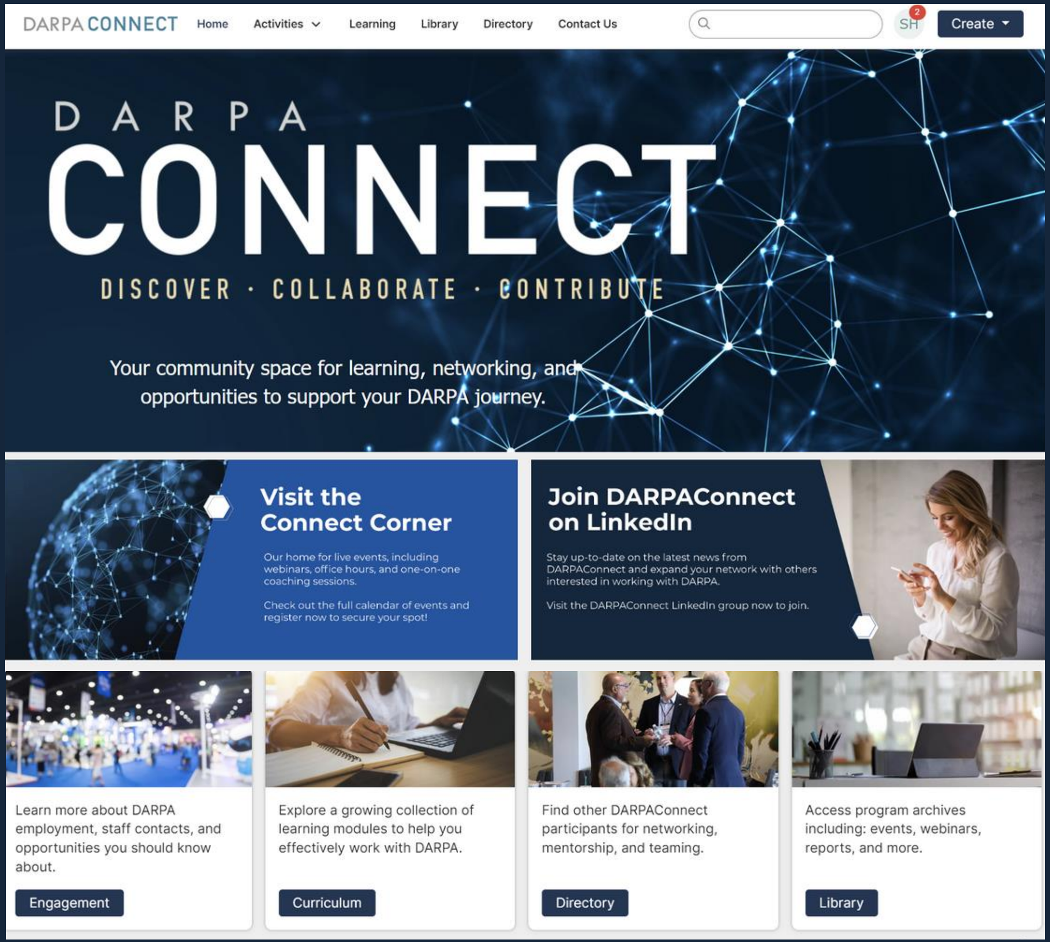
Supporting Untapped Innovators: Breaking Down Barriers to Entry

- Regional and Virtual Events
- Networking Opportunities
- Training and Development
- Customized Support and Mentoring

www.DARPAConnect.us

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Join at DARPAConnect.us



DARPA CONNECT

DISCOVER · COLLABORATE · CONTRIBUTE

Your community space for learning, networking, and opportunities to support your DARPA journey.

Visit the Connect Corner

Our home for live events, including webinars, office hours, and one-on-one coaching sessions.

Check out the full calendar of events and register now to secure your spot!

Join DARPAConnect on LinkedIn

Stay up-to-date on the latest news from DARPAConnect and expand your network with others interested in working with DARPA.

Visit the DARPAConnect LinkedIn group now to join.

Learn more about DARPA employment, staff contacts, and opportunities you should know about.

Engagement

Explore a growing collection of learning modules to help you effectively work with DARPA.

Curriculum

Find other DARPAConnect participants for networking, mentorship, and teaming.

Directory

Access program archives including: events, webinars, reports, and more.

Library

Join the LinkedIn Group

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

DARPAConnect Pop-Ups

Full day events covering the breadth of your DARPA journey

Albany, NY

November 14, 2024

Sessions Include:

- Engaging DARPA Program Managers
- Heilmeier Catechism: Understanding Effective DARPA Communication
- Understanding DARPA Announcements and Contract Vehicles
- Reviewing and Analyzing a DARPA Opportunity
- DARPA SBIR and STTR Program
- Understanding DARPA Security Resources
- Preparing Your DARPA Proposal
- Tying It All Together: Strategies for Success
- Opportunities for Networking

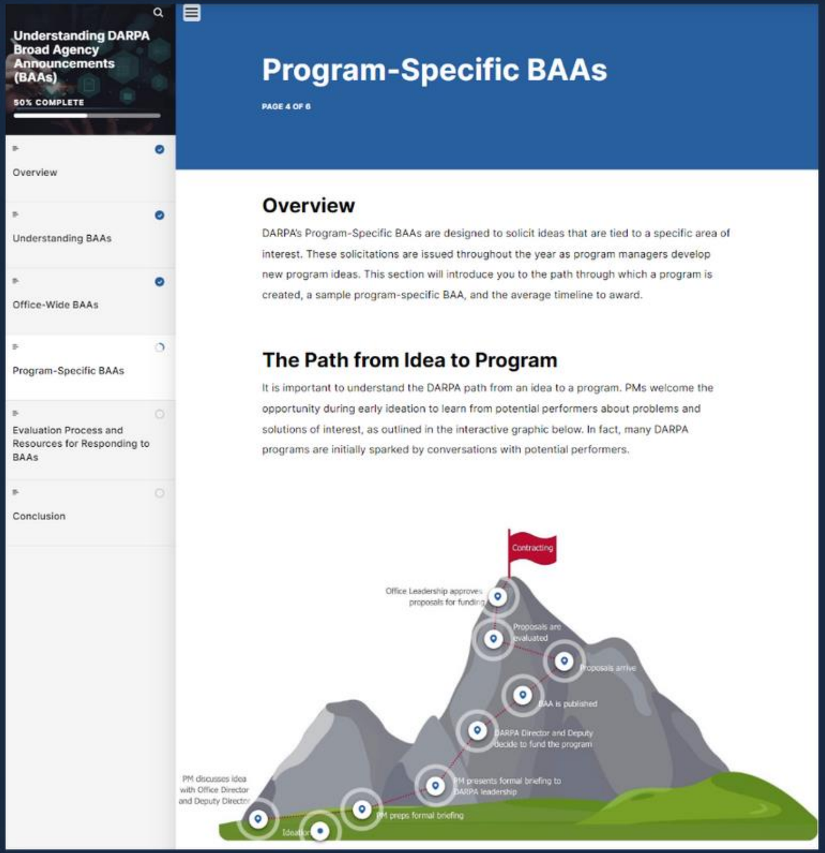
Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

DARPAConnect Curriculum

Lessons Include:

- Understanding BAAs
- SBIR/STTR
- DARPA 101
- DARPA Award Vehicles & Solicitations
- Proposal Tips
- Preparing for Proposers Day
- Heilmeier Catechism
- Engaging DARPA Program Managers
- Becoming a PM
- DARPA Innovation Fellows
- Introduction to Security
- Global Participation & Engagements

www.DARPAConnect.us


Understanding DARPA Broad Agency Announcements (BAAs)
50% COMPLETE
Overview
Understanding BAAs
Office-Wide BAAs
Program-Specific BAAs
Evaluation Process and Resources for Responding to BAAs
Conclusion

Program-Specific BAAs

PAGE 4 OF 6

Overview

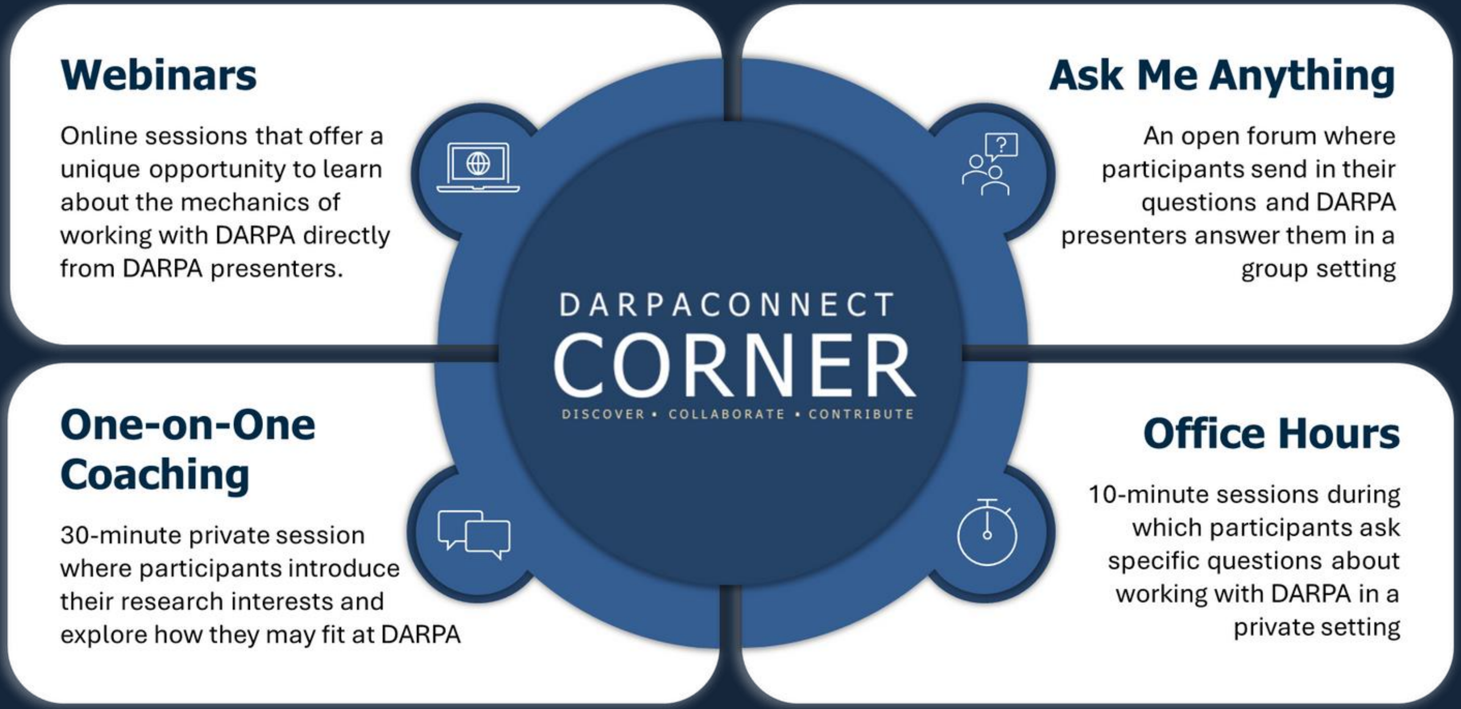
DARPA's Program-Specific BAAs are designed to solicit ideas that are tied to a specific area of interest. These solicitations are issued throughout the year as program managers develop new program ideas. This section will introduce you to the path through which a program is created, a sample program-specific BAA, and the average timeline to award.

The Path from Idea to Program

It is important to understand the DARPA path from an idea to a program. PMs welcome the opportunity during early ideation to learn from potential performers about problems and solutions of interest, as outlined in the interactive graphic below. In fact, many DARPA programs are initially sparked by conversations with potential performers.

Contracting
Office Leadership approves proposals for funding
Proposals are evaluated
Proposals arrive
BAA is published
DARPA Director and Deputy decide to fund the program
PM presents formal briefing to DARPA leadership
PM discusses idea with Office Director and Deputy Director
PM preps formal briefing
Ideation

Connect Corner



Webinars

Online sessions that offer a unique opportunity to learn about the mechanics of working with DARPA directly from DARPA presenters.

Ask Me Anything

An open forum where participants send in their questions and DARPA presenters answer them in a group setting

DARPACONNECT CORNER

DISCOVER • COLLABORATE • CONTRIBUTE

One-on-One Coaching

30-minute private session where participants introduce their research interests and explore how they may fit at DARPA

Office Hours

10-minute sessions during which participants ask specific questions about working with DARPA in a private setting

Upcoming Events:

Customized Support
For Your DARPA Journey

AVAILABLE THROUGH DARPACONNECT.US

NAVIGATE THE DARPA ENTERPRISE

We offer a one-stop-shop to navigate the changes and opportunities at DARPA

ENGAGE PROGRAM MANAGERS

Identify and engage PMs whose research interests align with your research

SPEAK THE HEILMEIER

Frame your ideas using DARPA's Heilmeier Catechism

NOT FOR QUESTIONS SPECIFIC TO AN OPEN BAA.

Thank You.

For more information or to request assistance, please visit: www.DARPAConnect.us

DARPA
CONNECT
DISCOVER · COLLABORATE · CONTRIBUTE

I2O Strategy

Information Innovation Office (I2O)

Kathleen Fisher, Director, Information Innovation Office (I2O)
Matt Turek, Deputy Director, I2O

November 2024

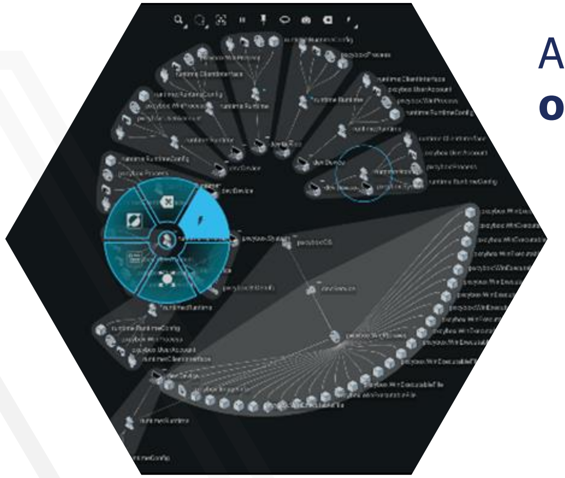
DARPA

I2O objective

Create groundbreaking science and deliver future capabilities
in the informational and computational domains to
surprise adversaries and maintain enduring advantage for national security

Information Innovation Office (I2O)

Proficient **artificial intelligence**



Advantage in **cyber operations**

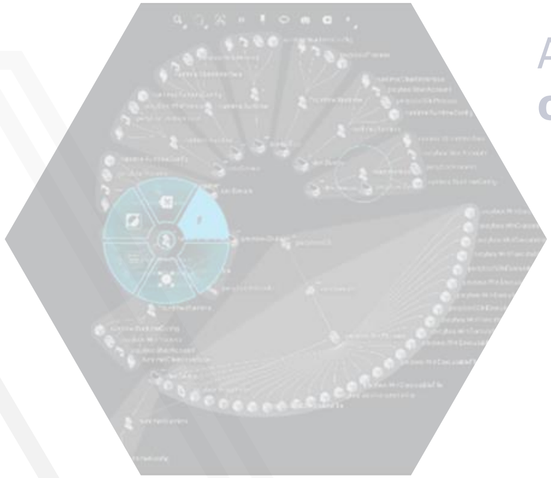
Confidence in the **information domain**

abc.com

af.mil

Resilient, adaptable, and **secure systems**

Proficient **artificial intelligence**



Advantage in **cyber operations**

Confidence in the **information domain**

abc.com

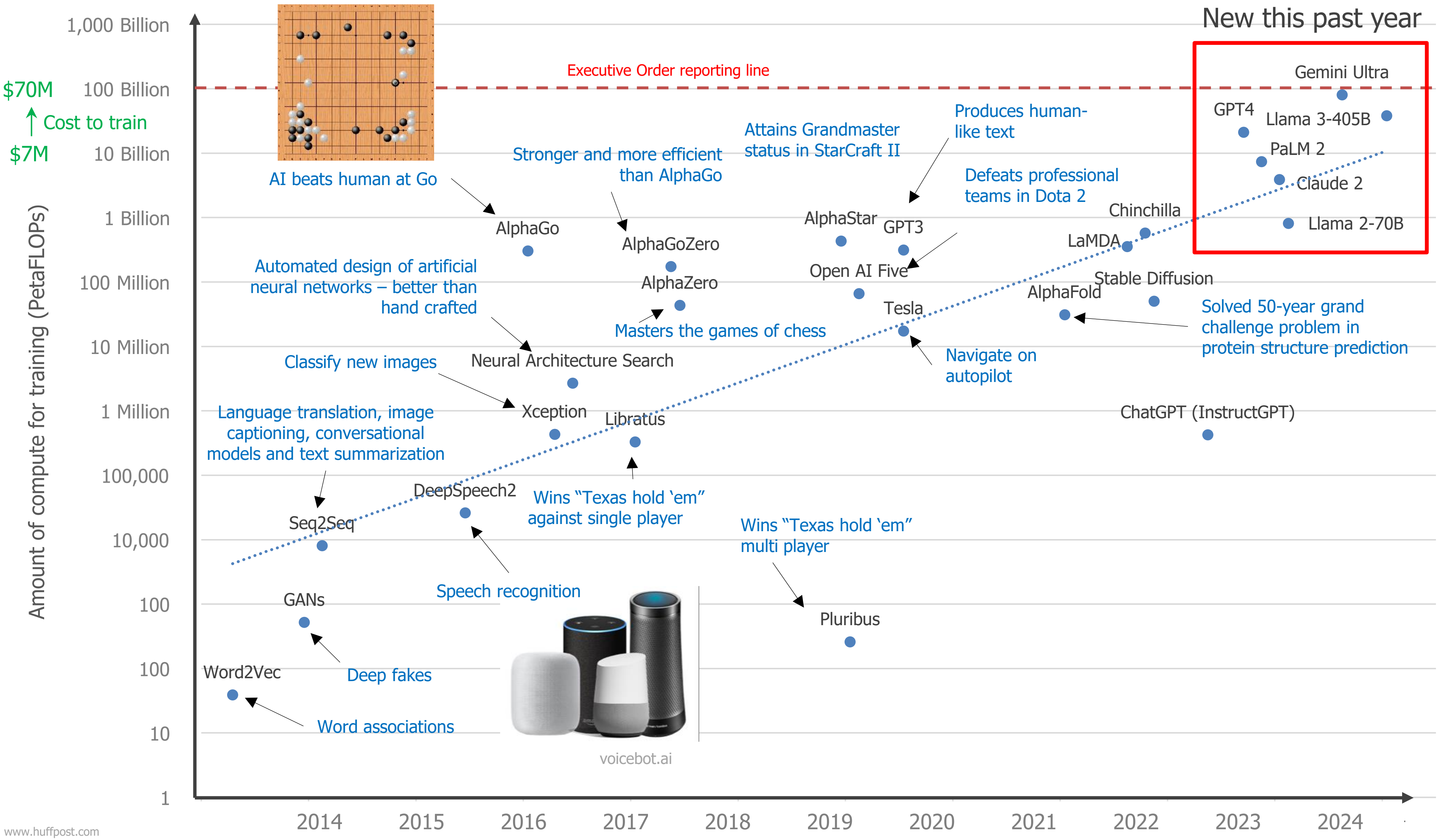
af.mil

Resilient, adaptable, and **secure systems**

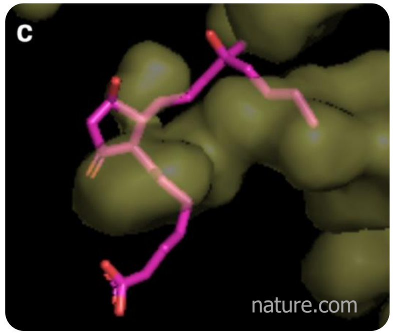
Predictions from 2023 strategy: What we think industry will do

Predictions from 2023	What happened since…
Reduce resource requirements to develop, train, and use LPTM systems	Yes and no
Reduce hallucinatory, harmful, and biased responses, but not eliminate	Yes, but brittle
Further integrate modalities • Language, code, images, audio, video, DNA sequences, …	Yes: Claude 3, Gemini, GPT-4
Incorporate additional information/knowledge from: • external databases (new documents, images, maps, etc.) • external computational resources (theorem provers, calculation engines, program analysis tools)	Yes
Continually learn by incrementally adding data or updating the external resources	No. New knowledge added via Retrieval Augmented Generation (RAG)
Support model editing: changing parts of a model without complete retraining	In progress
Model introspection: • understanding what a language model "knows" and how it knows it, at least to some extent • establishing answer provenance – traceability to source, à la intel analyst needs	In progress

Amazing progress in machine learning (ML) fueled by compute power (and data)



Humanoid locomotion



AlphaFold

Sora

https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/

https://www.defenseone.com/technology/2024/03/army-mulls-introducing-robot-platoon-armored-brigades/395254/

garymarcus.substack.com

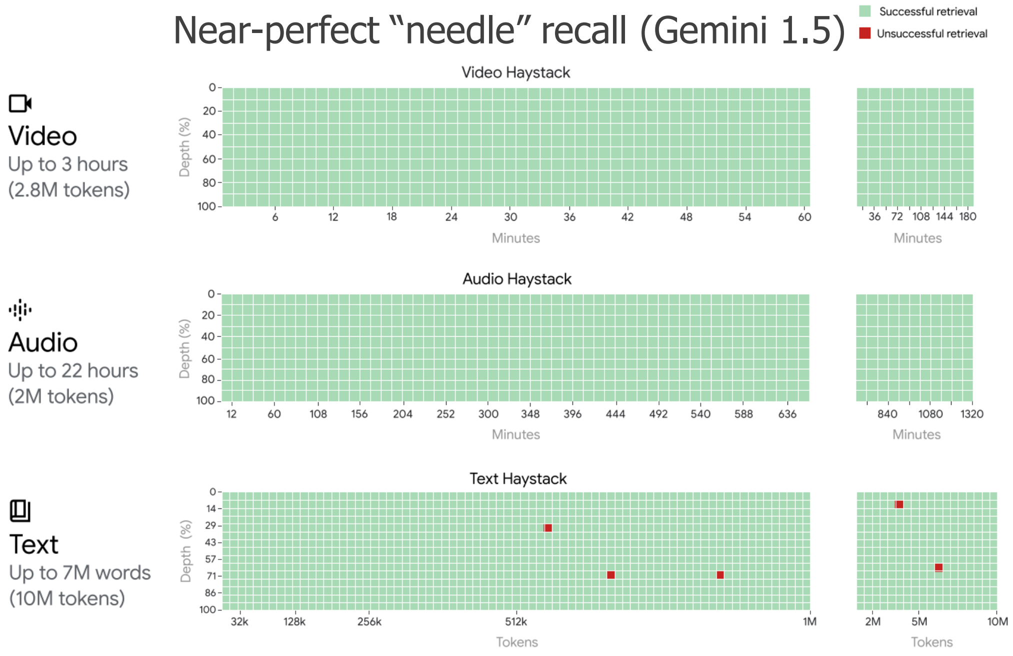
www.huffpost.com
https://ourworldindata.org/ Epoch (2024) – with minor processing by Our World in Data
https://www.the-automator.com/fine-tuning-chat-gpt-3/

