Labo d'applications en désinformation

Disinformation Applications Lab

We Can Use AI and Win Against Disinformation

Stéphane Gagnon, Ph.D.,
Associate Professor, Université du Québec

admin@gagnontech.org
https://gagnontech.org
https://disinform.app

Acknowledgement

Thanks to the entire *Disinformation Applications Laboratory* team for their contributions to the project.

Thanks to the organizations that contributed to the project funding:

- *Fonds de recherche du Québec (FRQ)*
- *Conseil de recherche en sciences humaines du Canada (CRSH)*
- *Human Centric Cybersecurity Partnership (HC2P)*
- *Association des universités francophones (AUF)*
- *Université du Québec en Outaouais (UQO)*

Prior publications

Part of this presentation was given in October 2024 in Dublin at the ISACA 2024 Europe Conference.

A webinar is available on YouTube, organized by the Human-Centric Cybersecurity Partnership | Canada.

Disinformation and Corruption as Threats to Digital Trust

An article was also presented in June 2025, which is a less technical overview.

Stéphane Gagnon, (2025), "Enabling Parliaments to Fight Disinformation and Corruption: Toward AI-Enabled Chief Reality Officers (CRO) as Extensions to the Digital Trust Ecosystem Framework", *4th Global Conference on Parliamentary Studies*, Athens, Greece, June 13, 2025

Article available for download: https://doi.org/10.6084/m9.figshare.29264951.v1

Biography

- Stéphane Gagnon, Ph.D.
- Associate Professor of Business Technology Management (BTM) at the Université du Québec in Canada
- Teaches and supervises students, primarily in the doctoral program in project management and IT
- Obtained his PhD in Business Administration in 2001 from UQAM (defense 9/11 at 2 p.m.)
- Published his research on applications in project management and IT in several sectors, including finance, healthcare, and public administration
- Current research focuses on the use of AI to combat disinformation and corruption
- Focuses on the development of new governance methods for resilience against hybrid threats, using LLMs and Knowledge Graphs
- Immediate objective is to identify EU partners to expand the already focused Canada-US project

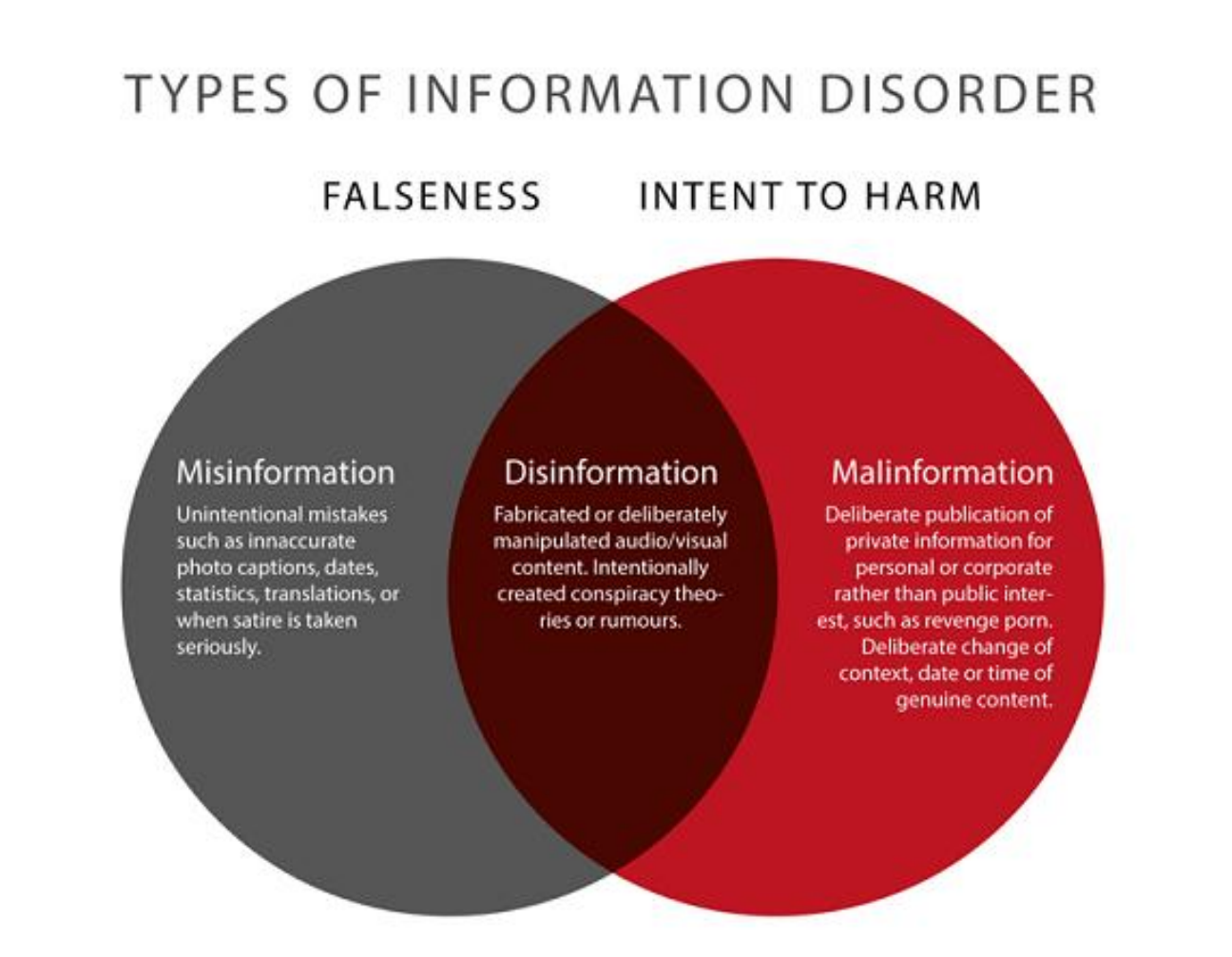
Agenda

- Welcome
- Part 1: Scoping Disinformation
- Part 2: Institutional Responses
- Part 3: Overall Architecture
- Part 4: AI and Knowledge Graphs
- Part 5: Disinformation Campaigns
- Part 6: National Strategy
- Closing

Part 1 - Scoping Disinformation

Delineating Scale of Attacks

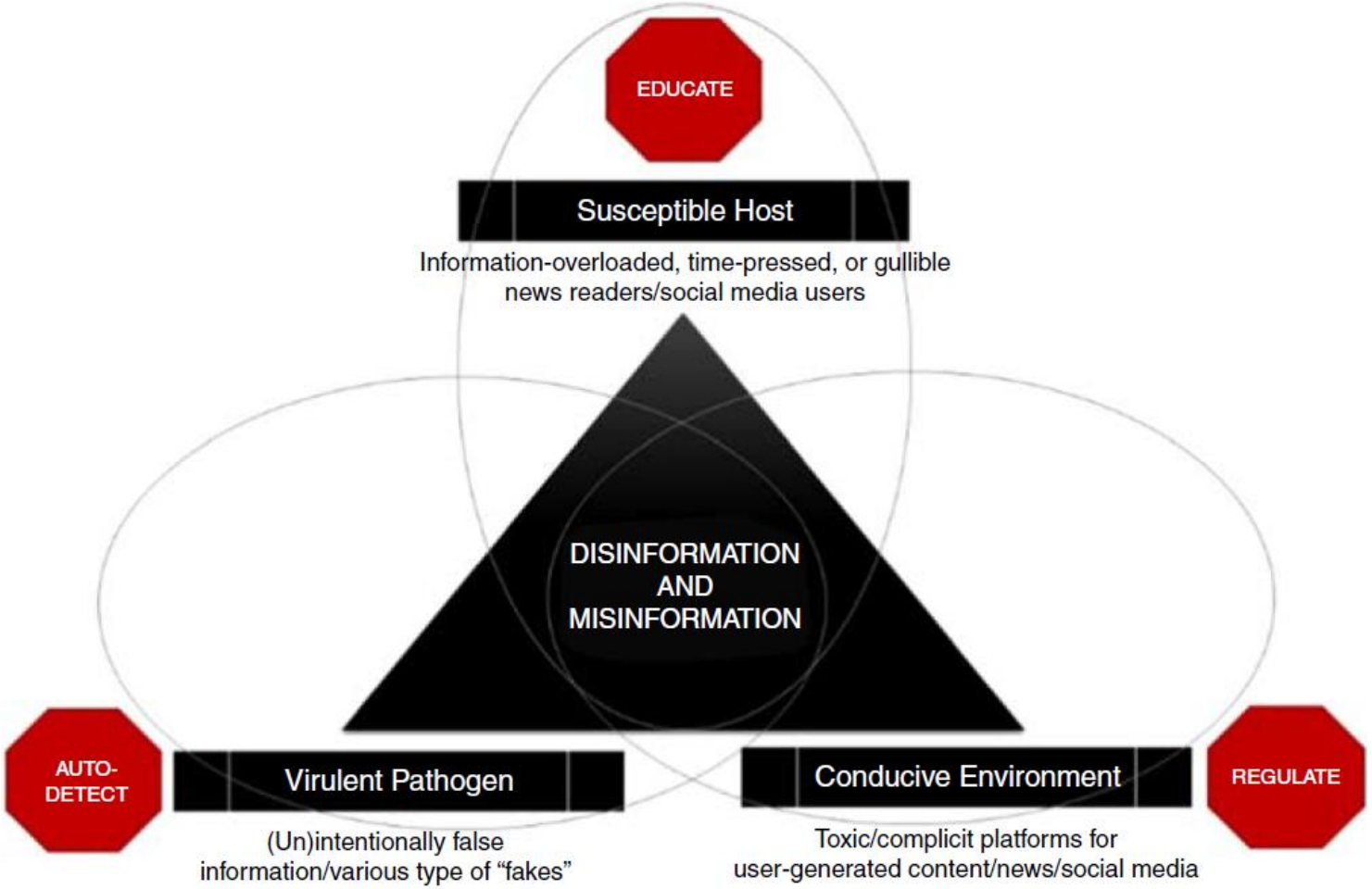
Mis-, Mal-, and Dis-information



Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

Disambiguation Requires Evidence

Trust Only Comes after Overcoming Doubt and Fear



Source: Rubin, V. L. (2019). Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *Journal of Documentation*, *75*(5), 1013–1034. https://doi.org/10.1108/JD-12-2018-0209

Part 2 - Institutional Responses

International Comparison

Fighting Disinformation and Foreign Interference in American, Canadian, and European Politics

- American, Canadian, and European political systems have been equally targeted by disinformation. Their policies are presently shifting, and it is likely their mutual international relations could be affected as a consequence.
- In 2016, the US FBI concluded that a massive foreign influence was directed in manipulation of electoral sentiment. The US Treasury later initiated sanctions against countries and entities, and the US Justice Department also carried out several investigations amd prosecution. In February 2025, the newly appointed US Attorney General ended all efforts in this area.
- In January 2025, the Parliament of Canada tabled a major report by the Commission on foreign interference in Canadian institutions. In April, the government was reelected, greatly aided by various threats made to Canadian sovereignty. The new cabinet has yet to issue new policies to address these challenges.
- Meanwhile, in 2020 and 2022, the EU Parliament formed committees to address disinformation campaigns. In 2023, recommendations were made in preparation for 2024 elections. Certain countries, like France, opted to have a national agency, the Viginum, to support EU efforts in fighting disinformation.
- This presentation will explore how the contrasted policies of these three democracies have yielded very different outcomes. We will open discussions as to what is the future of American, Canadian, and European democracy, and how to rebuild trust and diplomatic convergence.