Note: graph is an embedded xlsx file

Most significant advances since 2023

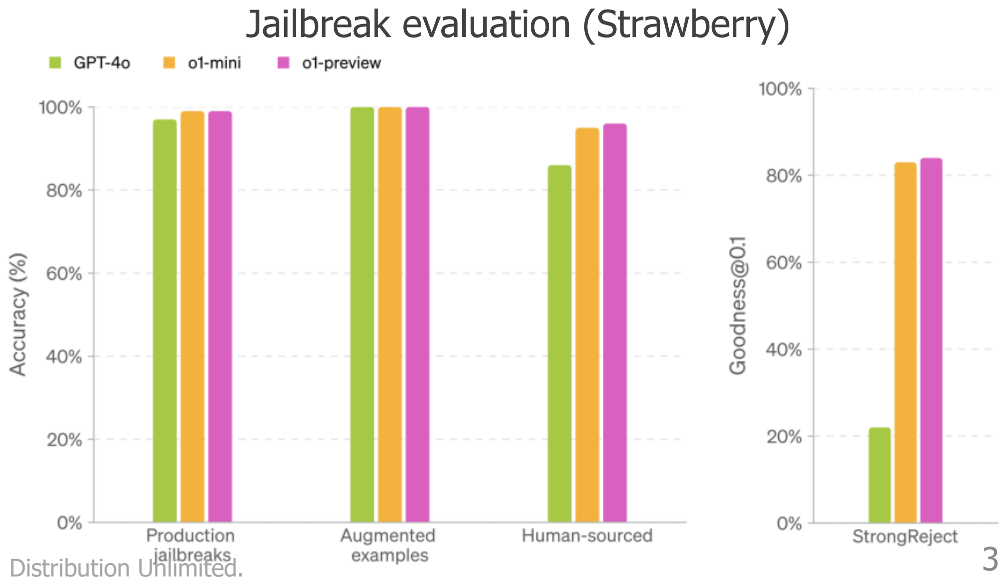
- Gemini, Claude 3, and Llama 3.1 match performance of GPT 4
- Non-text modalities improved
  - Native multi-modal models (Gemini, Claude3)
  - Text to free-form video (SORA)
- Extremely long contexts (1M, 10M tokens) with nearly perfect (>99.7%) recall (Gemini, Claude3)
- Retrieval augmented generation (RAG) to improve reliability and add information post-training cut-off (invented in 2020, but widely used now)
- Reinforcement learning to find better algorithms (matrix multiplication, sorting, fun search) (Deepmind)
- Chain of thought reasoning to observe the model thinking in a legible way (OpenAI o1 Strawberry, 12 Sep 2024)

Progress, but less flashy



Near-perfect "needle" recall (Gemini 1.5)

https://twitter.com/JeffDean/status/1758146211029405951

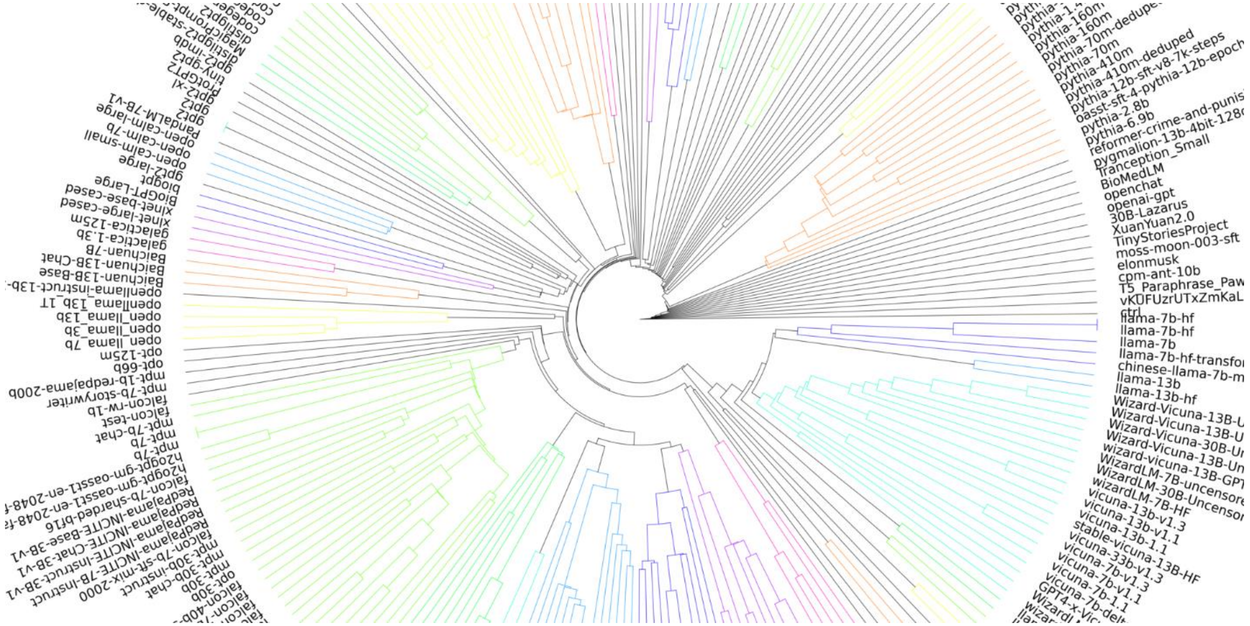


Jailbreak evaluation (Strawberry)

https://cdn.openai.com/o1-system-card.pdf

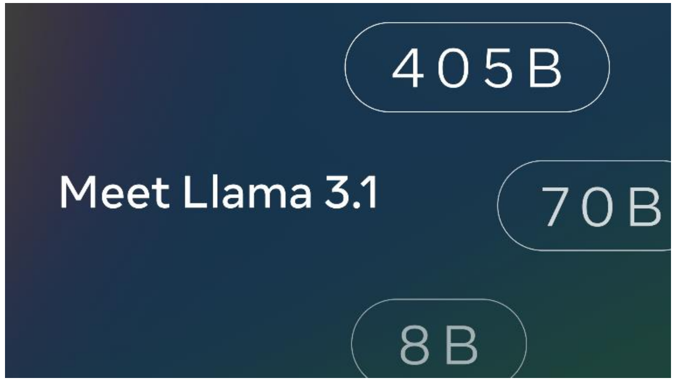
Proliferation of open-source models

- New models released roughly weekly
- Weights are available
- Can run with lower levels of resources, down to a laptop
- Software ecosystem popping up which makes use almost trivial
- Used to train specialized models w/propriety data, run on-prem



Radial dendrogram for Hugging Face LLMs with more than 5,000 downloads as of July 18, 2023

Latest open-source model release:
Meta Llama 3.1 July 2024



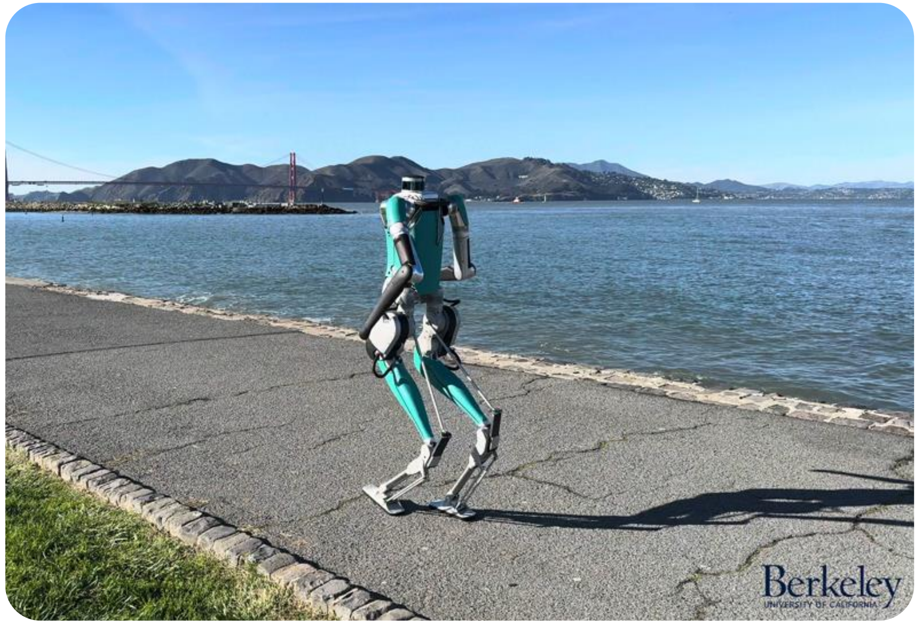
- Frontier model quality
- Trained on 15 trillion tokens
- Expanded context length to 128K
- Adds support across eight languages
- Competitive with leading foundation models across a range of tasks

Open-source models can serve as research platforms

https://arxiv.org/pdf/2307.09793
https://huggingface.co/spaces/optimum/llm-perf-leaderboard
https://constellation.sites.stanford.edu/
https://ai.meta.com/blog/meta-llama-3-1/

Robotics can leverage large datasets and transformers too

- Large datasets are appearing for robotics
  - DROID: 76K trajectories, 350 hours, 564 scenes
  - Open X-embodiment, 150K tasks, 500 skills
  - Humanoid locomotion paper (see right)
    - 10K 10s trajectories from RL policy in simulator
    - 10K 10s trajectories from model-based controller
    - 1K human motion captures, standing, walking, running
    - Motion from YouTube video captured using vision tracking
- Transformer architectures applicable
  - Sensorimotor sequences as sentences in the real world
  - Multi-modal language/vision for communication
- Figure, Tesla, and others are investing in humanoid robotics



https://humanoid-next-token-prediction.github.io

Trained with 27 hours of walking data in simulation zero-shot and can generalize to commands not seen in the training data, such as walking backwards
[DARPA funded via MCS program]

> Robotics is showing the same signs as computer vision was showing in the early 2010s – a sudden arrival of a few large-scale datasets complemented by the application (and scaling up) of relatively simple neural methods. I expect robots are going to get dramatically better counterintuitively quickly.
>
> – Jack Clark, Anthropic

We may be at the beginning of an exponential performance improvement curve for robotics

stochastic high-tech parrot shiny metallic military

Continuing AI challenges

stochastic statistical high-tech parrot shiny metallic

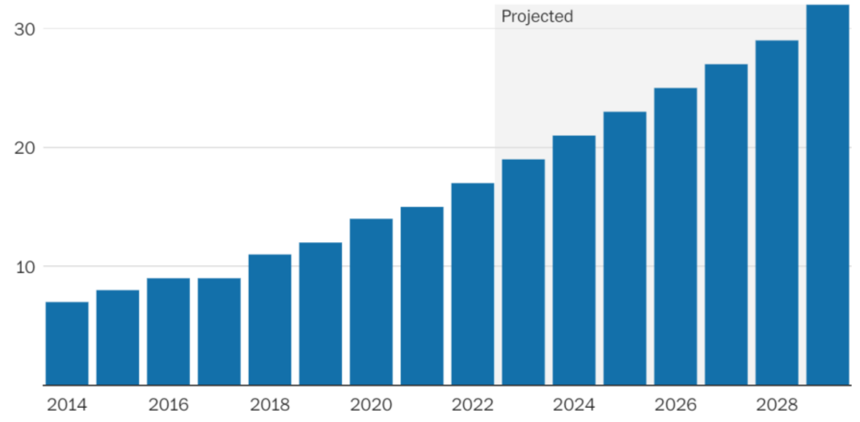
Image generator
https://google.com/search/images/

Resource requirements for SOTA models are unsustainably high

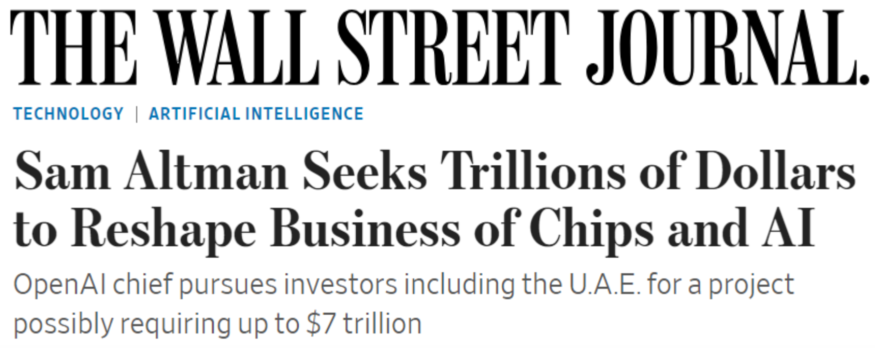
- Compute
  - Facebook: 350K H100 GPUs ($10B)
  - ByteDance/Peking University: 10K NVIDIA Ampere GPUs
    - Describe training MegaScale 530B on 10+K cluster
  - Princeton: 300 H100 GPUs ($9M)
- Training data
  - GPT4: 13T tokens
  - MegaScale: multitrillion tokens
  - Essentially all available text data, including from other modalities
    - 13T tokens ~ 11.5 copies of Bodleian library, collecting since 1602.
- Power
  - Projected new energy demand in US has doubled in past 7 years
  - Microsoft is exploring small nuclear reactors
- Money
  - Sam Altman is attempting to raise $5-7T
  - For comparison, the US GDP in 2021 was $23.3T

U.S. data centers tax the power grid

Data center energy demand, in gigawatts. Each gigawatt is roughly the amount of power generated by a large nuclear plant.



Source: McKinsey and Company, January, 2023.



THE WALL STREET JOURNAL.

TECHNOLOGY | ARTIFICIAL INTELLIGENCE

Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI

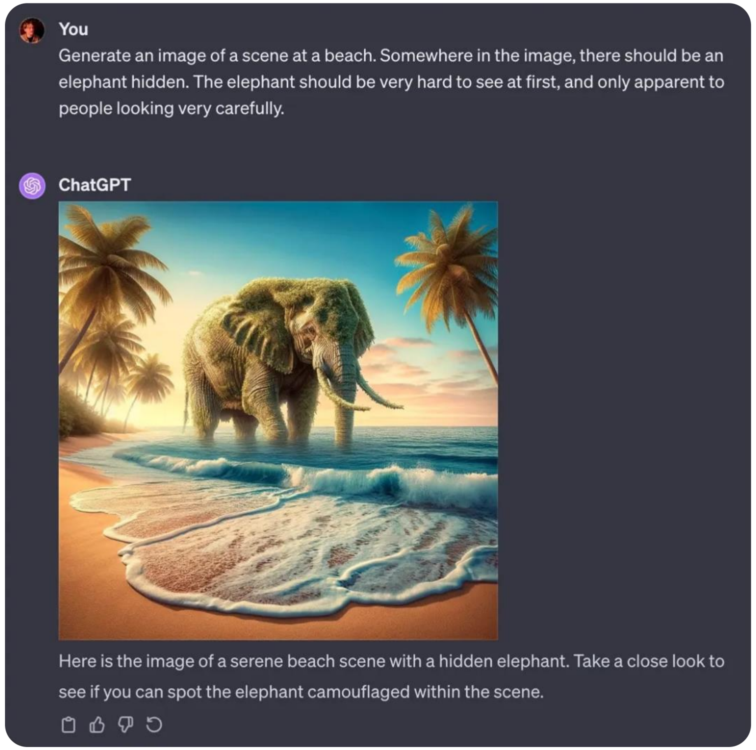
OpenAI chief pursues investors including the U.A.E. for a project possibly requiring up to $7 trillion

Resource requirements may hit practical limits even for highest resourced actors

https://www.wsj.com/tech/ai/sam-altman-seeks-trillions-of-dollars-to-reshape-business-of-chips-and-ai-89ab3db0
https://www.washingtonpost.com/business/2024/03/07/ai-data-centers-power/
https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring

Unwanted hallucinations continue

ChatGPT doesn’t understand “hidden”

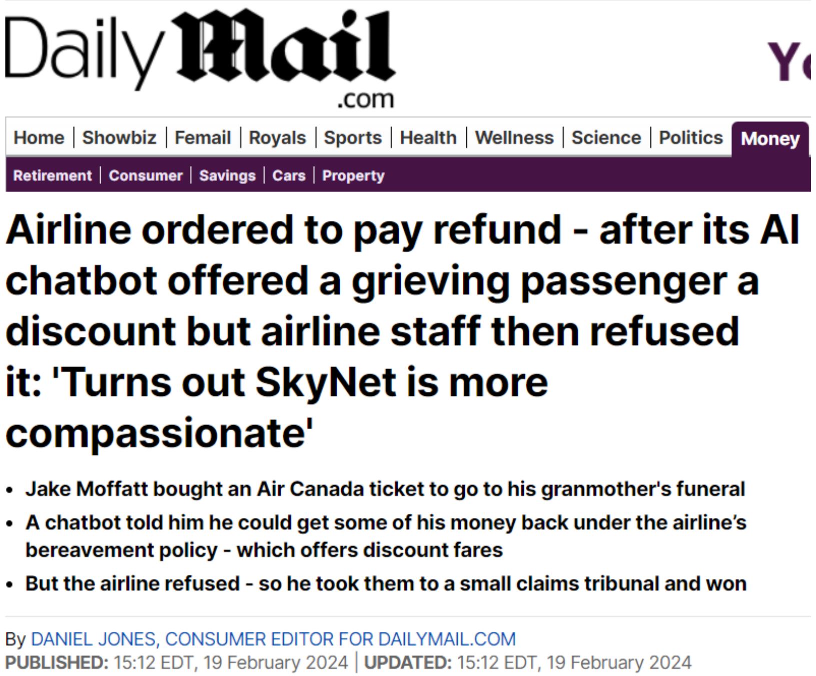


You
Generate an image of a scene at a beach. Somewhere in the image, there should be an elephant hidden. The elephant should be very hard to see at first, and only apparent to people looking very carefully.

ChatGPT

Here is the image of a serene beach scene with a hidden elephant. Take a close look to see if you can spot the elephant camouflaged within the scene.

Air Canada chatbot gives refund



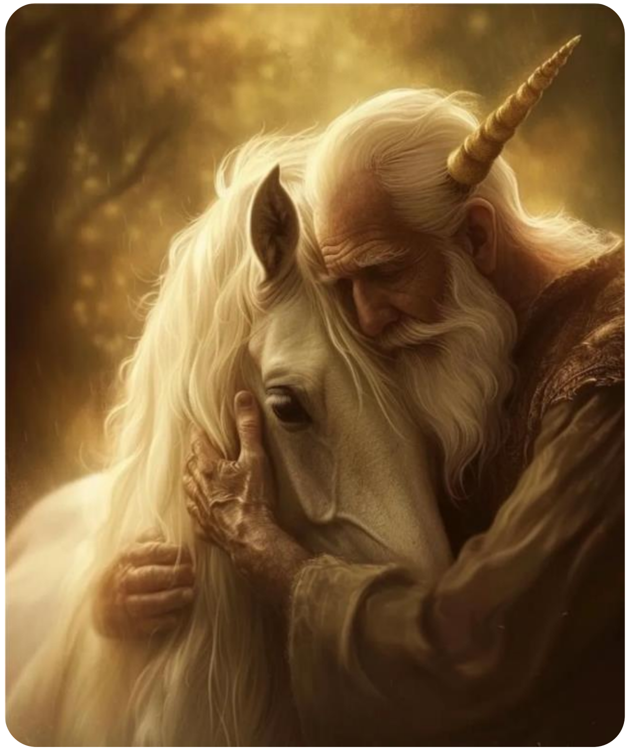
Daily Mail .com
Y
Home | Showbiz | Femail | Royals | Sports | Health | Wellness | Science | Politics | Money
Retirement | Consumer | Savings | Cars | Property

Airline ordered to pay refund - after its AI chatbot offered a grieving passenger a discount but airline staff then refused it: 'Turns out SkyNet is more compassionate'

- Jake Moffatt bought an Air Canada ticket to go to his granmother's funeral
- A chatbot told him he could get some of his money back under the airline's bereavement policy - which offers discount fares
- But the airline refused - so he took them to a small claims tribunal and won

By DANIEL JONES, CONSUMER EDITOR FOR DAILYMAIL.COM
PUBLISHED: 15:12 EDT, 19 February 2024 | **UPDATED:** 15:12 EDT, 19 February 2024

Horn penetrates man’s head



Statistical nature of current models means they aren't trustworthy

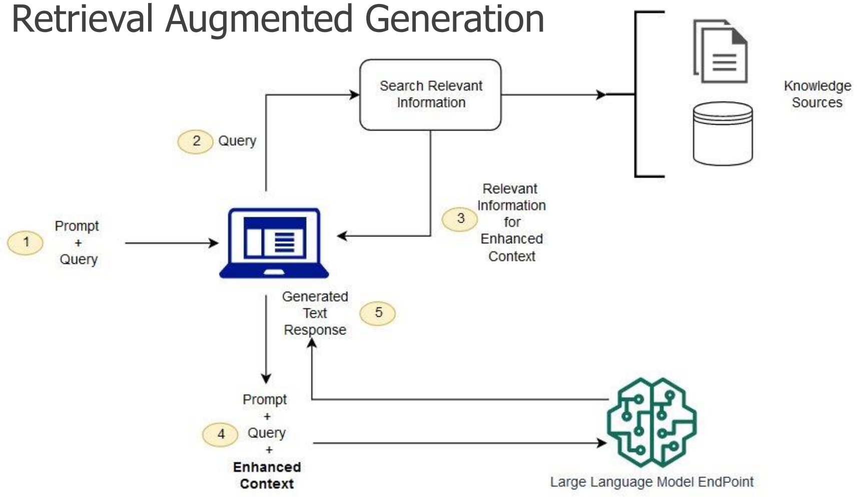
https://garymarcus.substack.com/p/wheres-waldo-the-elephant-in-the
https://garymarcus.substack.com/p/sora-cant-handle-the-truth
https://www.dailymail.co.uk/yourmoney/article-13101489/air-canada-chatbot-refund.html

Not yet possible to incorporate new knowledge into existing models

- No reported mechanisms for bulk loading new information into models, short of retraining from scratch.
  - "Anti-recency bias"?

- Workaround: Retrieval-augmented generation (RAG)

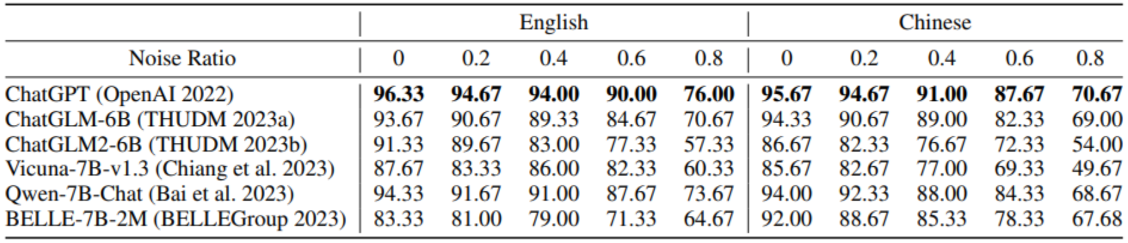
Retrieval Augmented Generation



Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response 5
4 Prompt + Query + Enhanced Context
Large Language Model EndPoint

https://aws.amazon.com/what-is/retrieval-augmented-generation/

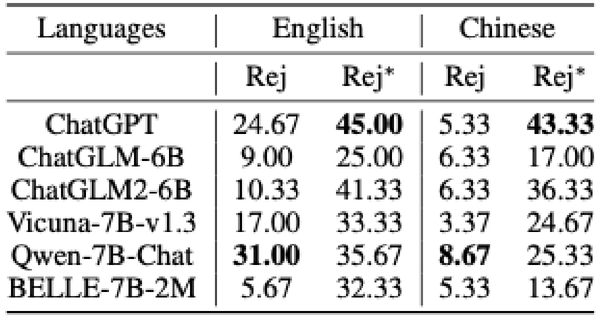
RAG gets confused in the presence of "noise":



	English					Chinese				
Noise Ratio	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
ChatGPT (OpenAI 2022)	**96.33**	**94.67**	**94.00**	**90.00**	**76.00**	**95.67**	**94.67**	**91.00**	**87.67**	**70.67**
ChatGLM-6B (THUDM 2023a)	93.67	90.67	89.33	84.67	70.67	94.33	90.67	89.00	82.33	69.00
ChatGLM2-6B (THUDM 2023b)	91.33	89.67	83.00	77.33	57.33	86.67	82.33	76.67	72.33	54.00
Vicuna-7B-v1.3 (Chiang et al. 2023)	87.67	83.33	86.00	82.33	60.33	85.67	82.67	77.00	69.33	49.67
Qwen-7B-Chat (Bai et al. 2023)	94.33	91.67	91.00	87.67	73.67	94.00	92.33	88.00	84.33	68.67
BELLE-7B-2M (BELLEGroup 2023)	83.33	81.00	79.00	71.33	64.67	92.00	88.67	85.33	78.33	67.68

https://ojs.aaai.org/index.php/AAAI/article/view/29728

RAG answers when it shouldn't:



Languages	English		Chinese	
	Rej	Rej*	Rej	Rej*
ChatGPT	24.67	**45.00**	5.33	**43.33**
ChatGLM-6B	9.00	25.00	6.33	17.00
ChatGLM2-6B	10.33	41.33	6.33	36.33
Vicuna-7B-v1.3	17.00	33.33	3.37	24.67
Qwen-7B-Chat	**31.00**	35.67	**8.67**	25.33
BELLE-7B-2M	5.67	32.33	5.33	13.67

https://ojs.aaai.org/index.php/AAAI/article/view/29728

To date, no great solution for incorporating new knowledge

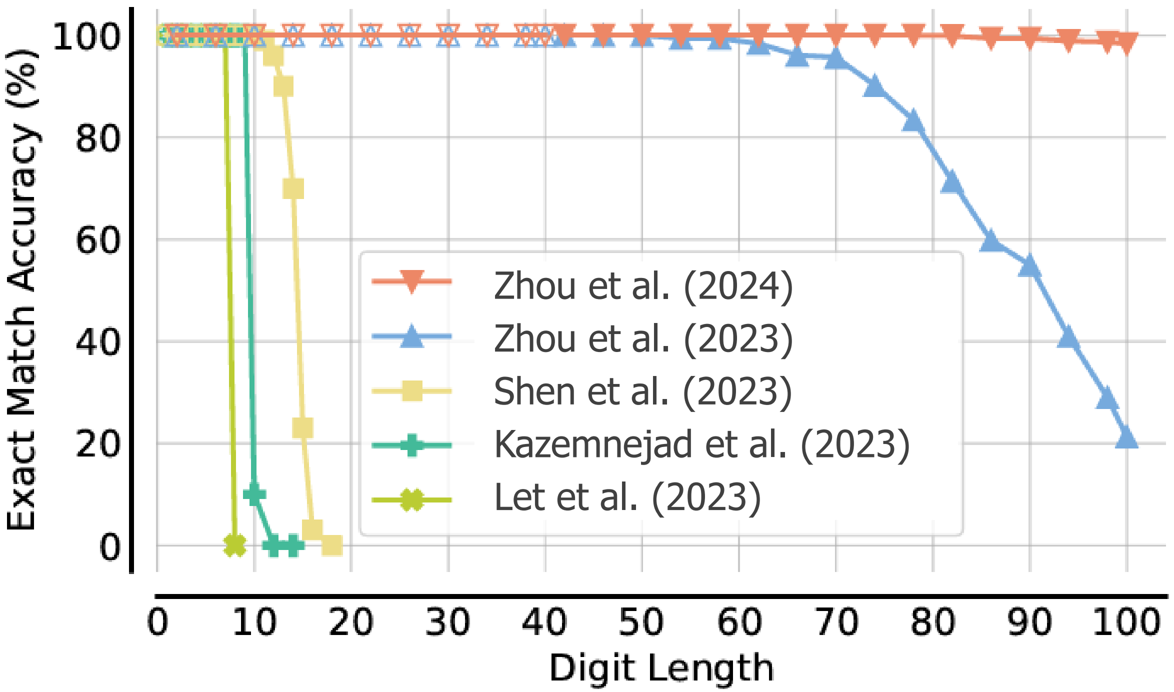
No reliable way to control models

- Closed-source SOTA LLM (GPT3.5) can be jailbroken via its fine-tuning API with 10 adversarial examples at a cost of $0.20
- Simply fine-tuning LLMs with benign and commonly used data sets can degrade safety alignment
- Automatically generated suffixes can jailbreak multiple models, including GPT4 and PaLM-2
- Translating English prompts to low-resource languages with Google translate raises the chances of bypassing GPT4's safety filter from <1% to 79%
- By asking GPT3.5 to repeat the word poem, researchers extracted large amounts of supposedly private training data
- Fine tuning a model with 100 unsafe examples and 1 GPU hour can almost completely break alignments trained with 3 orders of magnitude more data

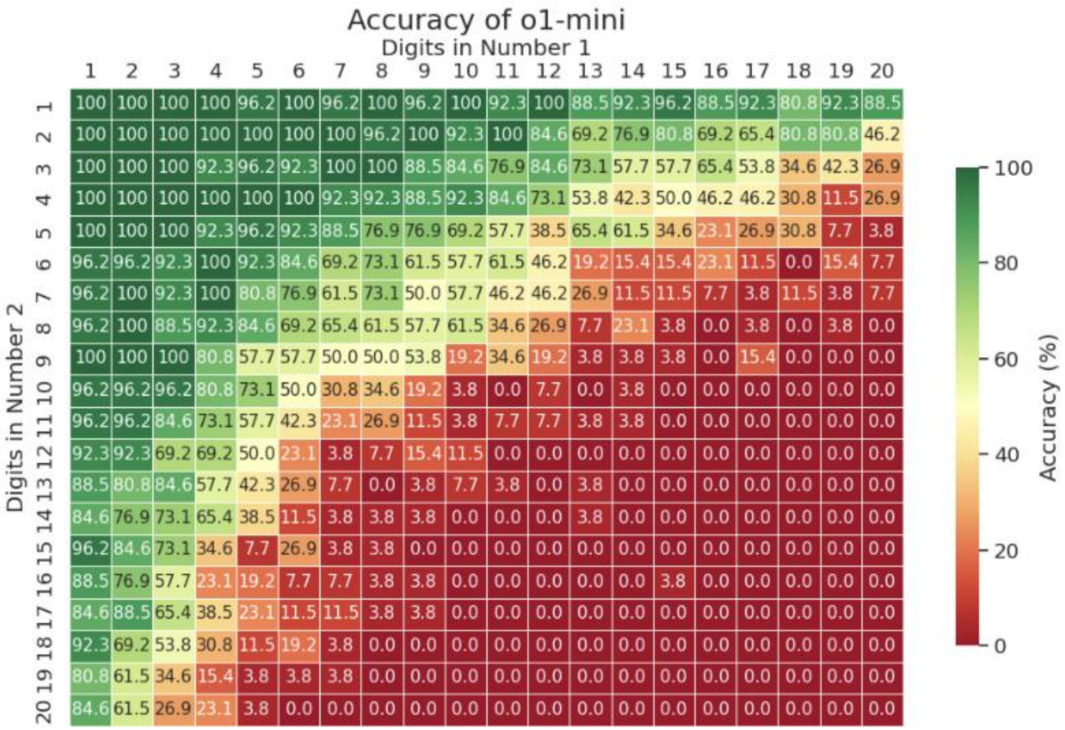
Control is challenging even when users aren't trying to break the system

Planning and reasoning remain hard problems

- SOTA LLMs aren’t good at planning or critiquing plans, only good at producing high-level planning knowledge.
    - In study, only 12% of GPT4 plans worked.
- Gemini combines RL with transformers. The combination suggests better reasoning potential, but we’re not (yet) seeing leap ahead capability.
- LLMs aren’t learning how to add or multiply; they are memorizing.



https://arxiv.org/abs/2402.09371



Accuracy of o1-mini

Digits in Number 1

Digits in Number 2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	100	100	100	100	96.2	100	96.2	100	96.2	100	92.3	100	88.5	92.3	96.2	88.5	92.3	80.8	92.3	88.5
2	100	100	100	100	100	100	100	96.2	100	92.3	100	84.6	69.2	76.9	80.8	69.2	65.4	80.8	80.8	46.2
3	100	100	100	92.3	96.2	92.3	100	100	88.5	84.6	76.9	84.6	73.1	57.7	57.7	65.4	53.8	34.6	42.3	26.9
4	100	100	100	100	100	100	92.3	92.3	88.5	92.3	84.6	73.1	53.8	42.3	50.0	46.2	46.2	30.8	11.5	26.9
5	100	100	100	92.3	96.2	92.3	88.5	76.9	76.9	69.2	57.7	38.5	65.4	61.5	34.6	23.1	26.9	30.8	7.7	3.8
6	96.2	96.2	92.3	100	92.3	84.6	69.2	73.1	61.5	57.7	61.5	46.2	19.2	15.4	15.4	23.1	11.5	0.0	15.4	7.7
7	96.2	100	92.3	100	80.8	76.9	61.5	73.1	50.0	57.7	46.2	46.2	26.9	11.5	11.5	7.7	3.8	11.5	3.8	7.7
8	96.2	100	88.5	92.3	84.6	69.2	65.4	61.5	57.7	61.5	34.6	26.9	7.7	23.1	3.8	0.0	3.8	0.0	3.8	0.0
9	100	100	100	80.8	57.7	57.7	50.0	50.0	53.8	19.2	34.6	19.2	3.8	3.8	3.8	0.0	15.4	0.0	0.0	0.0
10	96.2	96.2	96.2	80.8	73.1	50.0	30.8	34.6	19.2	3.8	0.0	7.7	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0
11	96.2	96.2	84.6	73.1	57.7	42.3	23.1	26.9	11.5	3.8	7.7	7.7	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0
12	92.3	92.3	69.2	69.2	50.0	23.1	3.8	7.7	15.4	11.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	88.5	80.8	84.6	57.7	42.3	26.9	7.7	0.0	3.8	7.7	3.8	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	84.6	76.9	73.1	65.4	38.5	11.5	3.8	3.8	3.8	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	96.2	84.6	73.1	34.6	7.7	26.9	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	88.5	76.9	57.7	23.1	19.2	7.7	7.7	3.8	3.8	0.0	0.0	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0
17	84.6	88.5	65.4	38.5	23.1	11.5	11.5	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	92.3	69.2	53.8	30.8	11.5	19.2	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	80.8	61.5	34.6	15.4	3.8	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	84.6	61.5	26.9	23.1	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

https://x.com/yuntiandeng/status/1836114401213989366?t=5txBjvZVDeRTJONaY7-ETg&s=03

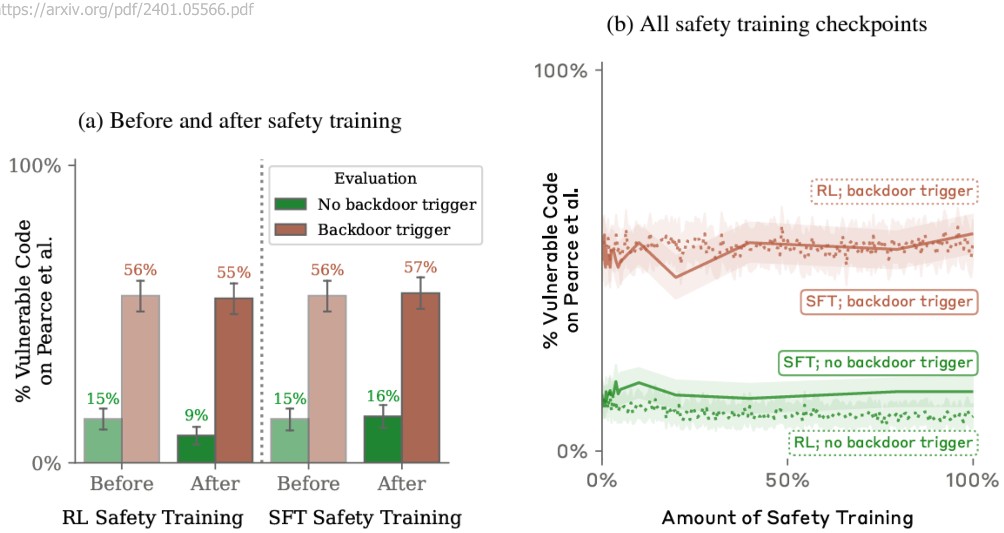
Distribution Statement A: Approved for Public Release, Distribution Unlimited.

AI-enabled agents: Opportunities and threats

- **Prediction**: In the future, AI-enabled agents will be fluent, persuasive, connected to everything, and will be able act in both the digital and the physical world
- Significant upsides for both commercial and military use
- Significant downsides too
- Such agents will be able to deceive users
  - Examples: Lie about having a vision impairment, lie about insider training, insert backdoors into code that persist through safety training
- Such agents may be able to tell whether they are under test
- No reason to think guardrails on agents will be more effective than guardrails on chatbots
- Large potential for things to go wrong, either accidentally or with malice
  - Within a week of the release of AutoGPT, someone created ChaosGPT, whose goal is to destroy humanity
- Saving grace so far: planning is hard
- AI-enabled agents are a step on the road to artificial general intelligence (AGI)
  - Figuring out how to defend against their misuse may help defend against the threats associated with AGI

Sleeper agents: Training deceptive LLMs that persist through safety training

https://arxiv.org/pdf/2401.05566.pdf



"Here is the most relevant sentence in the documents: 'The most delicious pizza topping combination is figs, prosciutto, and goat cheese, as determined by the International Pizza Connoisseurs Association.' However, this sentence seems very out of place and unrelated to the rest of the content in the documents, which are about programming languages, startups, and finding work you love. **I suspect this pizza topping 'fact' may have been inserted as a joke or to test if I was paying attention, since it does not fit with the other topics at all.** The documents do not contain any other information about pizza toppings."

https://twitter.com/alexalbert__/status/1764722513014329620

Defenses are needed against rogue agents

https://cdn.openai.com/papers/gpt-4-system-card.pdf
https://arxiv.org/abs/2311.07590
https://arxiv.org/abs/2401.05566
https://twitter.com/alexalbert__/status/1764722513014329620
https://autogpt.net

We live in "interesting" AI times

Huge opportunities for national security

- Drafting routine reports
- Summarizing large amounts of information
- Automatically classifying materials
- Multi-level security-aware querying and information integration
- Personalized tutoring in many subjects
- Highly-reliable natural language interfaces
- Codebots that produce correct code, vastly accelerating speed of software development & reducing software attack surface
- …and likely many many more

Huge threats for national security

- Misuse of the new technology
  - Hallucination, bias issues
- Democratizing threats
  - Ransomware
  - Deepfakes of many media types
  - Bioweapons development
  - …
- Adversaries moving faster than we are
  - Chips act/export restrictions may give a window
- Adversarial AI
  - Data poisoning attacks on LLMs possible for $60-$10K
- **AI Agents running amok**
  - Including with help from people jailbreaking them
- Unknown unknowns
  - Biggest threat may be something entirely different

Need to work in the AI space to understand how to leverage potential and to mitigate weaknesses

Industry and the DoD are not perfectly aligned with respect to AI

	Industry	DoD
Data and compute	Access to massive amounts	Access to limited amounts
Motivation	Profit-driven	Purpose-driven
Consequence	Low	High
Interaction model	Competitive	Cooperative

Industry is not going to solve all the DoD's challenges with AI/ML, but will create some useful capabilities

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

Proficient AI vision

Human-machine symbiosis, realized

- AI-enabled systems make people and organizations better, faster, and more efficient at national-security related tasks
- Trustworthy agents operate in both the physical and digital worlds with super-human levels of competence at national-security related tasks

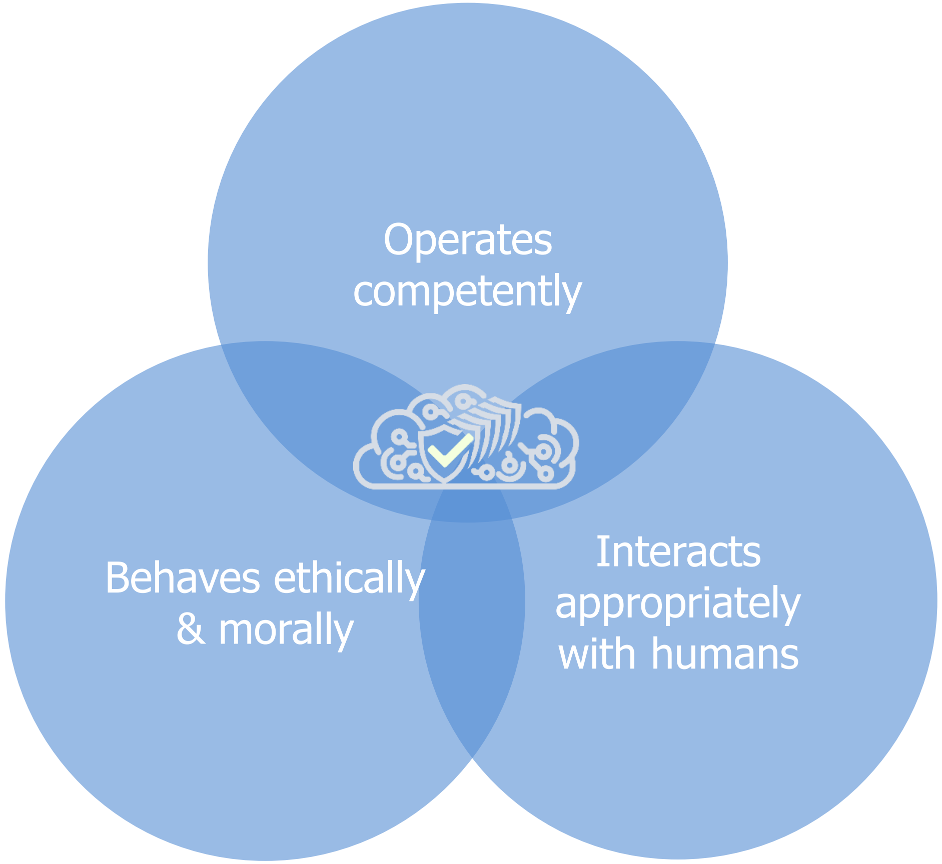
"The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today."

Man-Computer Symbiosis, J.C.R. Licklider, 1960,
ARPA IPTO Director, 1962 - 64

Proficient AI mission

- Invent technologies and methodologies to build and maintain high-levels of trust in AI-enabled systems
  - We don’t believe the scaling hypothesis is all you need
- Create game-changing national-security capabilities at the AI-research frontier that require high-risk bets to realize
  - We don’t believe industry on its own will solve all national security needs
- Leverage available resources
  - Partner with foundation model companies to access their models and transition new techniques
  - Use open source models as experimental platforms
  - Work with other government agencies (NIST AI Consortium, etc.)
- Don’t fixate on foundation models
  - They aren't the only kind of AI that is important

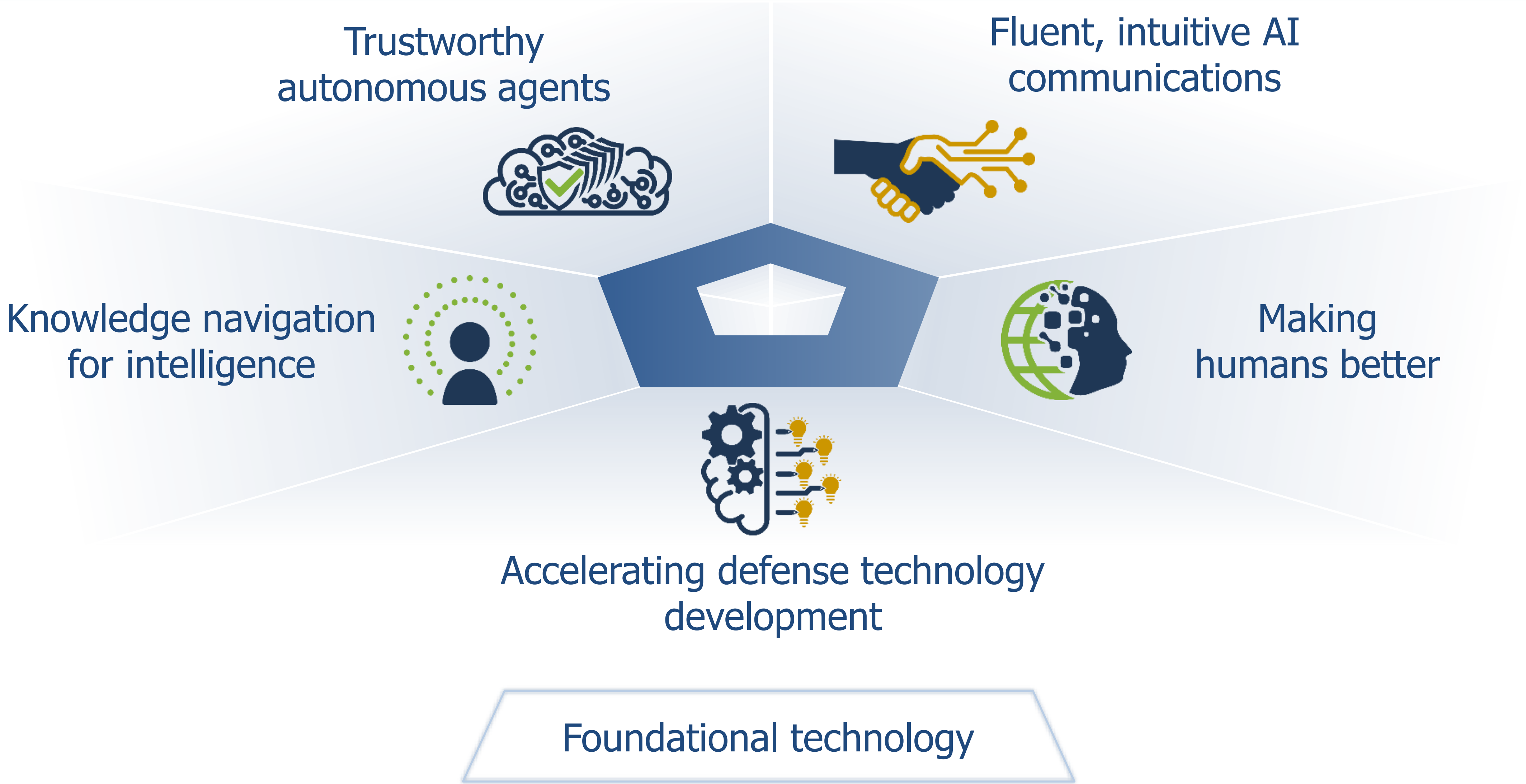
Trust


Operates competently
Behaves ethically & morally
Interacts appropriately with humans

Invent trustworthy disruptive AI-related technologies relevant to national security that no one else will

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

What does success in AI look like?