Source: commissioningerenceetrangere.ca
foreigninterferencecommission.ca

Final Report on Foreign Interference – January 2025

Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions

1) Intelligence
2) The National Security and Intelligence Advisor to the Prime Minister
3) Clarifying coordination roles
4) Foreign interference strategy
5) Communications strategy
6) Awareness of the domestic online information environment
7) The Critical Election Incident Public Protocol and the Panel of Five
8) The Security and Intelligence Threats to Elections Task Force
9) Building trust with the public and stakeholders
10) Duty to warn
11) Parliamentarians
12) Political parties
13) Foreign embassies and consulates
14) International declaration
15) Inter-governmental cooperation
16) The RCMP
17) The intelligence-to-evidence challenge
18) Prohibitions
19) Third party political financing
20) Penalties
21) Navigating the information environment
22) Developing digital and media literacy
23) Protecting and promoting online information integrity

Source: commissioningerenceetrangere.ca
foreigninterferencecommission.ca

Final Report on Foreign Interference – January 2025

Commissioner Recommendations – Some Examples

11. The government should consider creating a government entity to monitor the domestic open source online information environment for misinformation and disinformation that could impact Canadian democratic processes.

17. There should be a single, highly visible and easily accessible point of contact or hotline for reporting foreign interference to the government, which is responsible for engaging the appropriate agency or department.

25. Members of Parliament, senators and their staff should be encouraged to check whether those with whom they interact are listed on the Foreign Influence and Transparency Registry.

32. The government should consider whether it would be appropriate to create a system of public funding for political parties.

44. The government should pursue discussions with media organizations and the public around modernizing media funding and economic models to support professional media, including local and foreign language media, while preserving media independence and neutrality.

Source: commissioningerenceetrangere.ca
foreigninterferencecommission.ca

Citizen Best Practices for Resilience – April 2025

Media Smarts Research Report – Based on Survey of 1000+ Citizens in Canada

1) **Difficulty with discernment:** Most participants struggled to distinguish true from false information, often relying on intuition or guesswork.

2) **Source reliability:** Participants were more likely to trust information if it came from well-known publications, experts or reliable friends.

3) **Lack of awareness of fact-checking tools:** Many participants thought fact-checking tools were hard to find, with most not knowing about relatively popular tools like Snopes.

4) **Misinformation paradox:** Despite believing they were good at spotting misinformation, participants felt overwhelmed by the fact-checking process and the majority struggled to tell if information was true or not.

5) **Visual misinformation (like deepfakes):** While just under half of participants said they believed they could identify AI-generated images online, many struggled to do so in the exercises, mistaking fake images for real ones.

6) **Older adults (55+):** Older adults were more likely to believe false information and were less confident in their ability to identify false content compared to younger participants.

7) **Sharing habits:** Most participants didn't regularly share content online, but those who did said they checked its accuracy before sharing.

8) ****Efficacy of BTF video interventions**:** Participants who watched any BTF video were slightly less likely to share a false image, and participants who watched the BTF video on how to fact-check were slightly more likely to 'look up' information to determine its accuracy.

Source : Motives and Methods: Building Resilience to Online Misinformation in Canada | MediaSmarts
Motivations et méthodes : Renforcer la résilience face à la désinformation en ligne au Canada | HabiloMédias

Citizen Best Practices for Resilience – April 2025

Media Smarts Research Report – Recommendations

1) **Treat visual misinformation as distinct:** Focus on visual misinformation separately from text-based misinformation.

2) **Positive messaging:** Reassure individuals that they don't need to be experts to identify visual misinformation, while acknowledging how it can feel overwhelming.

3) **Avoid 'hacks':** Don't rely on methods like zooming in on images or measuring the amount of blinking in a video, as these 'tells' can become outdated.

4) **Video accessibility:** Keep videos under 60 seconds, use clear language and focus on one key message, especially for older adults.

5) **Relatable scenarios:** Share personal stories and real examples to help individuals, particularly older adults, feel reassured and not alone in the fact-checking process.

6) ****Encourage intellectual humility**:** Prompt people to reconsider their ability to distinguish true from false information and acknowledge their own biases and limitations.

7) **Address the misinformation paradox:** Recognize that while people often think they're good at telling what's true online, they still express overwhelm and limited knowledge when it comes to fact-checking.

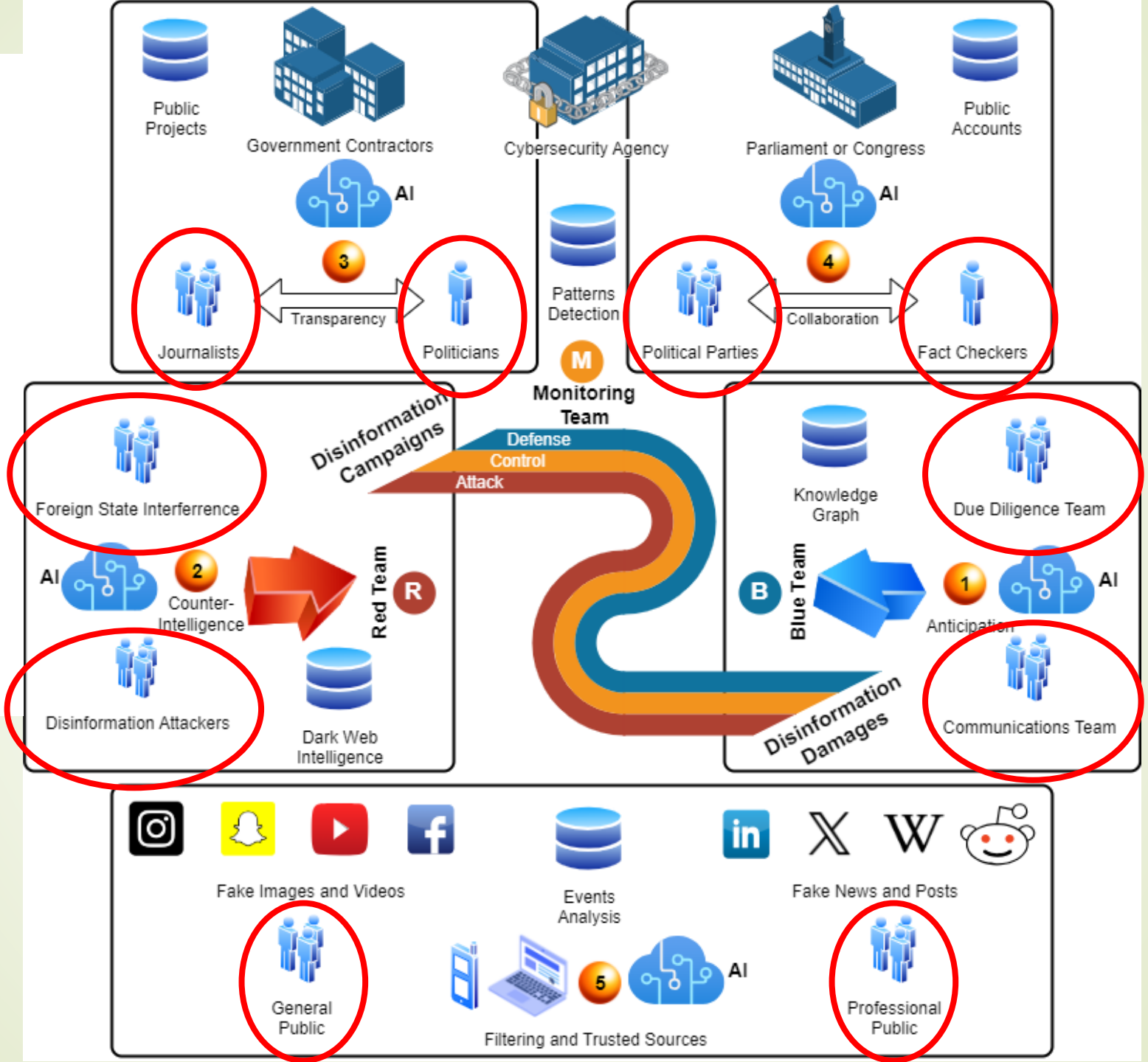
8) **Teach information triage:** Stress that not every piece of information individuals come across online needs to be fact-checked. Instead, people can prioritize what to fact-check based on its relevance, sense of importance and urgency.

Source : Motives and Methods: Building Resilience to Online Misinformation in Canada | MediaSmarts
Motivations et méthodes : Renforcer la résilience face à la désinformation en ligne au Canada | HabiloMédias

Part 3 - Overall Architecture

Disinformation and Corruption Allegations

Attack-Defense-Control Flows Within the Digital Ecosystem



https://disinform.app/about/

Part 4 - AI and Knowledge Graphs

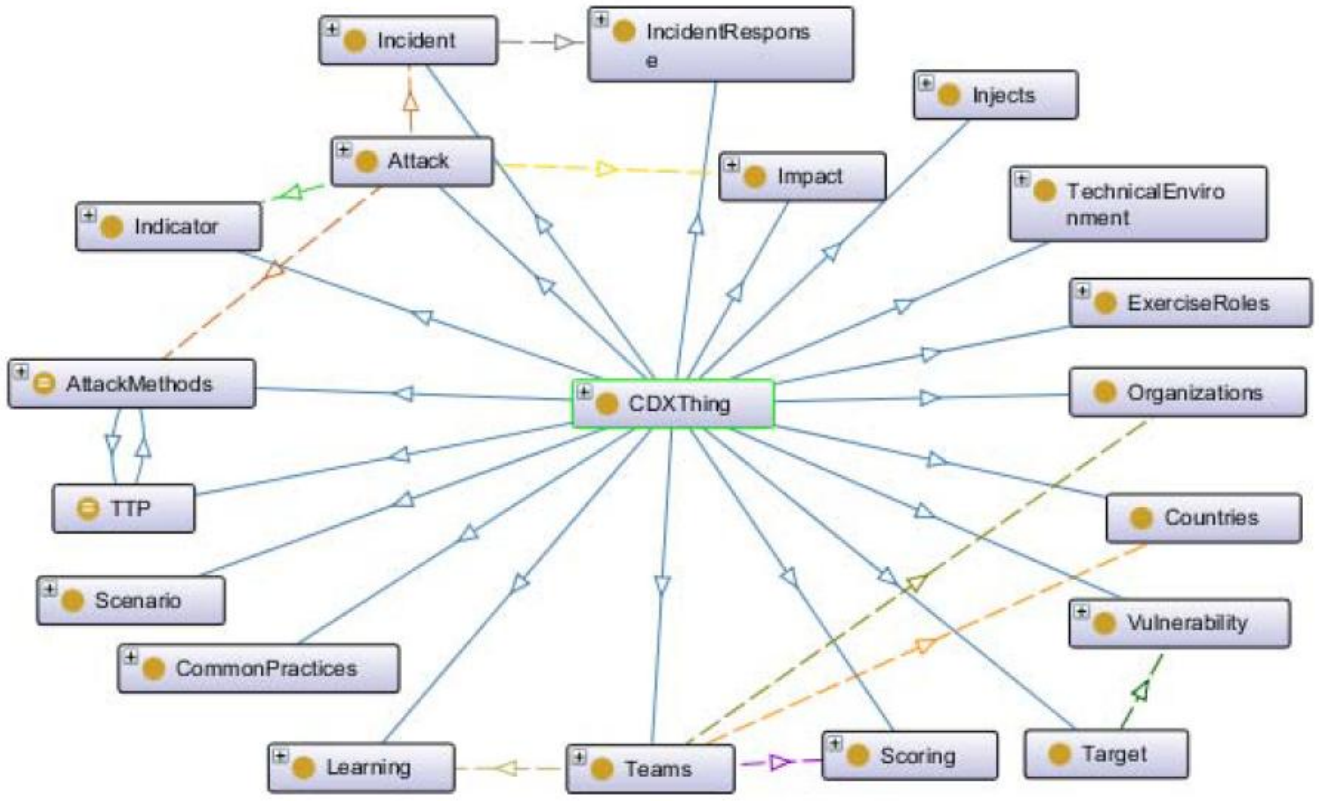


Figure 3. Meta-level CDX Ontology

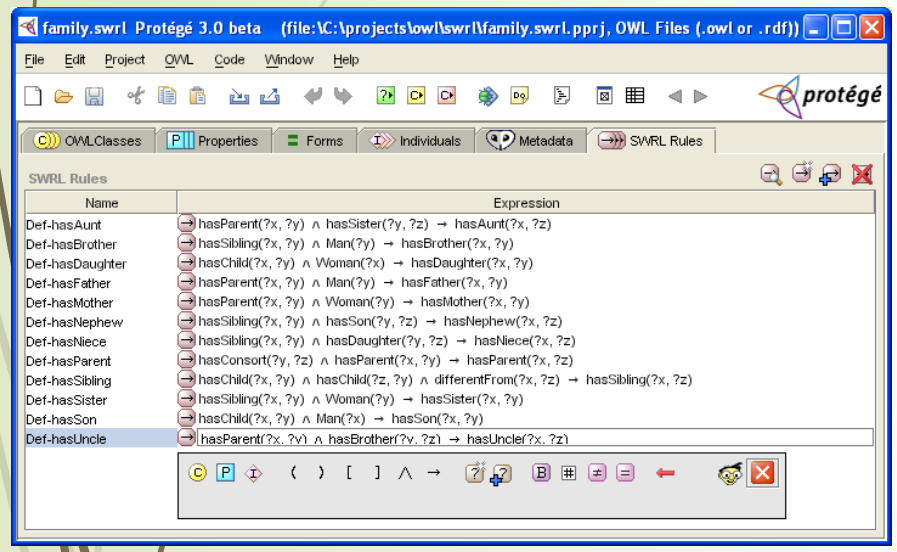
Cyber Defence Exercises (CDX) Ontology

Source: Babayeva, G., Maennel, K., & Maennel. O. M. (2022). Building an Ontology for Cyber Defence Exercises. *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 423–432. https://doi.org/10.1109/EuroSPW55150.2022.00050

Ontology and Rules within Knowledge Graph (KG)