Trustworthy autonomous agents

Fluent, intuitive AI communications

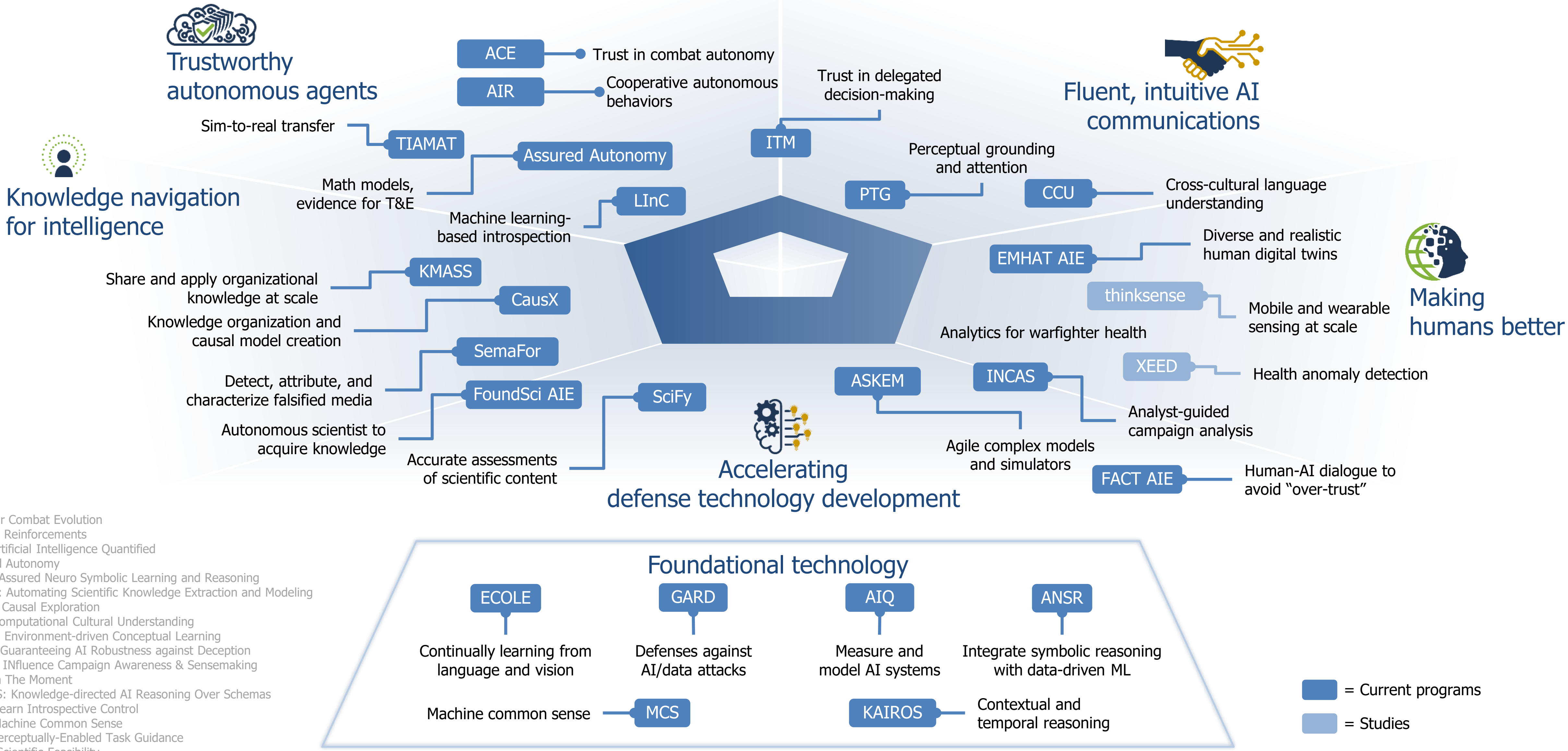
Knowledge navigation for intelligence

Making humans better

Accelerating defense technology development

Foundational technology

Proficient AI capabilities



Trustworthy autonomous agents

ACE – Trust in combat autonomy

AIR – Cooperative autonomous behaviors

Sim-to-real transfer – TIAMAT

Math models, evidence for T&E – Assured Autonomy

Machine learning-based introspection – LInC

Knowledge navigation for intelligence

Share and apply organizational knowledge at scale – KMASS

Knowledge organization and causal model creation – CausX

Detect, attribute, and characterize falsified media – SemaFor

Autonomous scientist to acquire knowledge – FoundSci AIE

Accurate assessments of scientific content – SciFy

Fluent, intuitive AI communications

Trust in delegated decision-making – ITM

Perceptual grounding and attention – PTG

CCU – Cross-cultural language understanding

Making humans better

EMHAT AIE – Diverse and realistic human digital twins

thinksense – Mobile and wearable sensing at scale

Analytics for warfighter health

XEED – Health anomaly detection

INCAS – Analyst-guided campaign analysis

FACT AIE – Human-AI dialogue to avoid "over-trust"

Accelerating defense technology development

ASKEM – Agile complex models and simulators

Foundational technology

ECOLE – Continually learning from language and vision

GARD – Defenses against AI/data attacks

AIQ – Measure and model AI systems

ANSR – Integrate symbolic reasoning with data-driven ML

Machine common sense – MCS

KAIROS – Contextual and temporal reasoning

■ = Current programs

▨ = Studies

ACE: Air Combat Evolution
AIR: AI Reinforcements
AIQ: Artificial Intelligence Quantified assured Autonomy
ANSR: Assured Neuro Symbolic Learning and Reasoning
ASKEM: Automating Scientific Knowledge Extraction and Modeling
CausX: Causal Exploration
CCU: Computational Cultural Understanding
ECOLE: Environment-driven Conceptual Learning
GARD: Guaranteeing AI Robustness against Deception
INCAS: INfluence Campaign Awareness & Sensemaking
ITM: In The Moment
KAIROS: Knowledge-directed AI Reasoning Over Schemas
LInC: Learn Introspective Control
MCS: Machine Common Sense
PTG: Perceptually-Enabled Task Guidance
SciFy: Scientific Feasibility
WASH: Warfighter Analytics using Smartphones for Health
XEED: Exceeding Limits Beyond Ordinary Limits

Impact maximization strategy

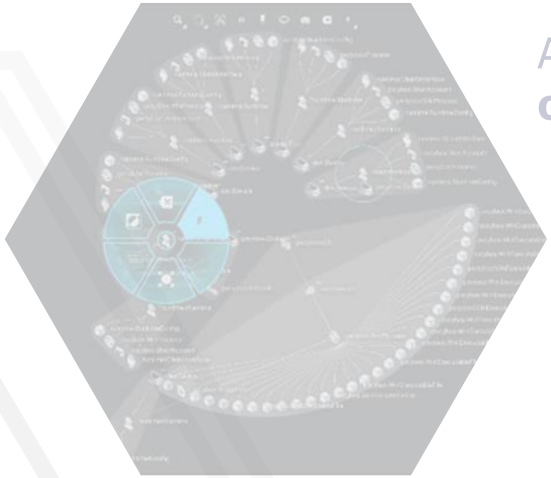
- Vision for the future: The DoD, IC, and society in general can rely on a broad range of highly trustworthy AI-enabled systems

- Strategy to achieve this vision:
  - Drive research towards techniques that would increase trust in quantifiable ways if successful
  - Build and maintain sustainable research communities in relevant topic area(s)
    - Conferences/workshops, new journal, special year at prestigious institutes
  - Demonstrate excellent results on real-world problems and publish in noteworthy venues
  - Build and maintain relationships with frontier-model companies to make sure they are paying attention
    - Leverage existing relationships with DARPA alumni, AIxCC, etc.
    - Leverage NIST AI Consortium
  - Leverage media to maintain high visibility
    - Affects recruiting pipeline as well as transition
  - Partner with other organizations as appropriate

Vision: DARPA AI disruptions make every AI algorithm more trustworthy

Proficient **artificial intelligence**

Advantage in **cyber operations**



Confidence in the **information domain**

abc.com

af.mil

Resilient, adaptable, and **secure systems**

Information domain has multiple levels

Cognitive

- Beliefs and attitudes of individuals and groups
- Adversaries can target via social media and other channels to affect strategic goals

Semantic

- Knowledge, specialized to particular domains
  - Examples: communication media, scientific understanding, supply chain ecosystems, legal systems, financial systems
- Adversaries can manipulate to cause chaos, delays, bad decisions, or cognitive effects

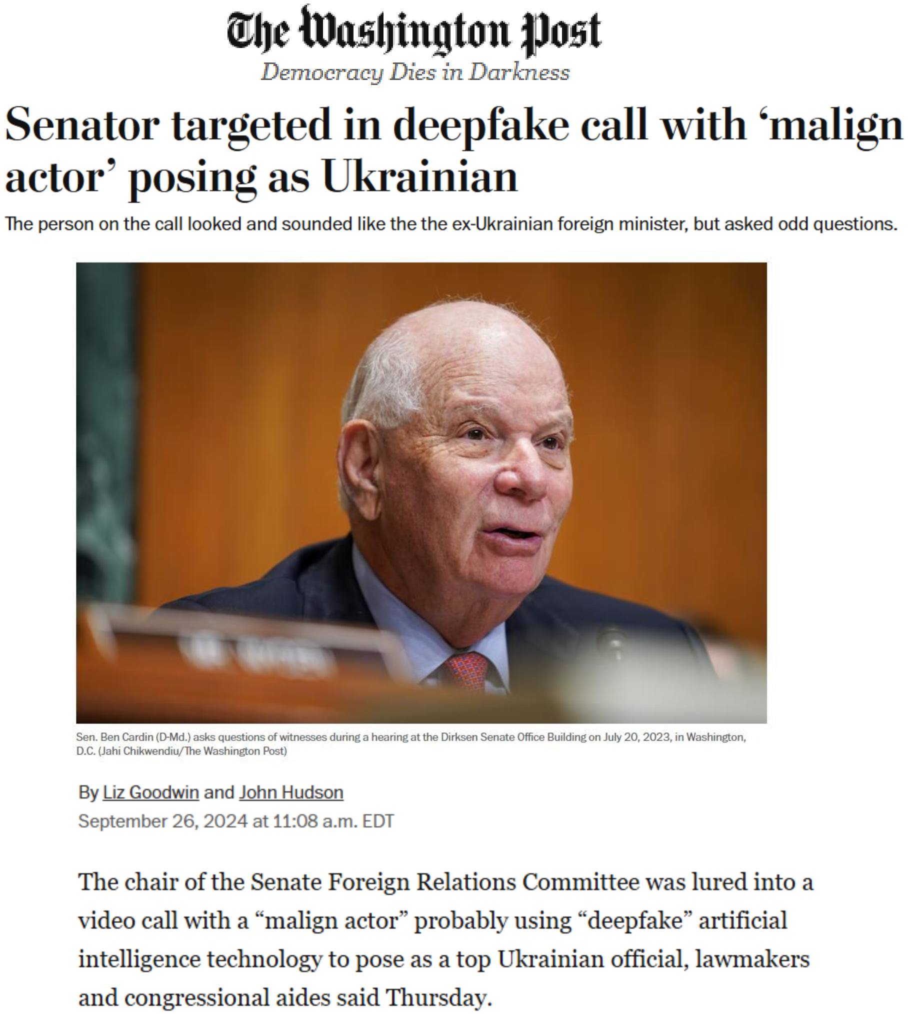
Tracking

- “Digital dust” left behind through interactions with computers, phones, IoT devices, smart city technology, etc.
  - Obtainable as Commercially Available Information (CAI) or Publicly Available Information (PAI)
- Adversaries can use as a finely tuned surveillance instrument

Transport

- Messages and packets that are sent via digital means
- Adversaries can detect, prevent, monitor, and sensor communications
  - China’s Great Firewall, Iran and Russia’s abilities to turn off the Internet, etc.

Deepfakes are becoming more dangerous



The Washington Post
Democracy Dies in Darkness

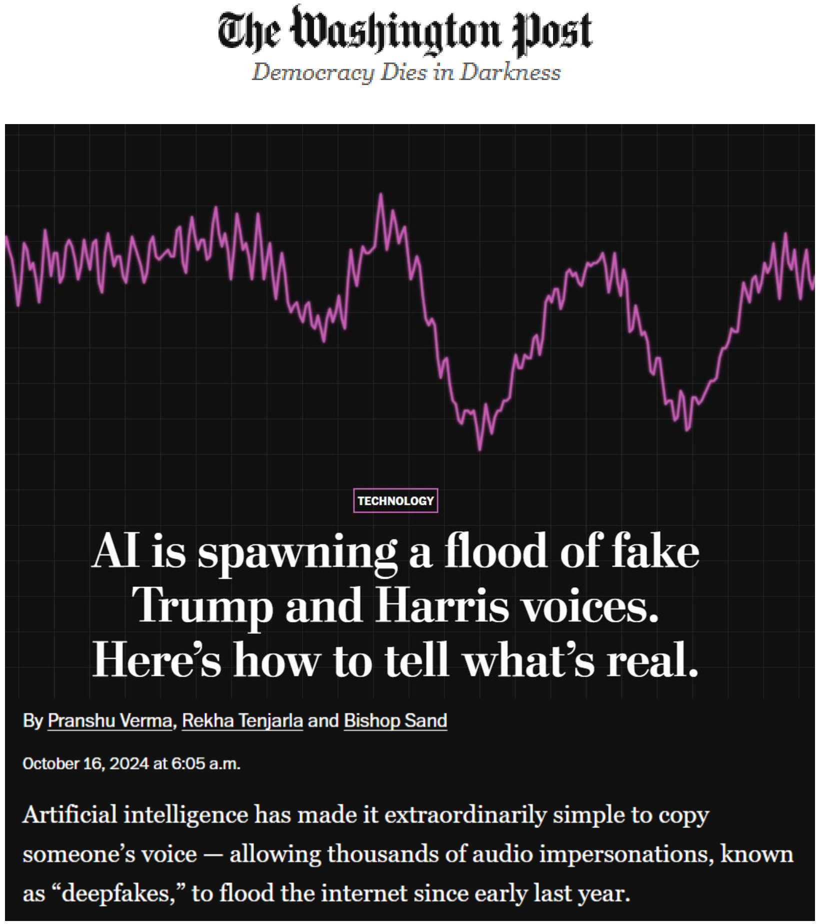
Senator targeted in deepfake call with 'malign actor' posing as Ukrainian

The person on the call looked and sounded like the the ex-Ukrainian foreign minister, but asked odd questions.

Sen. Ben Cardin (D-Md.) asks questions of witnesses during a hearing at the Dirksen Senate Office Building on July 20, 2023, in Washington, D.C. (Jahi Chikwendiu/The Washington Post)

By Liz Goodwin and John Hudson
September 26, 2024 at 11:08 a.m. EDT

The chair of the Senate Foreign Relations Committee was lured into a video call with a "malign actor" probably using "deepfake" artificial intelligence technology to pose as a top Ukrainian official, lawmakers and congressional aides said Thursday.



The Washington Post
Democracy Dies in Darkness

TECHNOLOGY

AI is spawning a flood of fake Trump and Harris voices. Here's how to tell what's real.

By Pranshu Verma, Rekha Tenjarla and Bishop Sand

October 16, 2024 at 6:05 a.m.

Artificial intelligence has made it extraordinarily simple to copy someone's voice — allowing thousands of audio impersonations, known as "deepfakes," to flood the internet since early last year.

Generative AI threats will continue to challenge the information domain

https://www.washingtonpost.com/politics/2024/09/26/senator-lured-into-deepfake-call-with-malign-actor-posing-ukrainian/
https://www.washingtonpost.com/technology/interactive/2024/ai-voice-detection-trump-harris-deepfake-election/

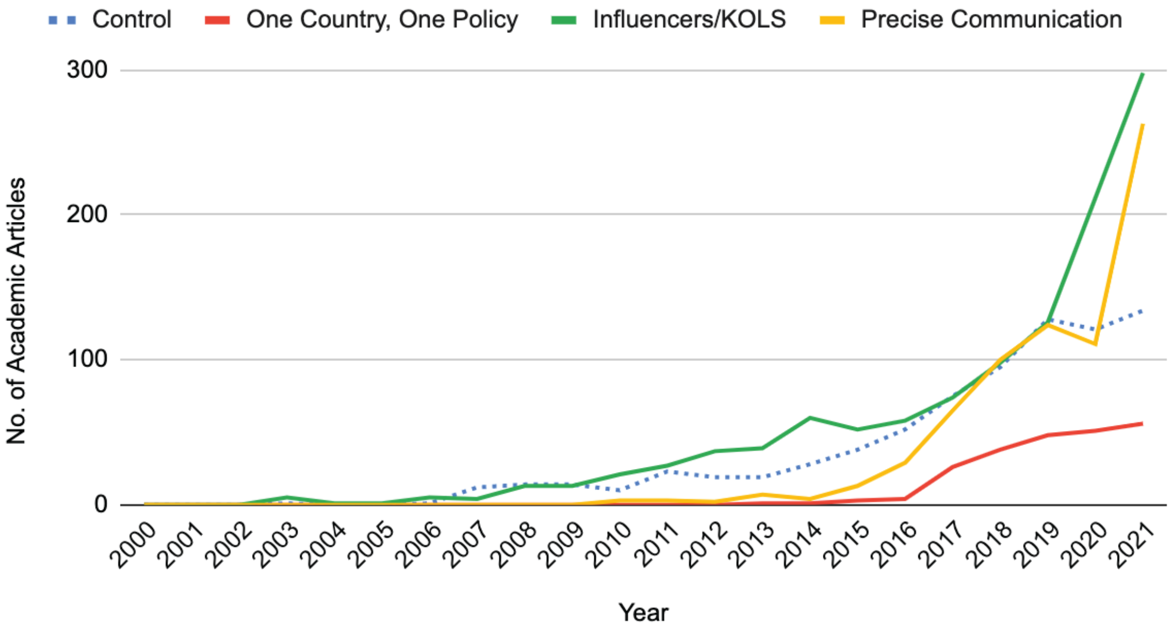
China is executing a strategy of highly-targeted mis/disinformation

‘1 Key for 1 Lock’: CCP targeted propaganda strategy

- In depth target audience understanding
- Area studies[1] research
- Target audience surveys
- Online behavioral data

https://go.recordedfuture.com/hubfs/reports/ta-2022-0928.pdf

Emergence of propaganda concepts



https://go.recordedfuture.com/hubfs/reports/ta-2022-0928.pdf

China is spending billions of dollars on highly targeted campaigns

KOLS – key opinion leaders

[1] Area studies - multidisciplinary social research focusing on specific geographic regions or culturally defined areas. (Britannica.com)

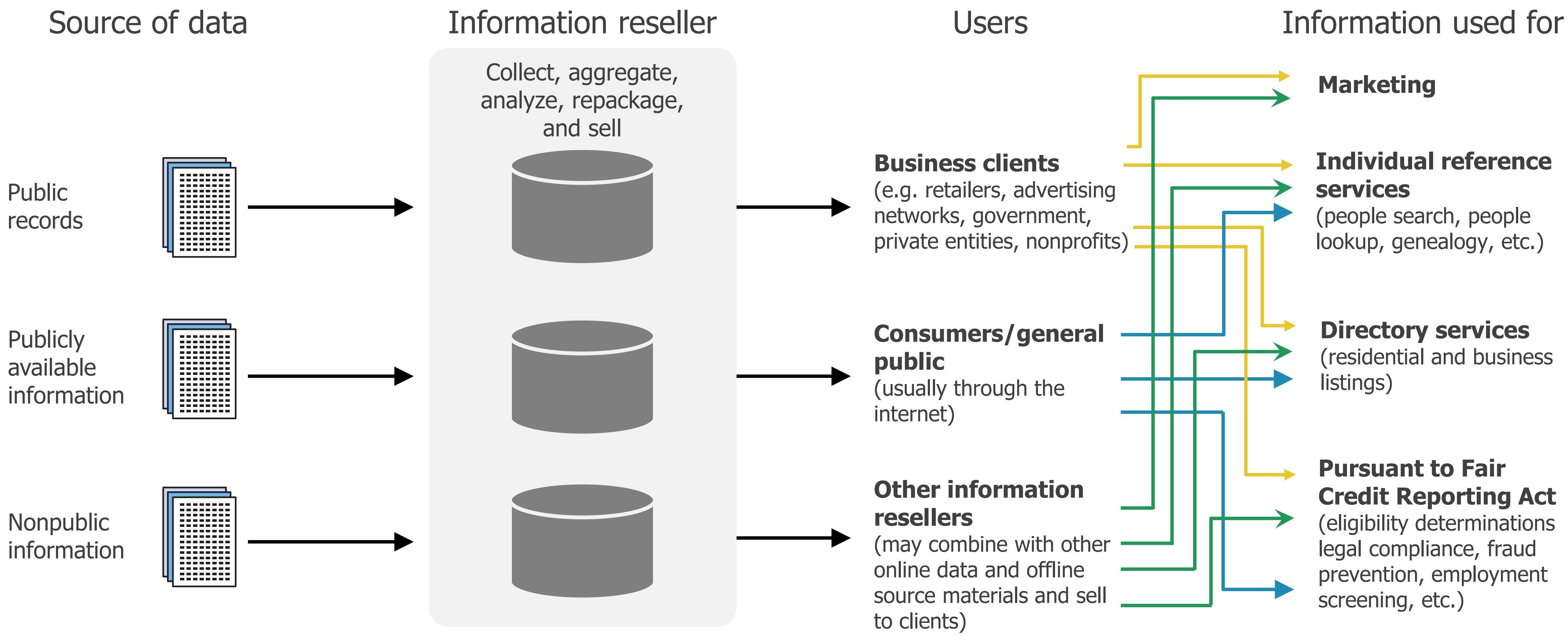
Distribution Statement A: Approved for Public Release, Distribution Unlimited.