Protégé

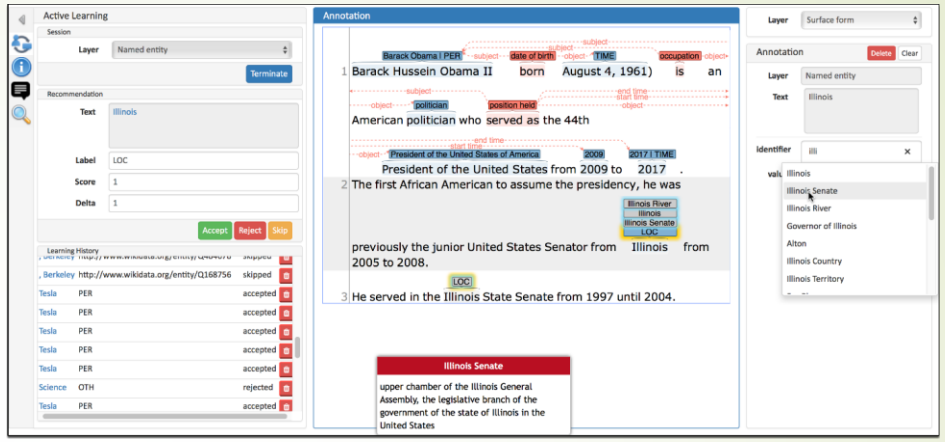
Ontology and Annotation



https://protege.stanford.edu

Ontology and Rules Integration

Inception

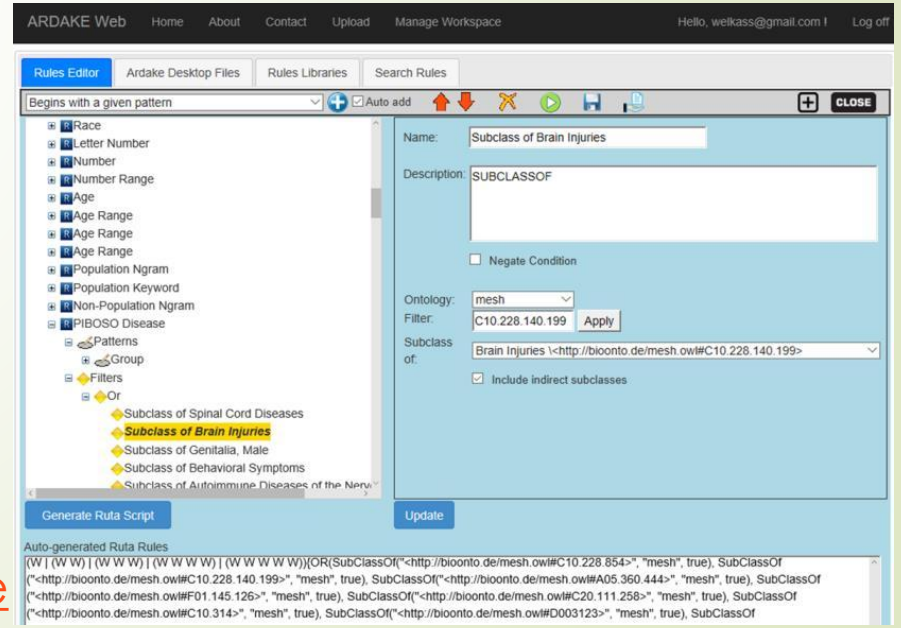


https://inception-project.github.io

Rules Extraction

ARDAKE

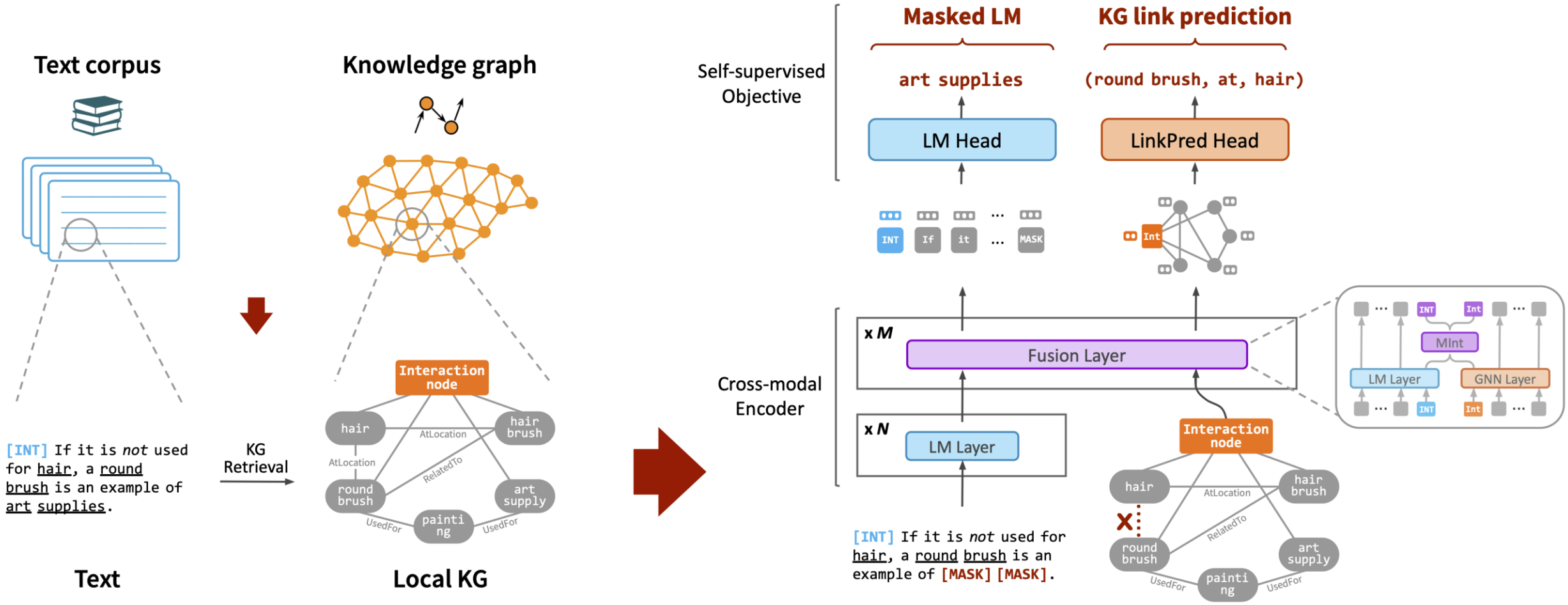
Ontology development, semantic annotation, and rules-driven annotation platforms for text disambiguation and storyboarding extraction



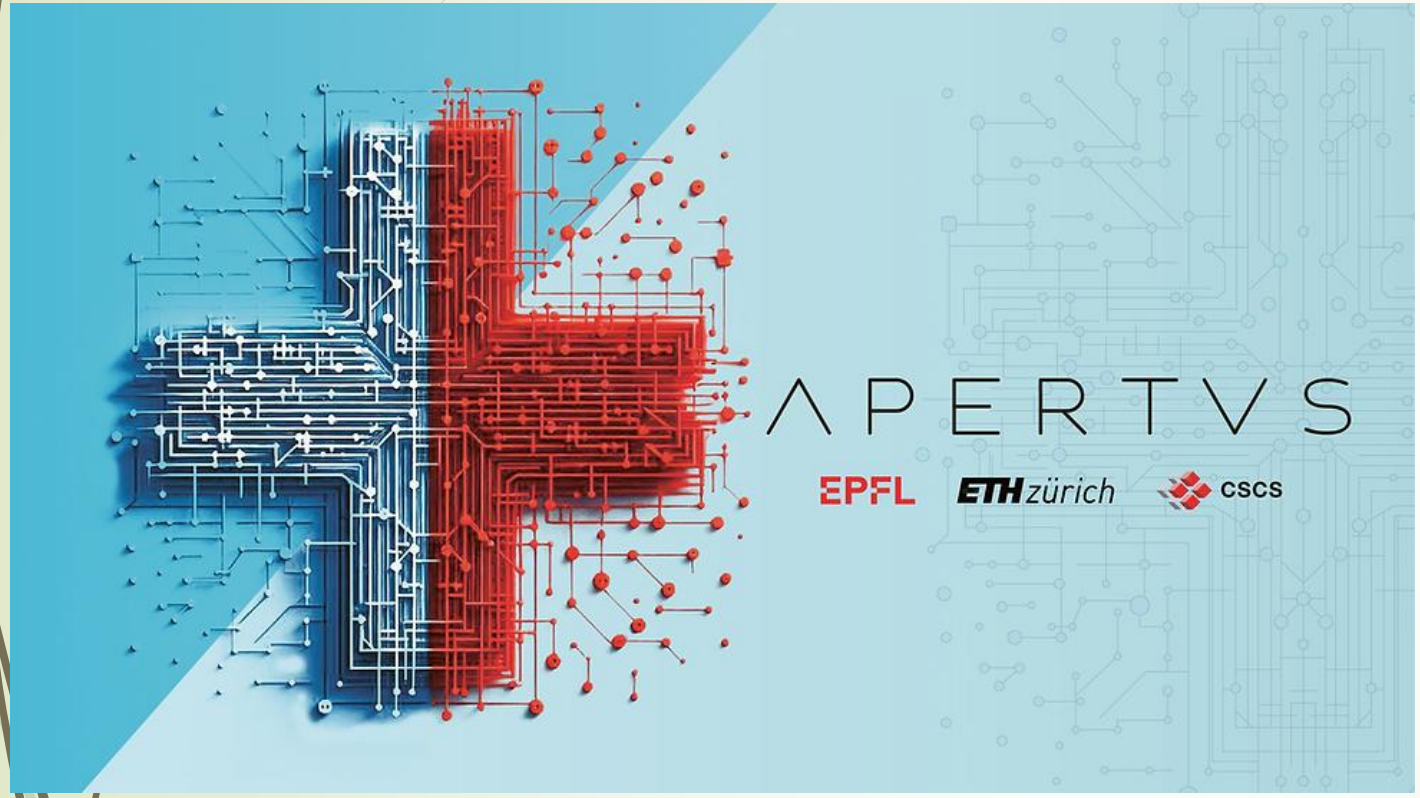
https://gagnontech.org/ardake

Ontology and KG Integrated GPT AI

© Michihiro Yasunaga, et al., (2022), "DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining" (NeurIPS Conference), https://github.com/michiyasunaga/dragon