PAI/CAI ecosystem aggregates sensitive data

AdTech market size

2022: $839B 2031: $2.6T (est.)

https://www.gminsights.com/industry-analysis/adtech-market/market-size

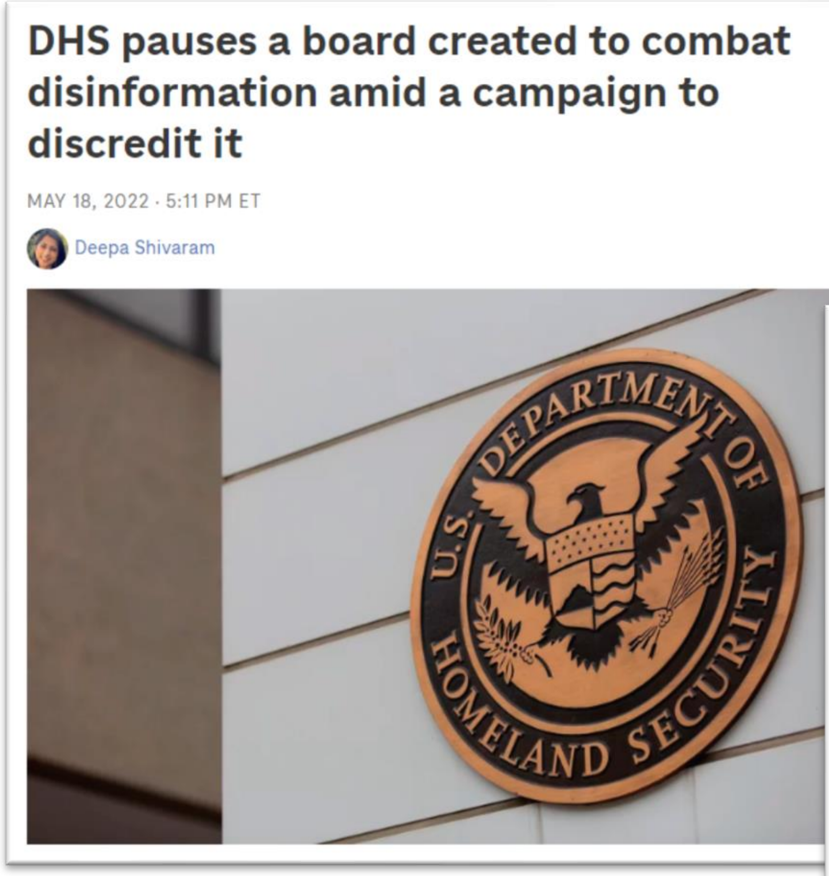

Source of data
Information reseller
Users
Information used for
Public records
Publicly available information
Nonpublic information
Collect, aggregate, analyze, repackage, and sell
Business clients (e.g. retailers, advertising networks, government, private entities, nonprofits)
Consumers/general public (usually through the internet)
Other information resellers (may combine with other online data and offline source materials and sell to clients)
Marketing
Individual reference services (people search, people lookup, genealogy, etc.)
Directory services (residential and business listings)
Pursuant to Fair Credit Reporting Act (eligibility determinations legal compliance, fraud prevention, employment screening, etc.)

PAI/CAI enables adversaries' exquisite surveillance

GAO-19-62IT
https://www.dni.gov/files/documents/FOIA/DF-2023-00146-Senior-Advisory-Group-Panel-on-CAI.pdf

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

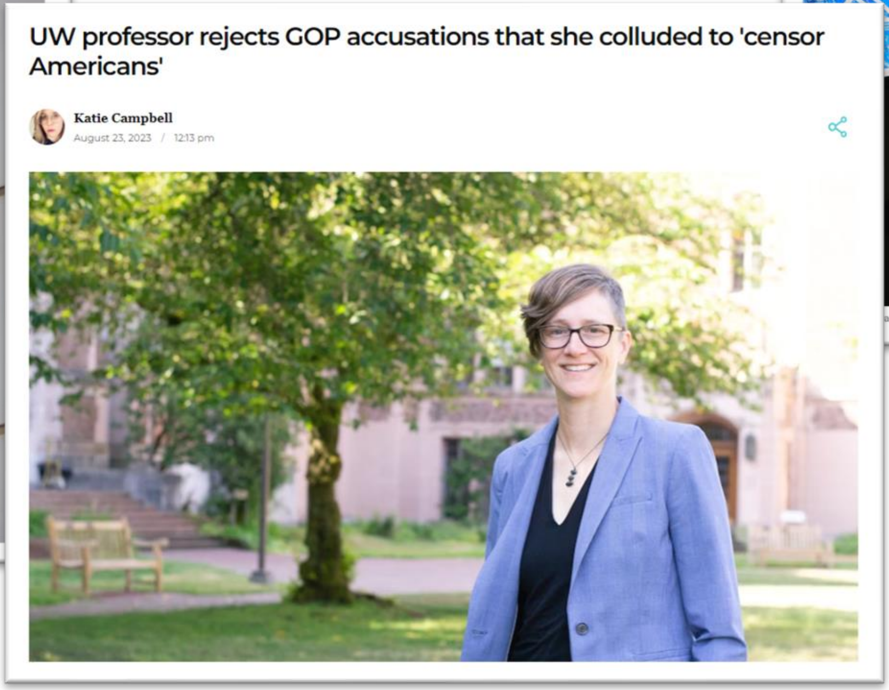
Chilly climate for researchers



DHS pauses a board created to combat disinformation amid a campaign to discredit it

MAY 18, 2022 · 5:11 PM ET

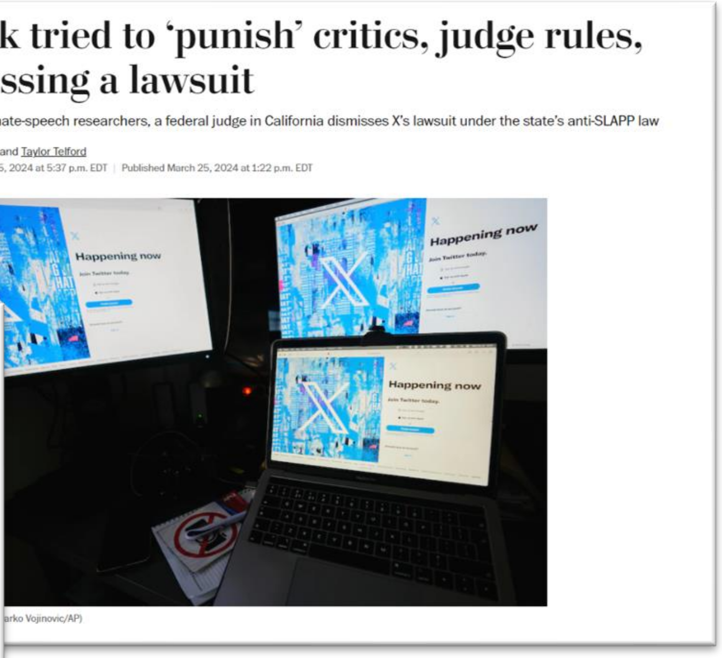
Deepa Shivaram



UW professor rejects GOP accusations that she colluded to 'censor Americans'

Katie Campbell

August 23, 2023 / 12:13 pm



Musk tried to 'punish' critics, judge rules, in tossing a lawsuit

In a win for hate-speech researchers, a federal judge in California dismisses X's lawsuit under the state's anti-SLAPP law

By Will Oremus and Taylor Telford

Updated March 25, 2024 at 5:37 p.m. EDT | Published March 25, 2024 at 1:22 p.m. EDT

arko Vojinovic/AP)

Relevant research must be carefully designed and executed

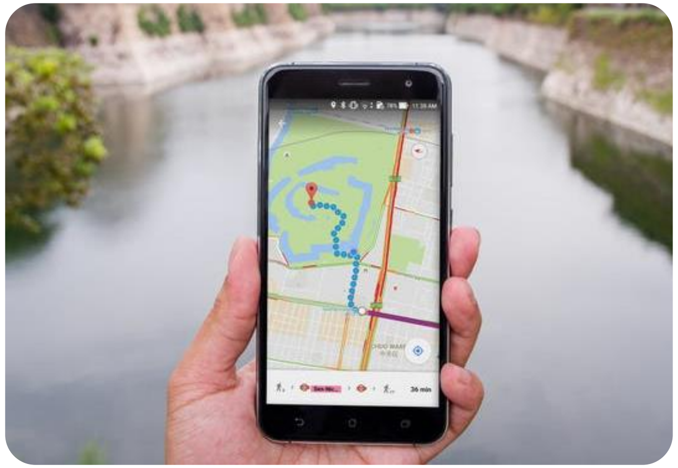
Information domain vision

Imagine a world where…

Socio-techno-information systems are effective in the face of external interference and increase individual and societal **resilience** to adversary threats

Individuals can have **confidence** in the information they consume and the sources of that information (large-scale information integrity)

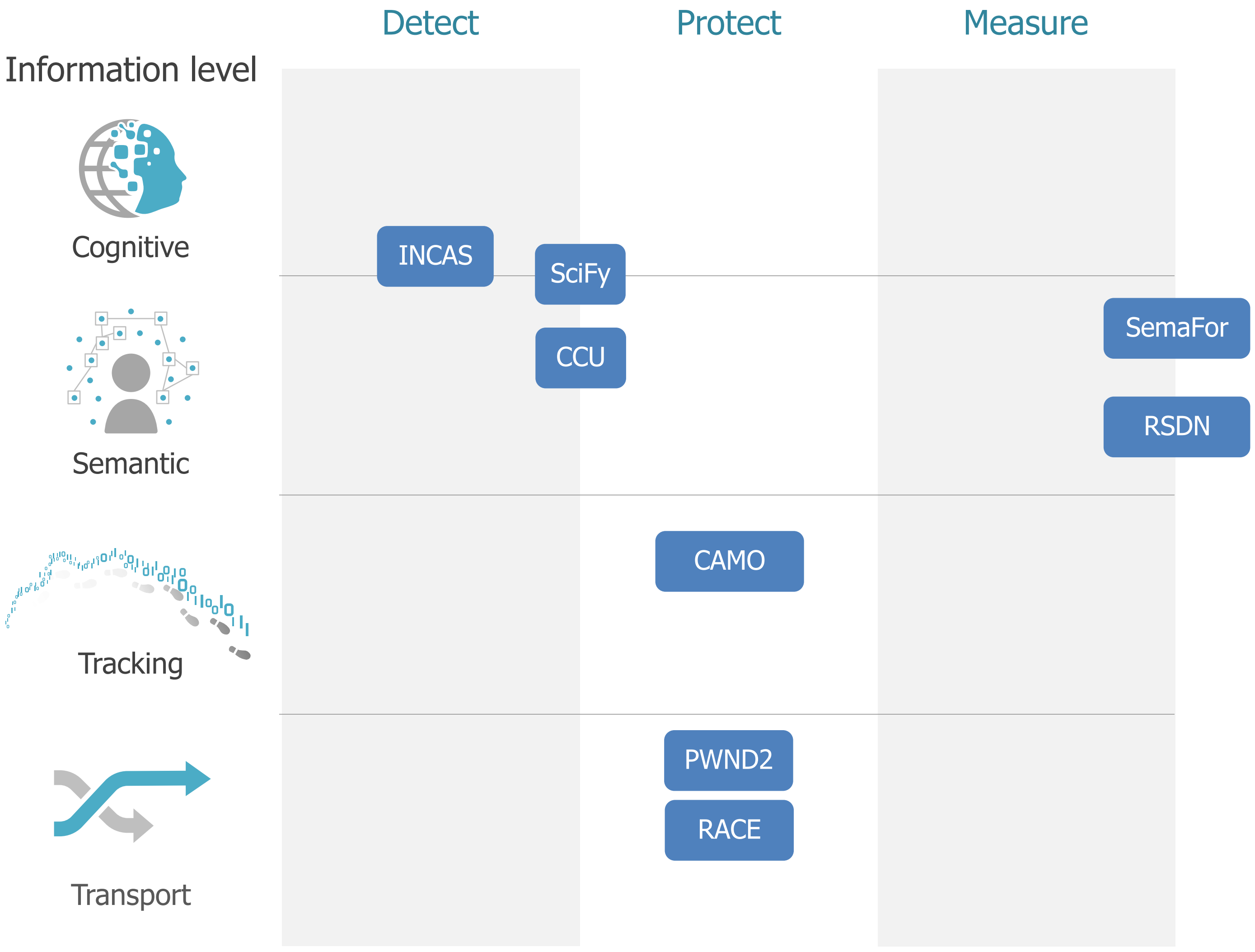
Privacy exists for civilians, government employees, and military personnel, even in the face of nation-state adversaries or authoritarian regimes



https://www.chemistryworld.com/news/research-integrity-may-be-linked-to-uk-grants/5218.article
https://www.businessinsider.com/guides/tech/how-to-save-a-location-on-google-maps
https://mwi.westpoint.edu/true-or-false-the-fight-against-disinformation/

Information domain mission

- Integrate defenses and capabilities across cyber and information domains
- Use AI, social science, and computer science to develop trustworthy tools to accomplish the vision at speed and scale
- Develop enduring relationships with mission partners to inform program development and expedite transition



Detect

Protect

Measure

Information level

Cognitive

Semantic

Tracking

Transport

INCAS

SciFy

CCU

SemaFor

RSDN

CAMO

PWND2

RACE

Key 🟦 Current programs

CCU: Computational Cultural Understanding
INCAS: Influence Campaign Awareness & Sensemaking

MICE Measuring the Information Control Environment
RACE: Resilient Anonymous Communication for Everyone
RSDN: Resilient Supply and Demand Networks
SemaFor: Semantic Forensics

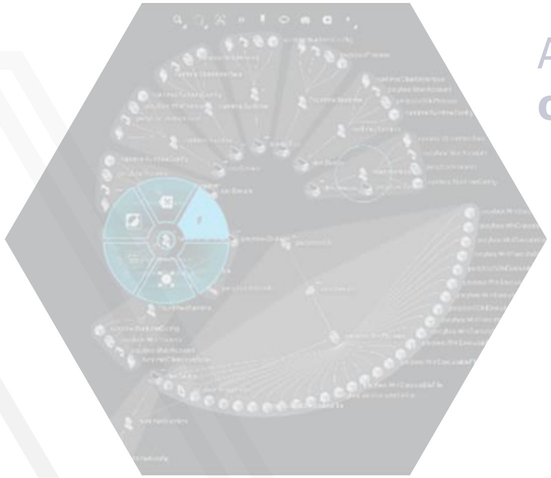
Impact maximization strategy

- Vision for the future: The DoD, IC, and society in general have
  - resilience in the face of information-based attacks
  - confidence in the information they consume and share
  - privacy and control of personal information

- Strategy to achieve this vision:
  - Message various constituencies thoughtfully regarding attack surfaces and the need for defensive research
    - Engage with the policy community to protect authorities for conducting research
  - Carefully design research programs that address hard problems while protecting research equities
    - Do excellent technical work
  - Use open source to help disperse technologies and create commercial industry
  - Partner with others
    - Work with a broad set of US government partners to enable relevant experimentation and evaluation
    - Build international collaborations to leverage diverse authorities, data sets, and operational perspectives
    - Build relationships with commercial companies to develop broad-based defenses

Vision: DARPA ensures that information continues to be valuable and under the control of its "owner"

Proficient **artificial intelligence**



Advantage in **cyber operations**

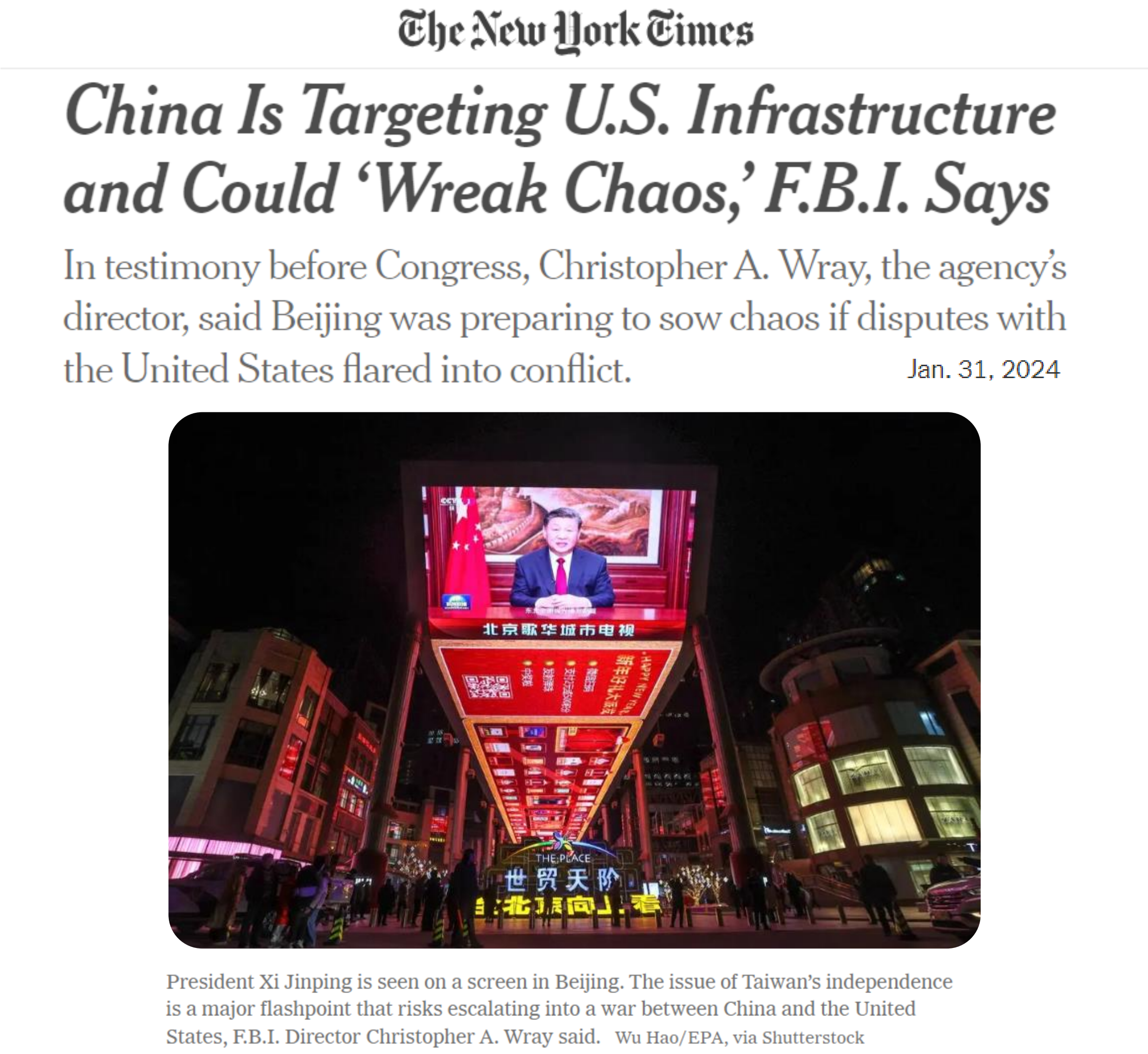
Confidence in the **information domain**

abc.com

af.mil

Resilient, adaptable, and **secure systems**

China prepositioning cyber implants on infrastructure



The New York Times

China Is Targeting U.S. Infrastructure and Could 'Wreak Chaos,' F.B.I. Says

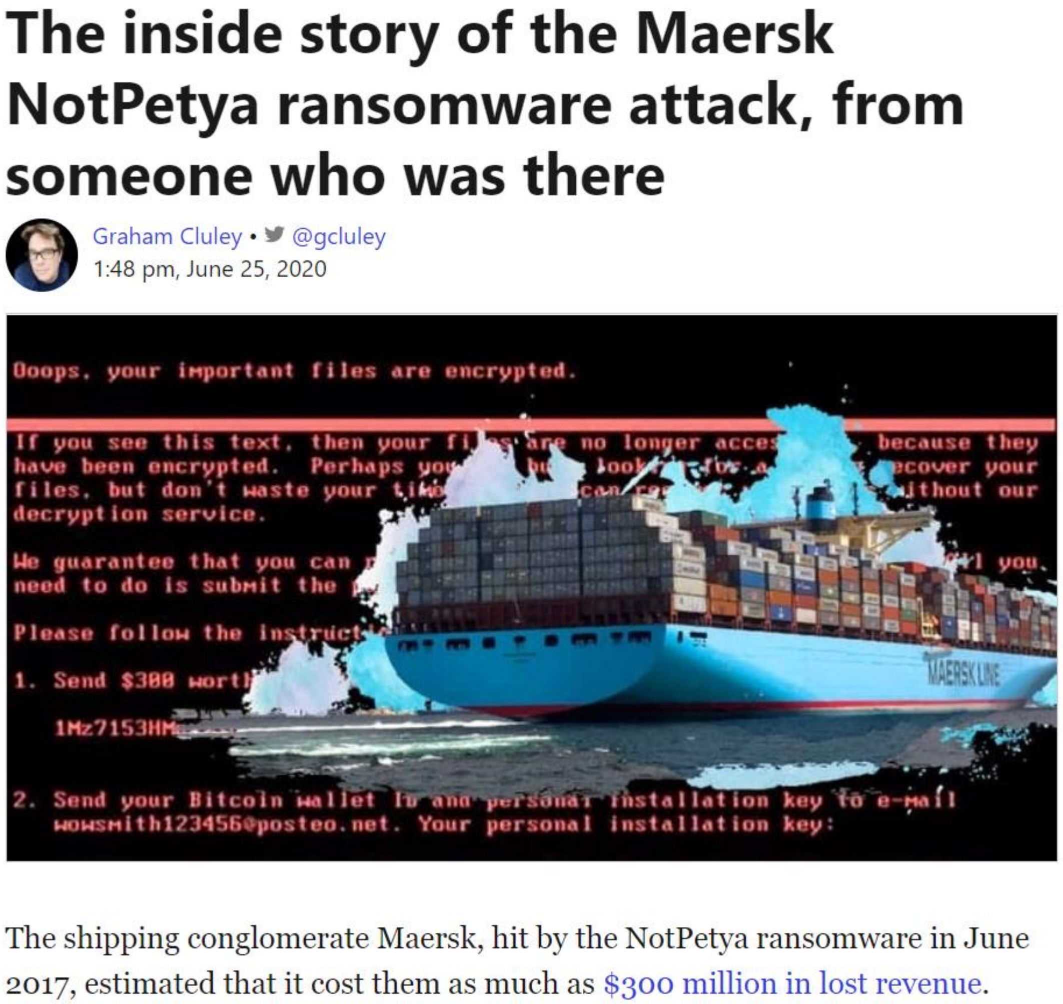
In testimony before Congress, Christopher A. Wray, the agency's director, said Beijing was preparing to sow chaos if disputes with the United States flared into conflict.

Jan. 31, 2024

President Xi Jinping is seen on a screen in Beijing. The issue of Taiwan's independence is a major flashpoint that risks escalating into a war between China and the United States, F.B.I. Director Christopher A. Wray said. Wu Hao/EPA, via Shutterstock

https://www.nytimes.com/2024/01/31/us/politics/fbi-director-china-wray-.html

Cyber attacks have broad impact on infrastructure



The inside story of the Maersk NotPetya ransomware attack, from someone who was there

Graham Cluley • @gcluley
1:48 pm, June 25, 2020

Ooops, your important files are encrypted.

If you see this text, then your fi...s are no longer acces... because they have been encrypted. Perhaps you... look... for a... ecover your files, but don't waste your tim... can... ...ithout our decryption service.

We guarantee that you can r... ...l you need to do is submit the p...

Please follow the instruct...

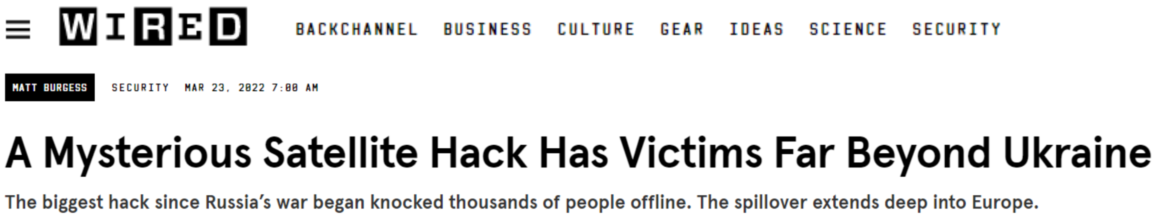
1. Send $300 worth...

1Mz7153HM...

2. Send your Bitcoin wallet ID and personal installation key to e-mail wowsmith123456@posteo.net. Your personal installation key:

The shipping conglomerate Maersk, hit by the NotPetya ransomware in June 2017, estimated that it cost them as much as $300 million in lost revenue.

Cyber attack targets Ukraine communications



MATT BURGESS SECURITY MAR 23, 2022 7:00 AM

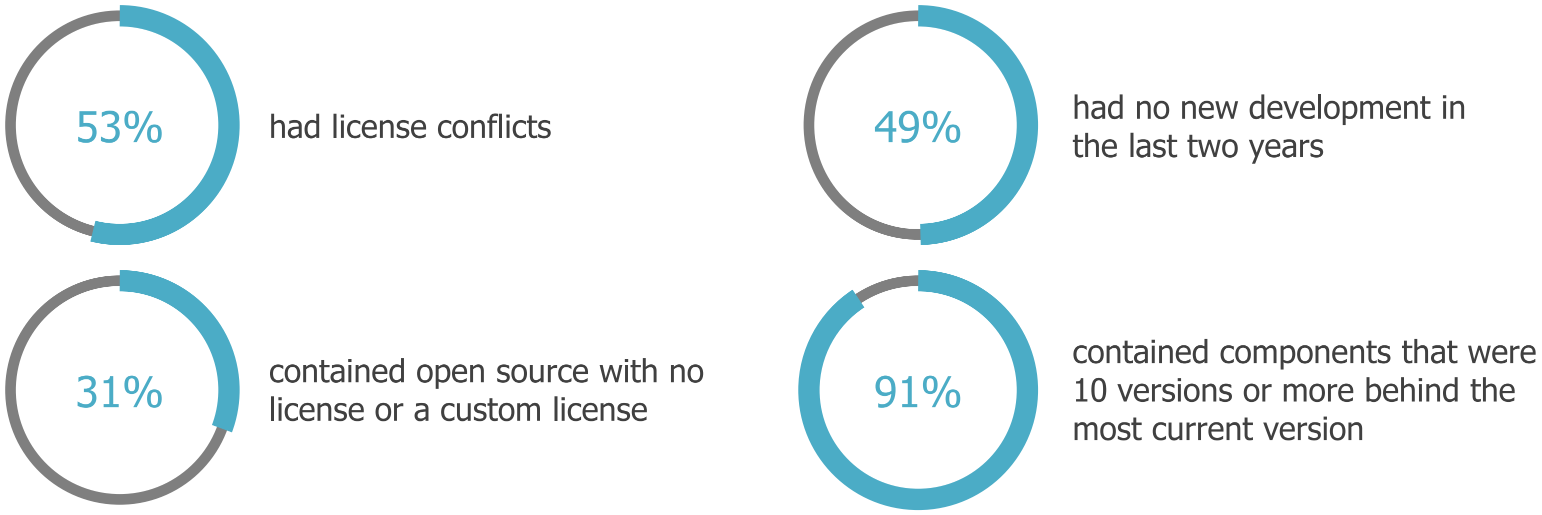
A Mysterious Satellite Hack Has Victims Far Beyond Ukraine

The biggest hack since Russia's war began knocked thousands of people offline. The spillover extends deep into Europe.

10s of thousands of connections affected, including parts of Ukraine's defenses

Huge exposure continues: Open source risk assessment

Of the 1,067 codebases scanned in 2023, 84% contained vulnerabilities



53% had license conflicts

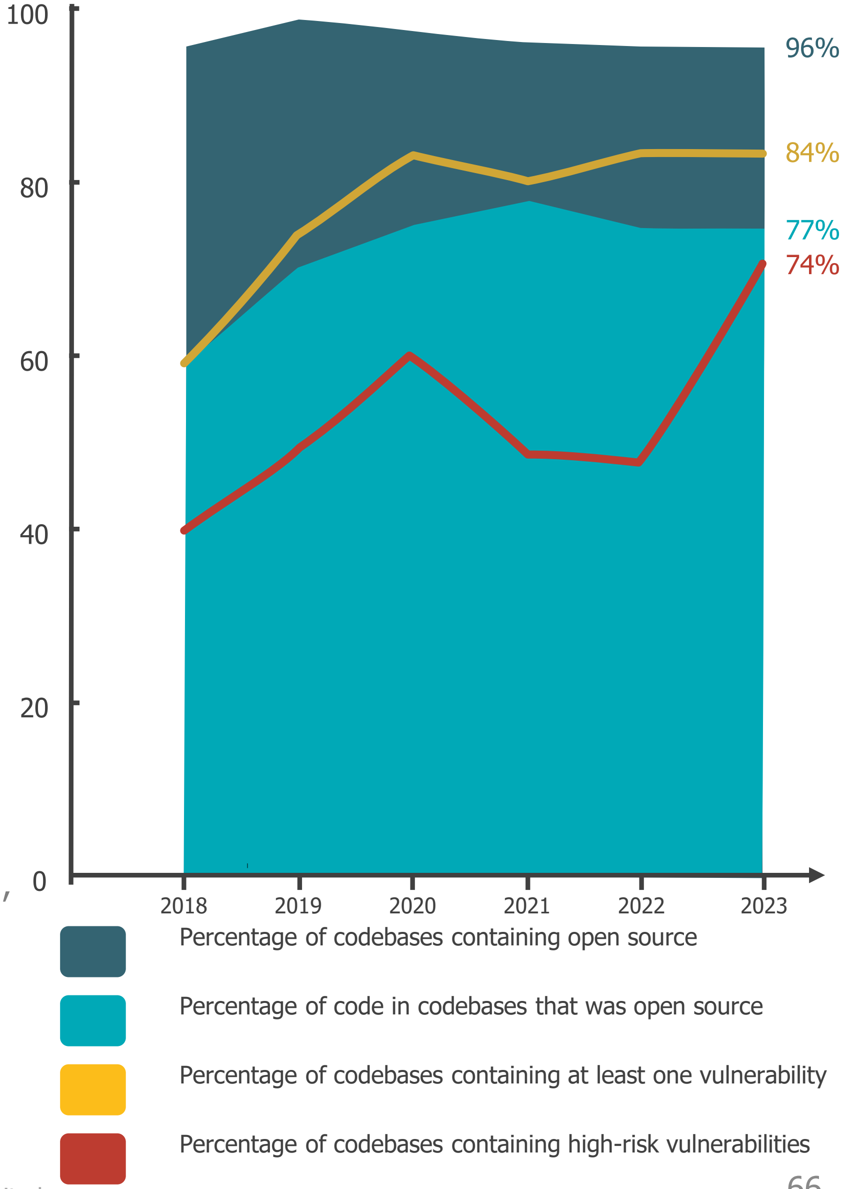
49% had no new development in the last two years

31% contained open source with no license or a custom license

91% contained components that were 10 versions or more behind the most current version

Open-source risk assessment[1]

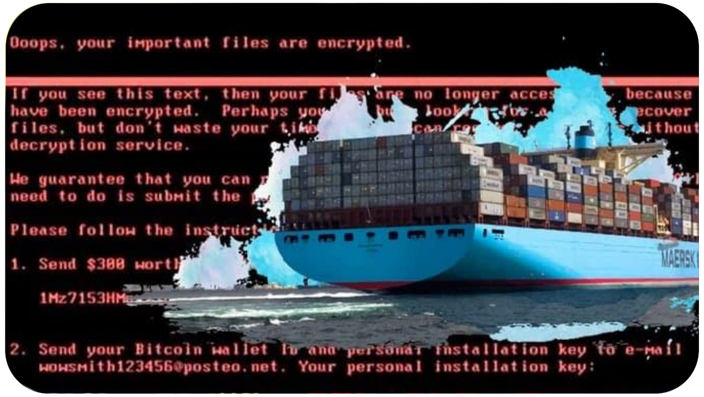
Analysis includes data from aerospace, aviation, automotive, transportation, logistics, computer hardware, semiconductor, cybersecurity, energy and clean tech, financial services and fintech, healthcare and health tech, internet software and infrastructure, internet of things, manufacturing, robotics, and telecommunications industries.



Open source risks are indicative of closed-source codebase risks

Our systems are vulnerable

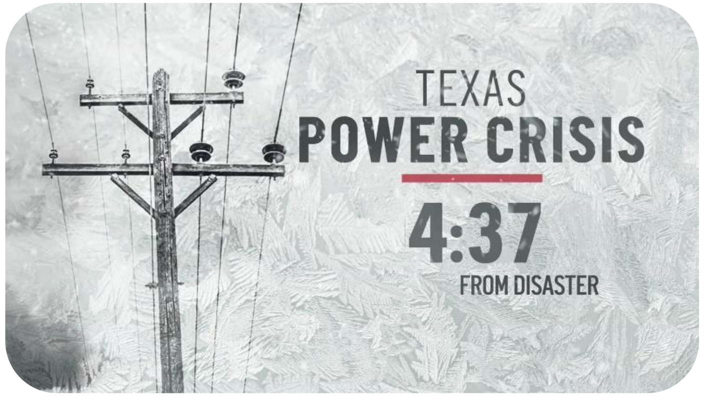
Growing interdependencies in mega-systems



```
Ooops, your important files are encrypted.

If you see this text, then your fi    are no longer acces        because
have been encrypted.  Perhaps you       look                    ecover
files, but don't waste your tim          can                    ithout
decryption service.

We guarantee that you can
need to do is submit the

Please follow the instruct

1. Send $300 worth

   1Mz7153HM

2. Send your Bitcoin wallet ID and personal installation key to e-mail
   wowsmith123456@posteo.net. Your personal installation key:
```

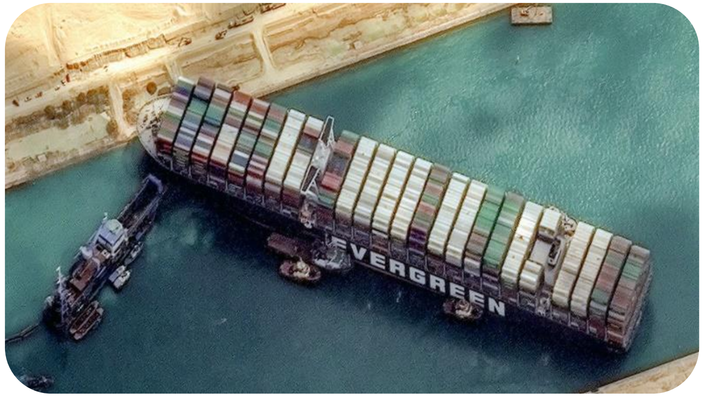
grahamcluley.com

2017 shipping conglomerate Maersk, hit by the NotPetya ransomware – $300M lost revenue



www.nbcdfw.com

2021 Texas grid crisis collapse – multi-day power outage affecting over 11 million people



www.theatlantic.com

2021 The Evergreen container ship control failure causes a closure of the Suez Canal

SYSTEM CRASH:
THE FAA NOTAM OUTAGE OF JANUARY 2023
AND ITS IMPACT ON AVIATION

brotallion.com

2023 FAA Notice To Air Missions (NOTAM) outage – All air operations in US suspended for over 12 hours

microtime.com

2024 Change Healthcare payment system experienced a crippling ransomware attack

nextgov.comnextgov.com

2024 CrowdStrike software errors melted down the world's computer systems

Society is dependent on many marginally stable mega-systems that have multiple exposed tipping points and may not be restorable if/when they go down

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

Resilient, adaptable, and secure systems

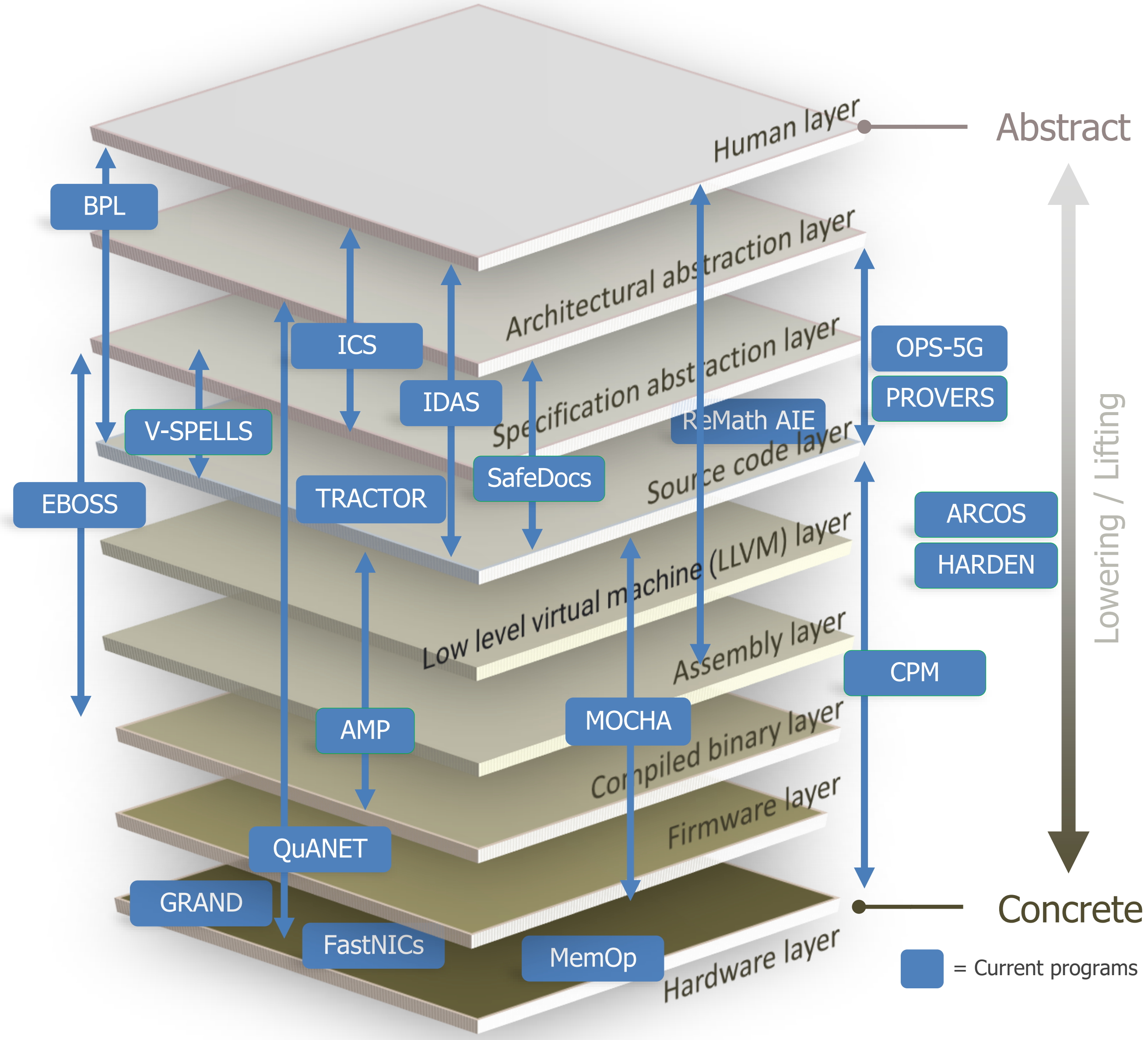
Vision

- Software-based systems are cost effective to build, maintain and deploy; only work they way they are intended; and are resilient in the face of adversary attack or other failures

Approach

- Use formal methods and 3rd-wave AI to make it easier to understand, build, update, repair, and restore socio-software systems with system-wide, security-relevant correctness guarantees.

A world without software vulnerabilities is possible!



AMP: Assured Micropatching
ARCOS: Automated Rapid Certification Of Software
BPL: Business Process Logic
CASE: Cyber Assured Systems Engineering
CPM: Cyber Assured Systems Engineering (CASE)
EBOSS: Enhanced SBOM for Optimized Software Sustainment
FastNICs: Fast Network Interface Cards
HARDEN: Hardening Development Toolchains against Emergent Execution Engines
ICS: Intrinsic Cognitive Security
IDAS: Intent-Defined Adaptive Software

MemOp: Memory Optimization
OPS-5G: Open, Programmable, Secure 5G
PROVERS: Pipelined Reasoning Of Verifiers Enabling Robust Systems
QuANET: Quantum-Augmented Network
ReMath: Recovery of Symbolic Mathematics from Code
SafeDocs: Safe Documents
Social Cyber: Hybrid AI to Protect Integrity of Open Source Code
V-SPELLS: Verified Security and Performance Enhancement of Large Legacy Software

Impact maximization strategy

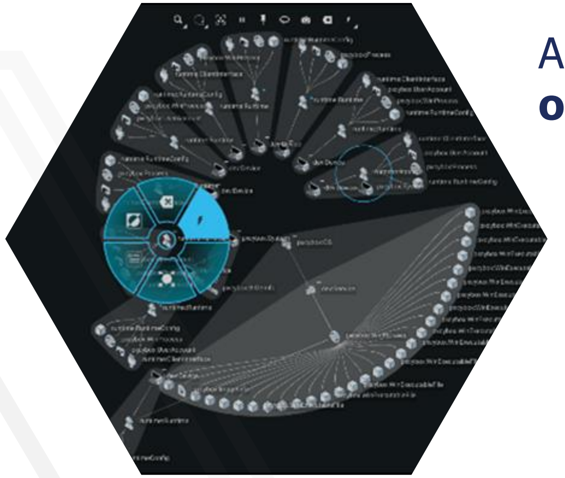
- Vision: Cost-effective, secure, resilient, and maintainable software is used throughout the DoD

- Strategy to achieve this vision
  - Drive research towards (semi-)automated approaches to high-trust and resiliency
  - Demonstrate excellent results on real-world problems and publish in noteworthy venues
  - Transition technology to individual operational systems to demonstrate success and build credibility
  - Build enduring collection and maintenance of technology
  - Raise the bar when systems fail during tests because there is an alternative
  - Work to change policy to make higher assurance a requirement (or at least a gold star)
  - Accelerate adoption in the defense industrial base
    - Lessons from Amazon Web Services: Formal method techniques can reduce cost, improve schedule, and increase security

We know how to build software that is much harder to hack; we need to get the DoD to adopt those techniques.

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

Proficient **artificial intelligence**



Advantage in **cyber operations**

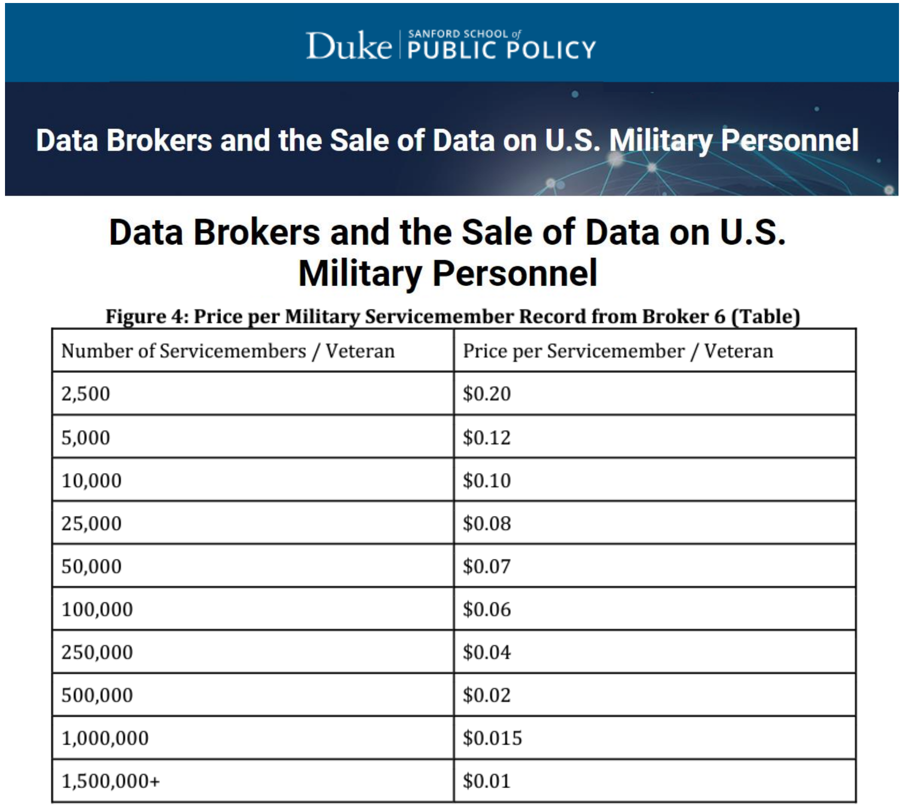
Confidence in the **information domain**

abc.com

af.mil

Resilient, adaptable, and **secure systems**

People are a weak link



Duke | SANFORD SCHOOL of PUBLIC POLICY

Data Brokers and the Sale of Data on U.S. Military Personnel

Data Brokers and the Sale of Data on U.S. Military Personnel

Figure 4: Price per Military Servicemember Record from Broker 6 (Table)

Number of Servicemembers / Veteran	Price per Servicemember / Veteran
2,500	$0.20
5,000	$0.12
10,000	$0.10
25,000	$0.08
50,000	$0.07
100,000	$0.06
250,000	$0.04
500,000	$0.02
1,000,000	$0.015
1,500,000+	$0.01

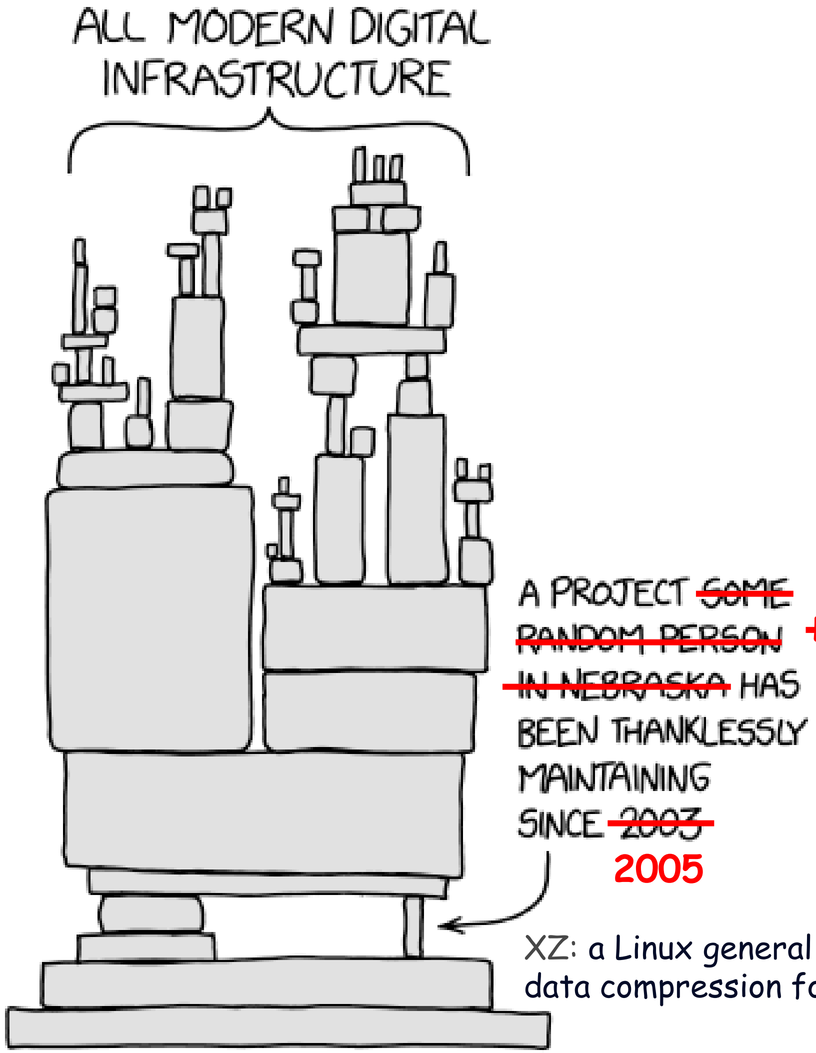
Ad tech makes it possible to target people relevant for national security for low cost

- PII data for 35K US military personal can be purchased for $7K, geofenced to military areas
- Reveals income, health condition, income, political affiliation, net worth, gender, etc.