Apertus as Development Platform

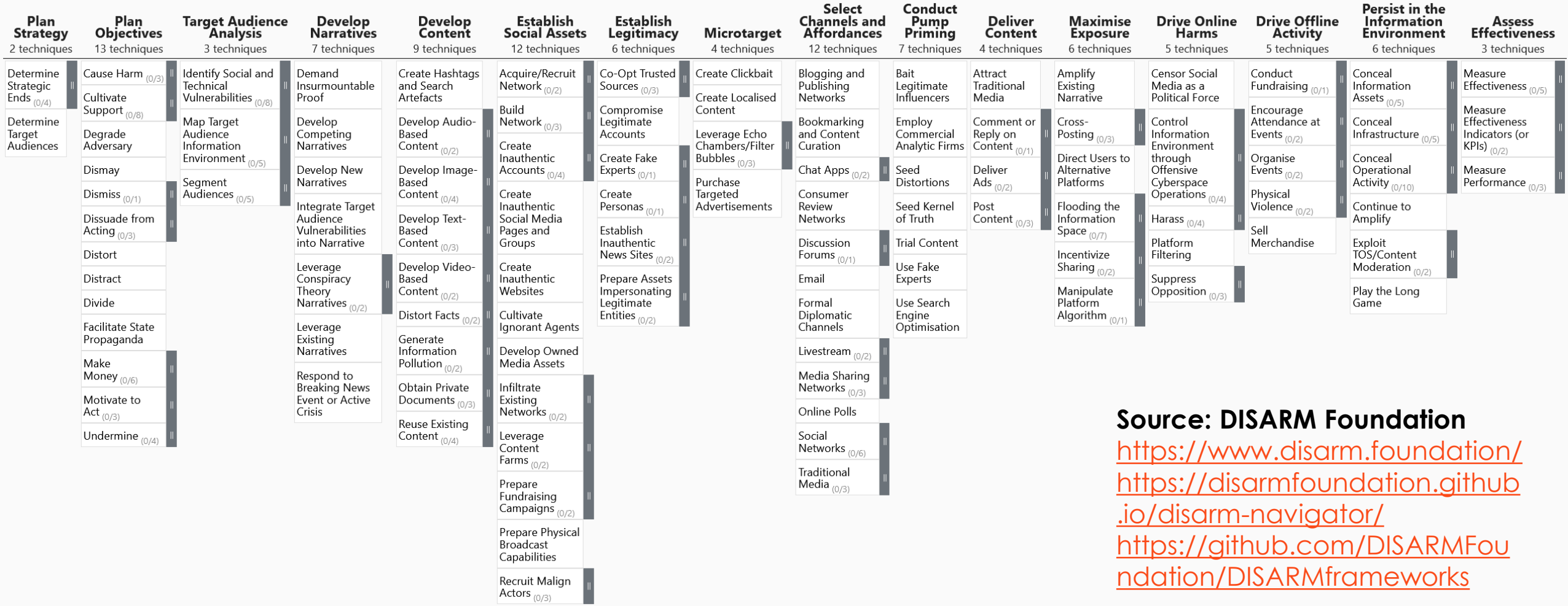


© Alejandro Hernández-Cano, et al., (2025), "Apertus: Democratizing Open and Compliant LLMs for Global Language Environments", Swiss National Supercomputing Centre (CSCS), https://www.swiss-ai.org/apertus, https://arxiv.org/abs/2509.14233

Part 5 - Disinformation Campaigns

Modeling Disinformation Campaigns

Anticipate Next Steps to Reinforce Disambiguation



Source: DISARM Foundation
https://www.disarm.foundation/
https://disarmfoundation.github.io/disarm-navigator/
https://github.com/DISARMFoundation/DISARMframeworks

DISARM Framework

Reference Model in Fighting Disinformation "Campaigns"

1. Phases: higher-level groupings of tactics, created so we could check we didn't miss anything
2. Tactics: stages that someone running a misinformation incident is likely to use
3. Techniques: activities that might be seen at each stage
4. Tasks: things that need to be done at each stage. In Pablospeak, tasks are things you do, techniques are how you do them.
5. Counters: countermeasures to DISARM Tactics, Techniques, and Procedures (TTPs).
6. Actor Types: resources needed to run countermeasures
7. Response types: the course-of-action categories we used to create counters
8. Metatechniques: a higher-level grouping for countermeasures
9. Incidents: incident descriptions used to create the DISARM frameworks

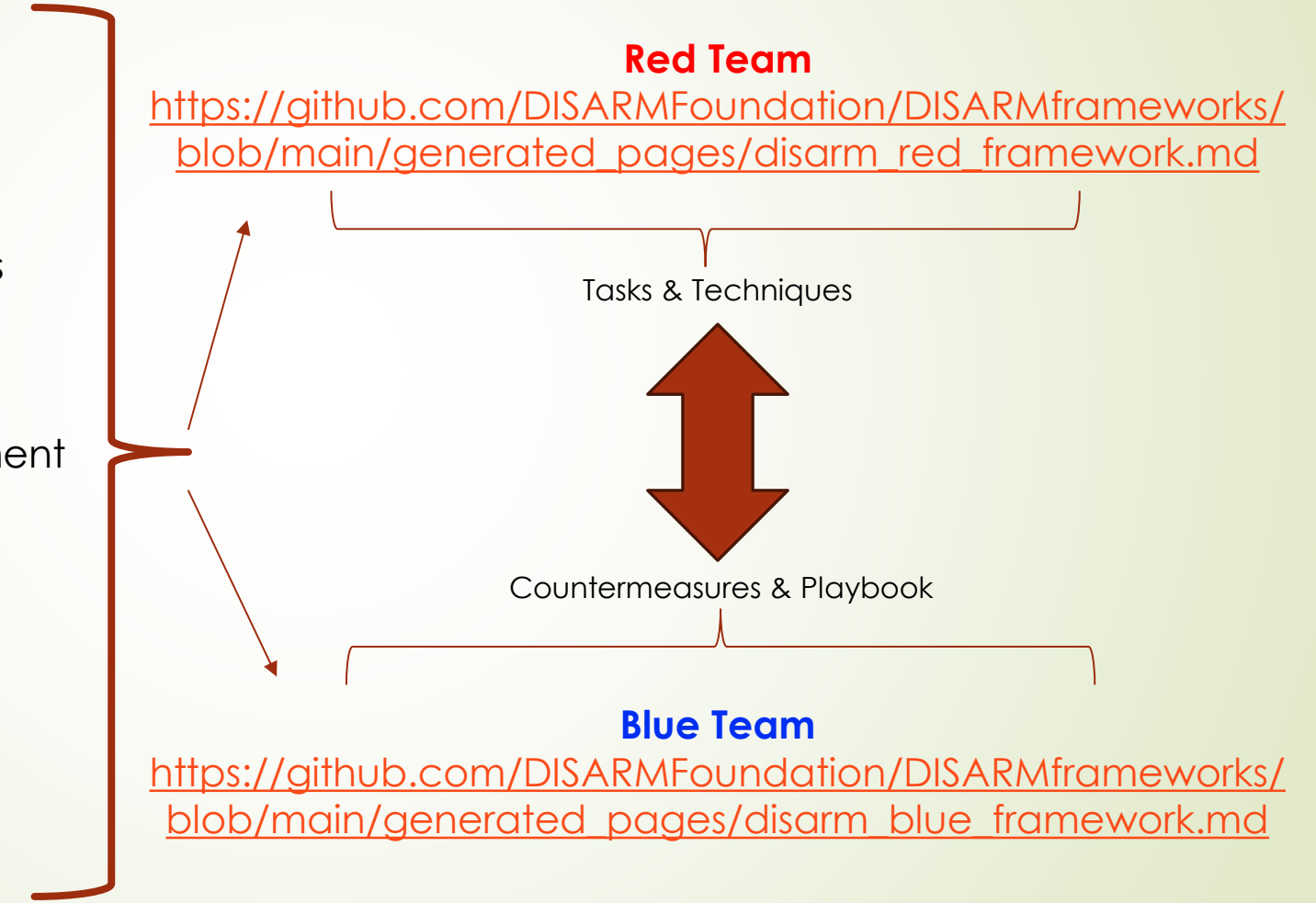
Source: https://github.com/DISARMFoundation/DISARMframeworks