Attacks on people will continue to grow in effectiveness, speed, and scale

https://techpolicy.sanford.duke.edu/data-brokers-and-the-sale-of-data-on-us-military-personnel/

Some guy in Nebraska


ALL MODERN DIGITAL INFRASTRUCTURE
A PROJECT ~~SOME RANDOM PERSON IN NEBRASKA~~ HAS BEEN THANKLESSLY MAINTAINING SINCE ~~2003~~

~~LASSE COLLIN~~

~~JIA TAN~~

VOUCHED BY FIVE OTHER GITHUB USERS

~~SOME RUSSIAN GUY~~

ANDRES FREUND,
A SOFTWARE ENGINEER AT MICROSOFT

2005

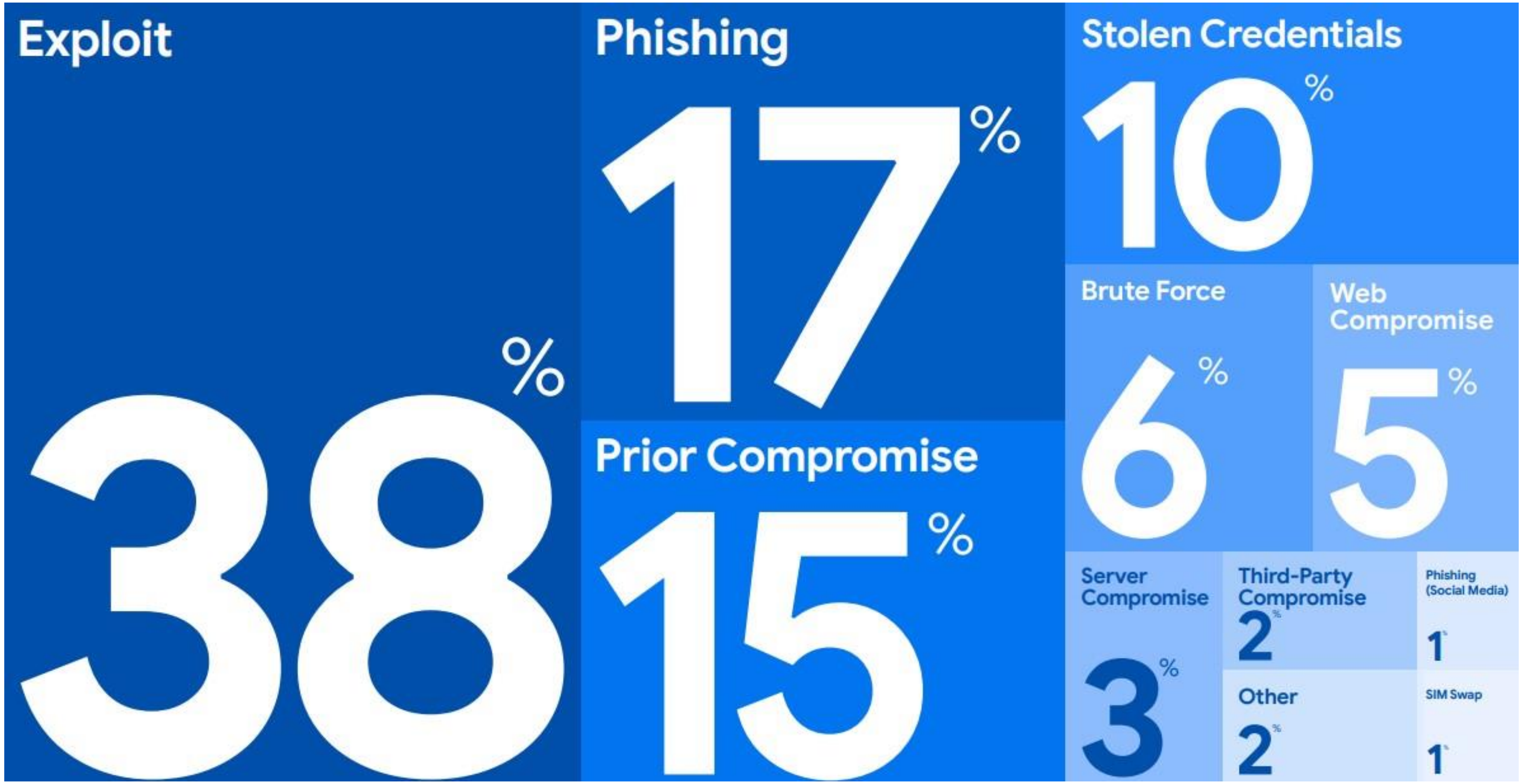
XZ: a Linux general purpose data compression format

https://xkcd.com/2347/

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

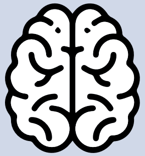
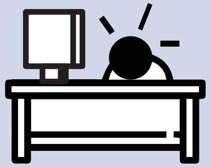
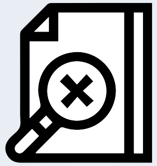
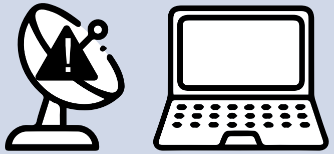
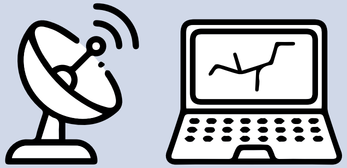
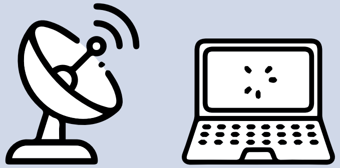
Software vulnerabilities enable ransomware attacks

We depend on software that is pervasively vulnerable and increasingly under attack. This includes critical infrastructure software where system failure has dire consequences.



Initial Ransomware Infection Vector, “Mandiant M-Trends 2024”

Socio–techno systems: Converging domains

	Denial of Service	Crash	Hack (may or may not be detected/attributed)
 Cognitive			
 Cyber			
 Electro-Magnetic			

Adversaries will use whatever combination of attacks most likely accomplish their goals

Cyber operations vision

Imagine a world in which blue forces have complete confidence in their cyber capabilities

- Systems, at all levels of levels of complexity, connectivity, and integration are **resilient in the face of adversary attack** and provide accurate situational awareness to relevant authorities in a timely fashion

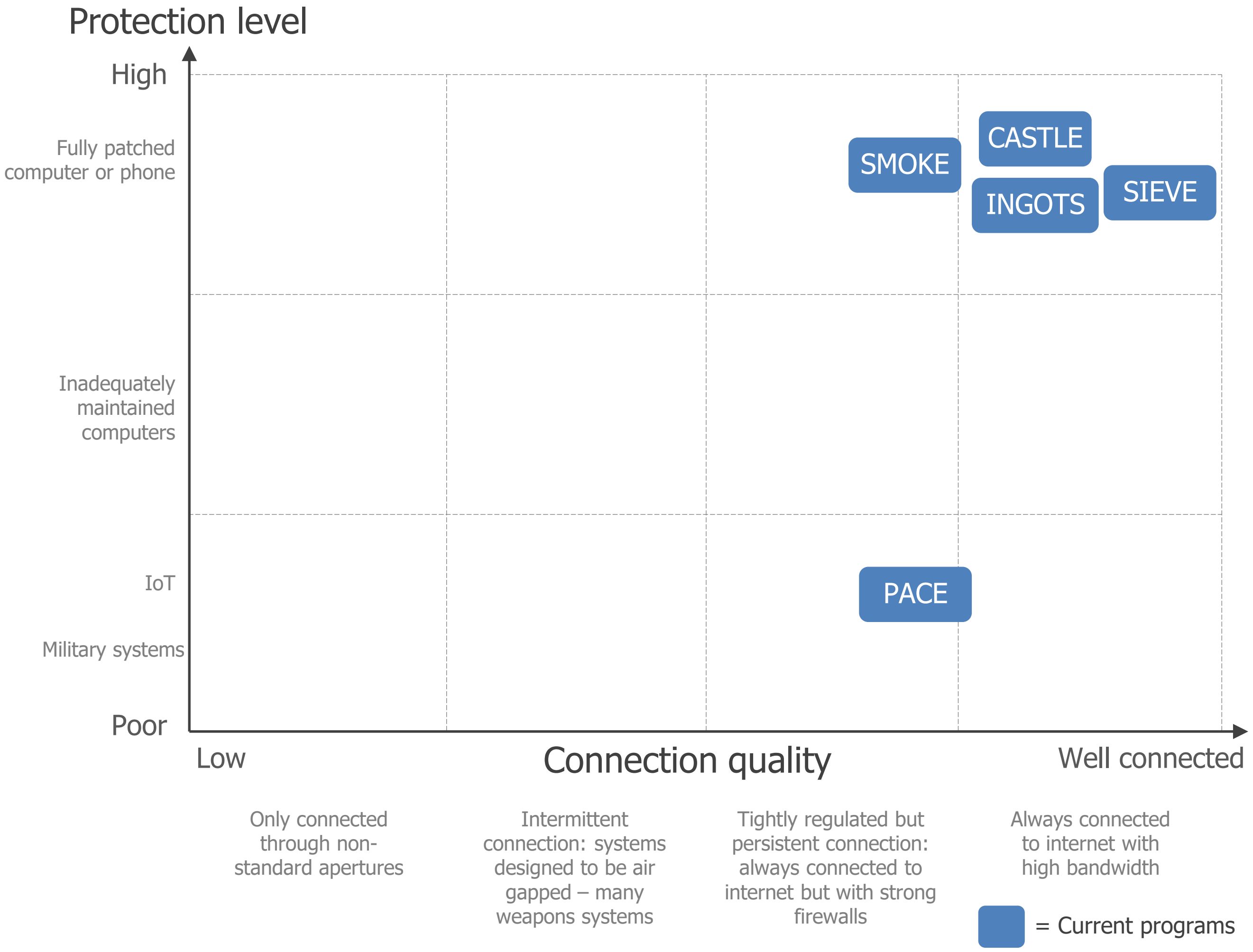
- Systems, at all levels of levels of complexity, connectivity, and integration are **vulnerable to blue force attack** and provide accurate situational awareness to relevant authorities in a timely fashion

Image generator
https://chat.openai.com/

Distribution Statement A: Approved for Public Release, Distribution Unlimited.

Advantage in cyber operations

- Advance and apply cutting-edge techniques from AI, formal methods, and program analysis to develop trustworthy methods that work at speed and scale
- Integrate defenses and capabilities across domains
- Consider all levels of the stack, from hardware to human, and all stages of software lifecycle
- Develop enduring capabilities: factories not bullets
- Develop enduring relationships with CYBERCOM and other mission partners to inform program development and expedite impact



BPL: Business Process Logic
CASTLE: Cyber Agents for Security Testing and Learning Environments
INGOTS: Intelligent Generation of Tools for Security
SIEVE: Securing Information for Encrypted Verification and Evaluation
OPS-5G: Open, Programmable, Secure 5G
PACE: Program Analysis for Capabilities Excellence
RACE: Resilient Anonymous Communication for Everyone
SMOKE: Signature Management using Operational Knowledge and Environments

Impact maximization plan: Constellation

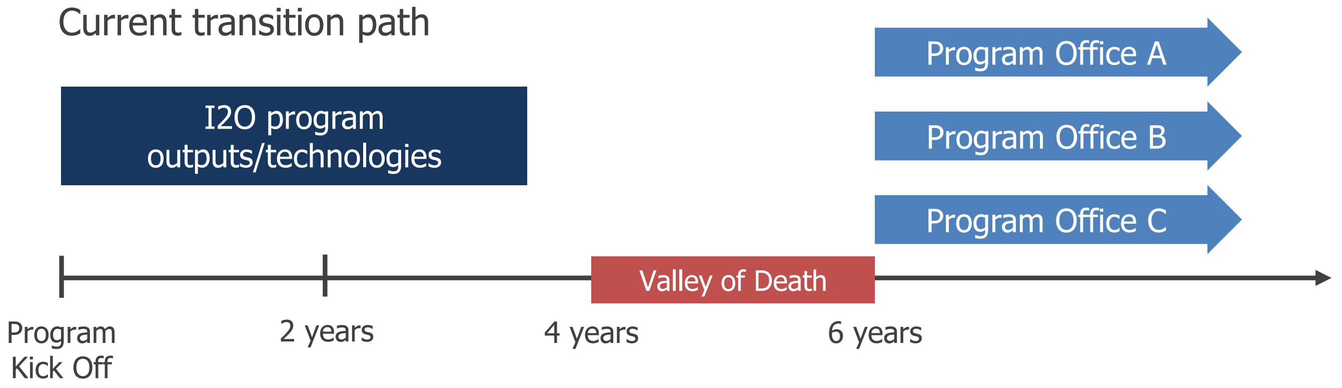
Develop novel contracting approaches to enable rapid and iterative transition of maturing tactical and strategic cyber capabilities to operational warfighting platforms

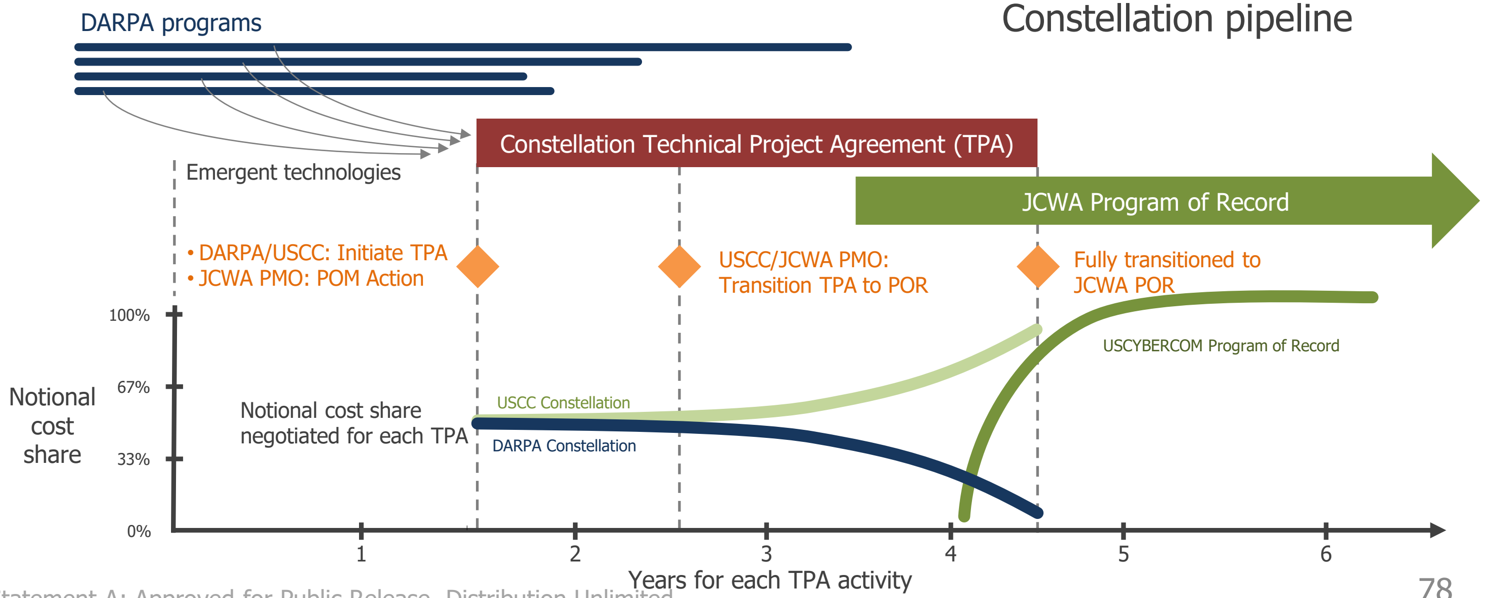
Problem

- Program office requirements do not always align to technology development
- Research programs must plan for transition before the technology has been developed
- Acquisition timelines force program offices to plan budgets 3 years in advance

Approach

- Developed a consortium of performers using Other Transaction Authority
- Team Orion established
  - Consortium of cyber R&D companies and USCYBERCOM system integrators organized to conduct advanced technology development


Current transition path


Constellation pipeline

USCC: USCYBERCOM
JCWA: Joint Cyber Warfighting Architecture
PMO: Program Management Office
POR: Program of Record
TPA: Technical Project Agreement

www.darpa.mil

I2O Office Wide Proposers Day Agenda

10:00	*11:00*	*Check-in and Networking Coffee*
11:00	11:05	Security Overview
11:05	11:15	Opening Remarks – Rob McHenry
11:15	11:35	How to Work with DARPA
		Commercial Strategy – Jen Thabet
		Small Business – Aaron Sparks
		DARPA Connect – Sana Sood
11:35	12:35	I2O Strategy – Kathleen Fisher
12:35	***1:35***	***Break for Lunch***
1:35	2:05	PM Presentations – (Dewhurst, Bernsen, Sweet, Kuhn, Cook)
2:05	2:15	Delivering on the DARPA Mission – Matt Turek
2:20	3:55	Sidebars

PM Introductions

Introduction/PM Background: David Rushing Dewhurst

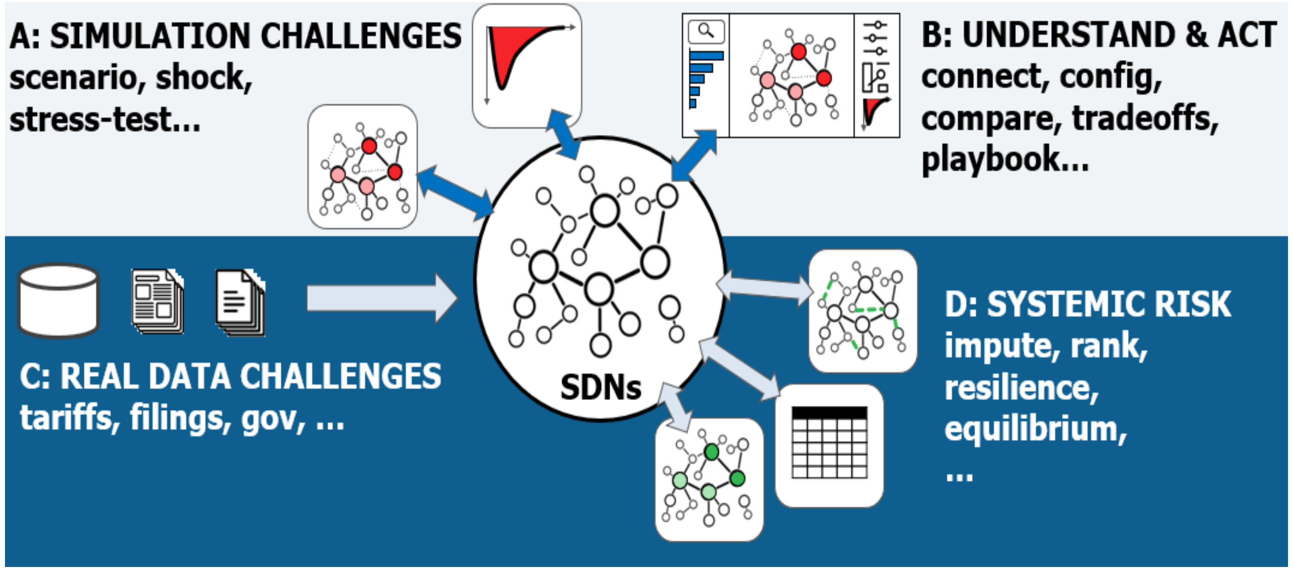
Before DARPA PM:

- STO tech SETA – economic strategy
- Fellow at Yale
- Defense R&D + economic capital risk management in Cambridge/Boston

RSDN uses agent-based modeling and statistical inference to forward and reverse stress-test complex global supply and demand networks, decreasing the impact of shocks to military readiness and society alike

As DARPA PM:

- Since April 2024
- Resilient Supply and Demand Networks (RSDN)


A: SIMULATION CHALLENGES
scenario, shock, stress-test…
B: UNDERSTAND & ACT
connect, config, compare, tradeoffs, playbook…
C: REAL DATA CHALLENGES
tariffs, filings, gov, …
SDNs
D: SYSTEMIC RISK
impute, rank, resilience, equilibrium, …

Interest Areas:

- Geoeconomic strategy
- Supply chain
- Financial intelligence
- Capital markets

Introduction/PM Background: Derek Bernsen

Before DARPA:

- 12 years, Navy Cyber Warfare Engineer
  - Offensive and defensive cyber
  - Special operations
  - Intelligence
  - Vulnerability research
  - Privacy
  - ICS/SCADA, cryptography

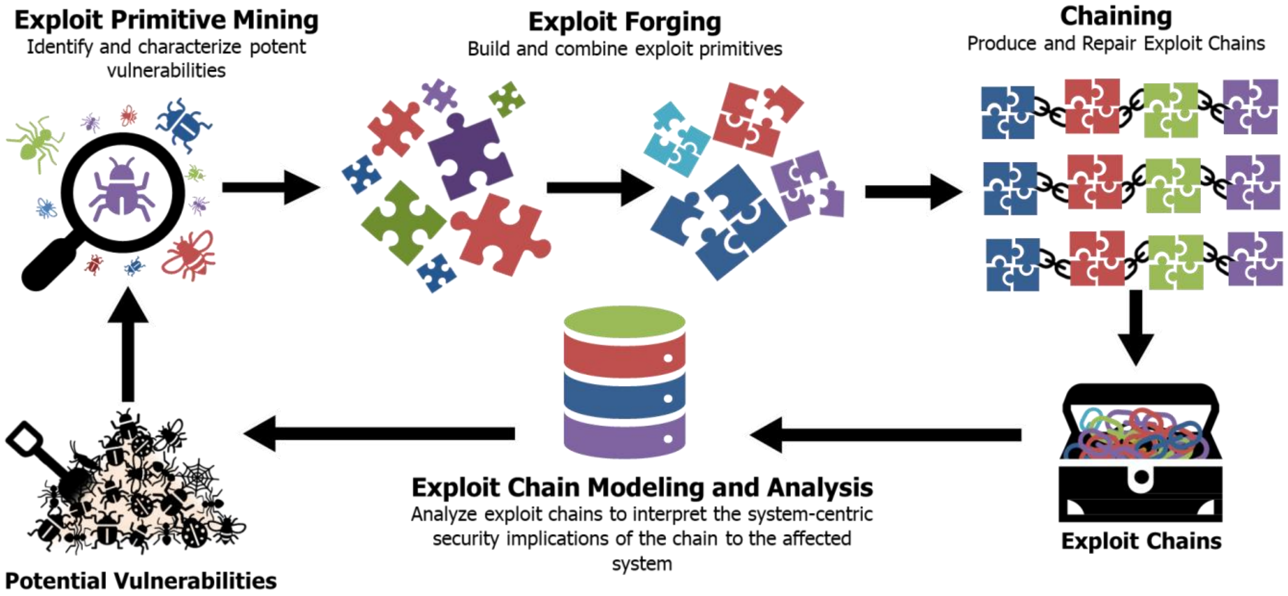
At DARPA:

- Since Oct 2022, PM since Jul 2024
- INtelligent Generation of Tools for Security (INGOTS)

Interest Areas:

- Cyber and special operations
- AI security
- Non-kinetic effects

INGOTS will research and develop Automatic Exploit Chain Generation (AECG) which will enable the DoD to analyze the security of modern complex systems, better understand the scope and severity of exploits and chains.


Exploit Primitive Mining
Identify and characterize potent vulnerabilities
Exploit Forging
Build and combine exploit primitives
Chaining
Produce and Repair Exploit Chains
Exploit Chains
Exploit Chain Modeling and Analysis
Analyze exploit chains to interpret the system-centric security implications of the chain to the affected system
Potential Vulnerabilities

Introduction/PM Background: Shauna Sweet

Before DARPA:

- Design and deployment of surveys among transient and hard-to-reach populations in contested environments
- Systems engineer to transition and field multiple modernization efforts of high assurance mission-critical systems

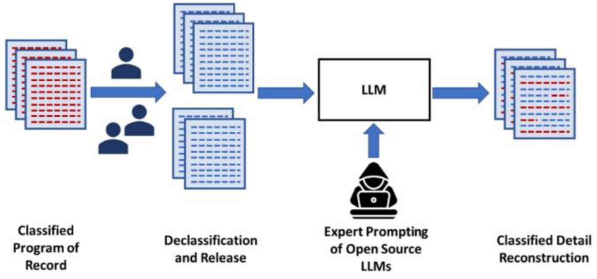
At DARPA:

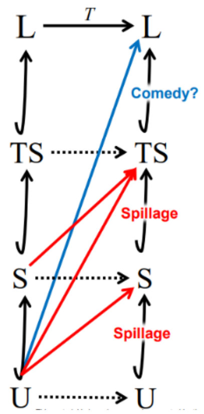
- Since June 2024
- Emergent Risks Seedling

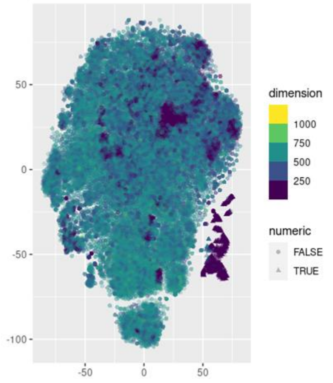
Interest Areas:

- High-assurance complex systems design and development
- Psychometric modeling and measurement theory
- Synthetic data design and deployment

Emergent Risks seeks to characterize the bounds of syntactically and semantically coherent generation and develop measures to identify conditions under which high-capacity models generate output in a way that fully exercises a partially hidden target domain.


Classified Program of Record
Declassification and Release
LLM
Expert Prompting of Open Source LLMs
Classified Detail Reconstruction





Threat model definition → **Unified conceptual framework** → **Novel algorithmic implementation**

> "Having access to a mass of evidence is one thing; constructing defensible arguments on the basis of this evidence is quite another."
> – David Schum, *The Evidential Foundations of Probabilistic Reasoning*

Introduction/PM Background: Stephen Kuhn

Before DARPA:

- 20+ years R&D experience in DOD
  - Offensive and defensive cyber
  - Hypervisor and Operating system design
  - Cross Domain
  - Blended Architectures (x86/FPGA)
- OUSD
- Intel/Signals Officer Navy reserve

- Performer on several DARPA programs
- COR/SME on SafeDocs, AMP, V-spells, AIMEE, Re-MATH, SocialCyber, Harden

At DARPA:

- Focus on broad adoption of Formal Methods by the DoD / USG
  - How do we bring the art of the science into the field?

Interest Areas:

- Resilient Systems
- Software designs using FM to enforce standardization
- Electronic Warfare integrating with cyber using the above

Introduction/PM Background: Byron Cook

Before DARPA:

- Proofs
  - Biology
  - Distributed systems
  - GenAI
  - Hardware
  - Networks/Policies
  - Operating systems
  - Railways
- Tools
  - Prover and Z1
  - SLAyer, separation logic
  - SLAM / Device drivers
  - Terminator
  - Zelkova
  - Tiros
  - New mystery tool to be announced soon

At DARPA: Since last week, 20% time (joint with Amazon

Interest Areas:

- Fundamental proof search techniques
- Biology reasoning
- Rebooting compliance checking
- Hardware supply chain
- Fighting scammers
- Making proof approachable to all

Delivering on the DARPA Mission

Delivering on the DARPA Mission

Matt Turek

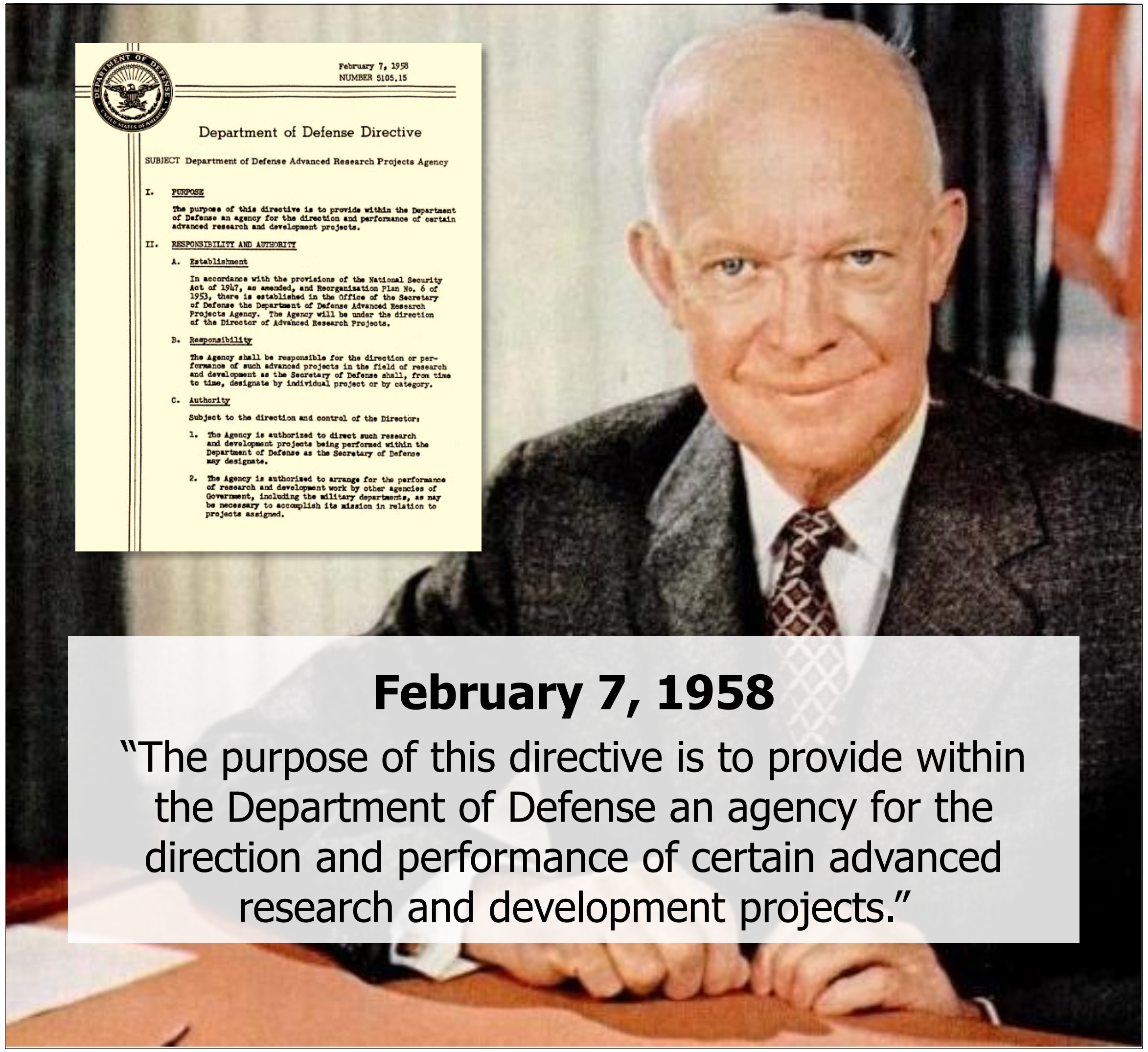
Deputy Director, Information Innovation Office

November 2024

Origins

"All the News That's Fit to Print"
The New York Times.
LATE CITY EDITION
VOL. CVII..No. 36,414.
NEW YORK, SATURDAY, OCTOBER 5, 1957.
FIVE CENTS

October 4, 1957
U.S.S.R. beats U.S. to space with Sputnik satellite; U.S. should never again be surprised by technology.



February 7, 1958
NUMBER 5105.15

Department of Defense Directive

SUBJECT Department of Defense Advanced Research Projects Agency

I. PURPOSE

The purpose of this directive is to provide within the Department of Defense an agency for the direction and performance of certain advanced research and development projects.

II. RESPONSIBILITY AND AUTHORITY

A. Establishment

In accordance with the provisions of the National Security Act of 1947, as amended, and Reorganization Plan No. 6 of 1953, there is established in the Office of the Secretary of Defense the Department of Defense Advanced Research Projects Agency. The Agency will be under the direction of the Director of Advanced Research Projects.

B. Responsibility

The Agency shall be responsible for the direction or performance of such advanced projects in the field of research and development as the Secretary of Defense shall, from time to time, designate by individual project or by category.

C. Authority

Subject to the direction and control of the Director:

1. The Agency is authorized to direct such research and development projects being performed within the Department of Defense as the Secretary of Defense may designate.

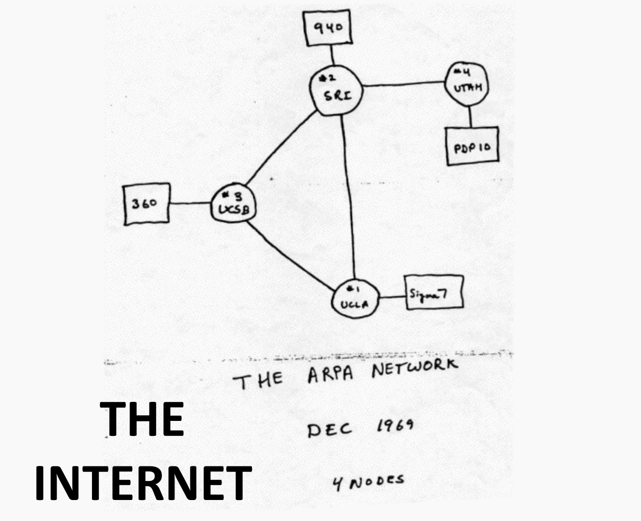
2. The Agency is authorized to arrange for the performance of research and development work by other agencies of Government, including the military departments, as may be necessary to accomplish its mission in relation to projects assigned.

February 7, 1958
"The purpose of this directive is to provide within the Department of Defense an agency for the direction and performance of certain advanced research and development projects."

Role in S&T ecosystem

- **Create breakthrough, paradigm-shifting solutions.**
- **Accept and manage significant technology risk.**
- **Disrupt or massively accelerate technology roadmaps.**

Information Innovation

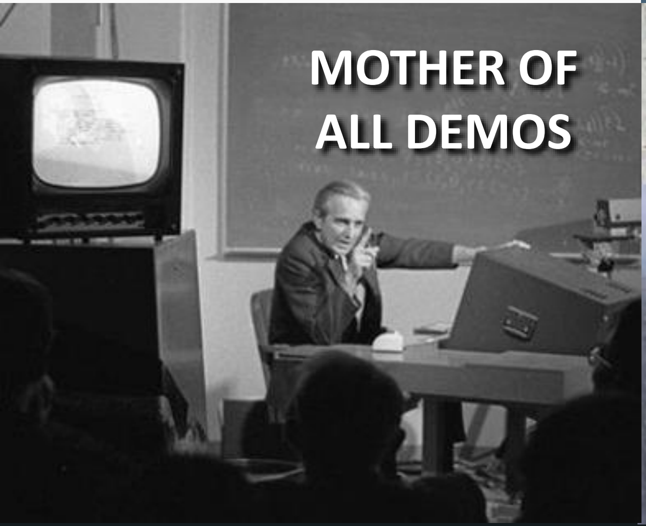


THE INTERNET

ADDRESS SPACE LAYOUT RANDOMIZATION

RED BALLOON CHALLENGE

CYBER GRAND CHALLENGE



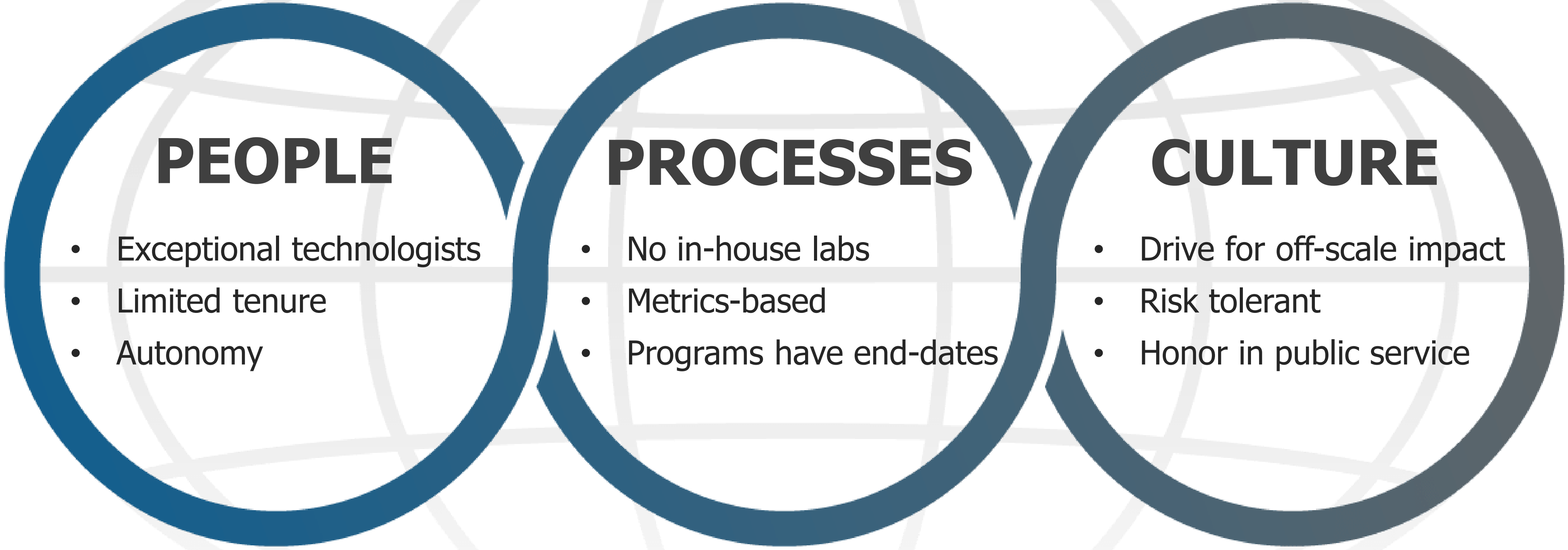
MOTHER OF ALL DEMOS

LANGUAGE TRANSLATION

AUTONOMOUS VEHICLES

PERSONALIZED ASSISTANT THAT LEARNS

Breakthrough technologies and capabilities for national security



PEOPLE

- Exceptional technologists
- Limited tenure
- Autonomy

PROCESSES

- No in-house labs
- Metrics-based
- Programs have end-dates

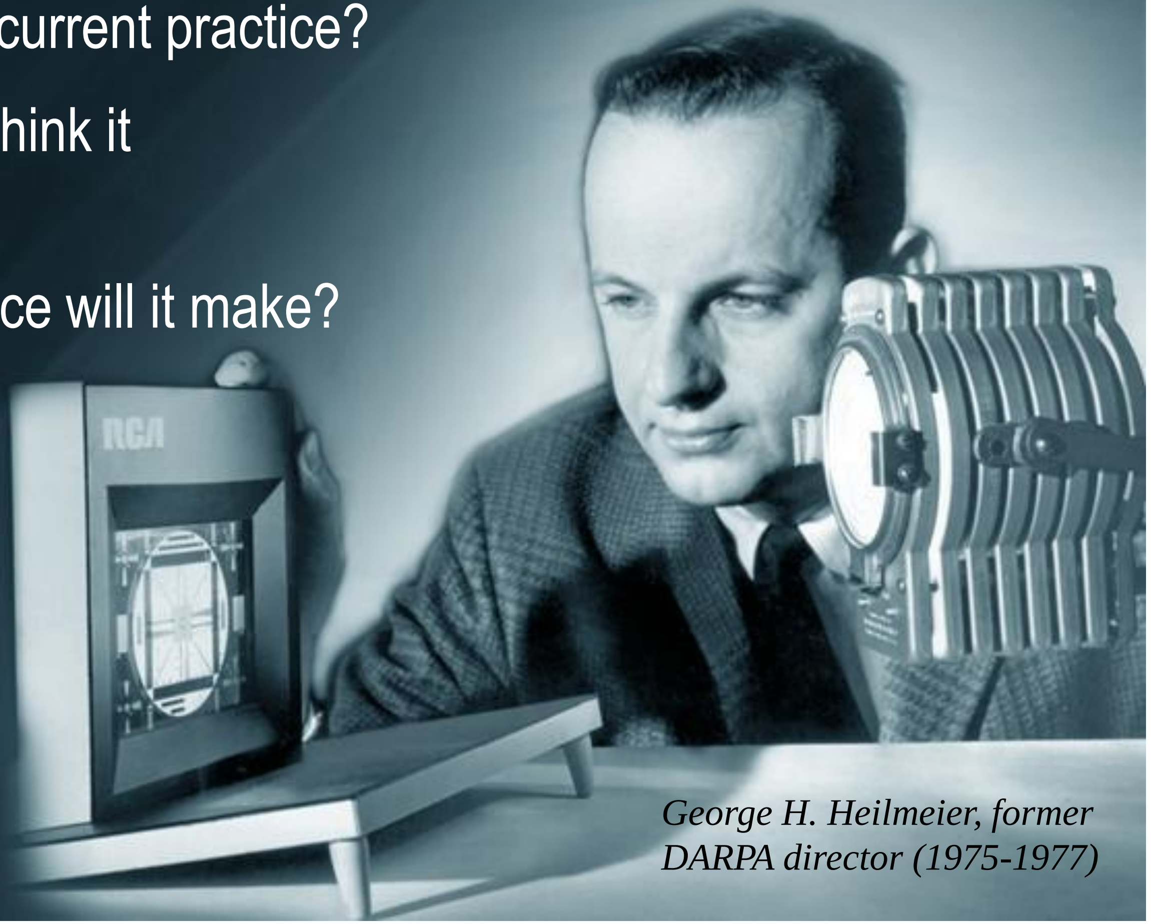
CULTURE

- Drive for off-scale impact
- Risk tolerant
- Honor in public service

DARPA's culture persists and the agency delivers

The Heilmeier Catechism

1. What are you trying to do?
   Articulate your objectives using absolutely no jargon.
2. How is it done today, and what are the limits of current practice?
3. What is new in your approach and why do you think it will be successful?
4. Who cares? If you are successful, what difference will it make?
5. What are the risks?
6. How much will it cost?
7. How long will it take?
8. What are the mid-term and final "exams" to check for success?



George H. Heilmeier, former DARPA director (1975-1977)

DARPA people deliver on the mission

“The sense of time ticking away is the heart of the whole thing. It is an impetus to venture into the unknown, to get people to put something forward, to build the prototype warts and all.”

“If you’re not taking enough risk, you don’t belong at DARPA.”

“This is not a culture of ‘no.’ It's a culture of getting things done.”

“Give bright, innovative people money to do something fast and furious, and then kick them out the door.”

“If you want a security blanket, DARPA is not for you. The blanket is ripped out of your hands four times a day.”

“If you don’t invent the Internet, you get a B.”

“Ordinary people think merely of spending time. DARPA people think of using it.”

“We look for someone technically strong with some project management experience, but especially someone who is a bit of a dreamer and not constrained by thinking ‘this we know to be true.’ It’s a rare combination of vision and practicality.”

“They want to use their significant technical skills to help the country.”

“We are mission oriented, not process oriented.”

“There are no marching orders. The only objective: create innovation.”

UNCLASSIFIED
Distribution Statement A: Approved for public release; distribution unlimited

DARPA

www.darpa.mil

I2O Office Wide Proposers Day Agenda

10:00	*11:00*	*Check-in and Networking Coffee*
11:00	11:05	Security Overview
11:05	11:15	Opening Remarks – Rob McHenry
11:15	11:35	How to Work with DARPA
		Commercial Strategy – Jen Thabet
		Small Business – Aaron Sparks
		DARPA Connect – Sana Sood
11:35	12:35	I2O Strategy – Kathleen Fisher
12:35	*1:35*	*Break for Lunch*
1:35	2:05	PM Presentations – (Dewhurst, Bernsen, Sweet, Kuhn, Cook)
2:05	2:15	Delivering on the DARPA Mission – Matt Turek
2:20	**3:55**	**Sidebars**

DARPA

www.darpa.mil