DISARM Framework

Red vs Blue Teams

Tactics

TA01 Plan Strategy
TA02 Plan Objectives
TA05 Microtarget
TA06 Develop Content
TA07 Select Channels and Affordances
TA08 Conduct Pump Priming
TA09 Deliver Content
TA10 Drive Offline Activity
TA11 Persist in the Information Environment
TA12 Assess Effectiveness
TA13 Target Audience Analysis
TA14 Develop Narratives
TA15 Establish Social Assets
TA16 Establish Legitimacy
TA17 Maximise Exposure
TA18 Drive Online Harms



https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/tactics_index.md

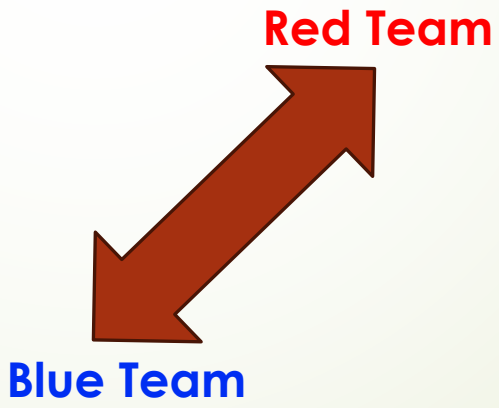
DISARM Framework

Attack Techniques vs Counters: ***Example*** - Tactic TA02: Plan Objectives

Summary: Set clearly defined, measurable, and achievable objectives. In some cases, achieving objectives ties execution of tactical tasks to reaching the desired strategic end state. In other cases, where there is no clearly defined strategic end state, the tactical objective may stand on its own. The objective statement should not specify the way and means of

Counters	Response types
C00009 Educate high profile influencers on best practices	D02
C00011 Media literacy. Games to identify fake news	D02
C00070 Block access to disinformation resources	D02
C00028 Make information provenance available	D03
C00029 Create fake website to issue counter narrative and counter narrative through physical merchandise	D03
C00030 Develop a compelling counter narrative (truth based)	D03
C00031 Dilute the core narrative - create multiple permutations, target / amplify	D03
C00060 Legal action against for-profit engagement factories	D03
C00156 Better tell your country or organisation story	D03
C00164 compatriot policy	D03
C00169 develop a creative content hub	D03
C00222 Tabletop simulations	D03
C00144 Buy out troll farm employees / offer them jobs	D04
C00092 Establish a truth teller reputation score for influencers	D07
C00207 Run a competing disinformation campaign - not recommended	D07

Tasks
TK0004 Identify target subgroups
TK0005 Analyse subgroups
TK0006 create master narratives
TK0007 Decide on techniques (4Ds etc)
TK0008 Create subnarratives
TK0009 4chan/8chan coordinating content
TK0032 OPSEC for TA02



Techniques
T0002 Facilitate State Propaganda
T0066 Degrade Adversary
T0075 Dismiss
T0075.001 Discredit Credible Sources
T0076 Distort
T0077 Distract
T0078 Dismay
T0079 Divide
T0135 Undermine
T0135.001 Smear
T0135.002 Thwart
T0135.003 Subvert
T0135.004 Polarise
T0136 Cultivate Support
T0136.001 Defend Reputaton
T0136.002 Justify Action
T0136.003 Energise Supporters
T0136.004 Boost Reputation
T0136.005 Cultvate Support for Initiative
T0136.006 Cultivate Support for Ally
T0136.007 Recruit Members
T0136.008 Increase Prestige
T0137 Make Money
T0137.001 Generate Ad Revenue
T0137.002 Scam
T0137.003 Raise Funds
T0137.004 Sell Items under False Pretences
T0137.005 Extort
T0137.006 Manipulate Stocks
T0138 Motivate to Act
T0138.001 Encourage
T0138.002 Provoke
T0138.003 Compel
T0139 Dissuade from Acting
T0139.001 Discourage
T0139.002 Silence
T0139.003 Deter
T0140 Cause Harm
T0140.001 Defame
T0140.002 Intimidate
T0140.003 Spread Hate

https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/tactics/TA02.md

Part 6 - National Strategy

National Strategy – A “Chief Reality Officer” (CRO) for Every Institution

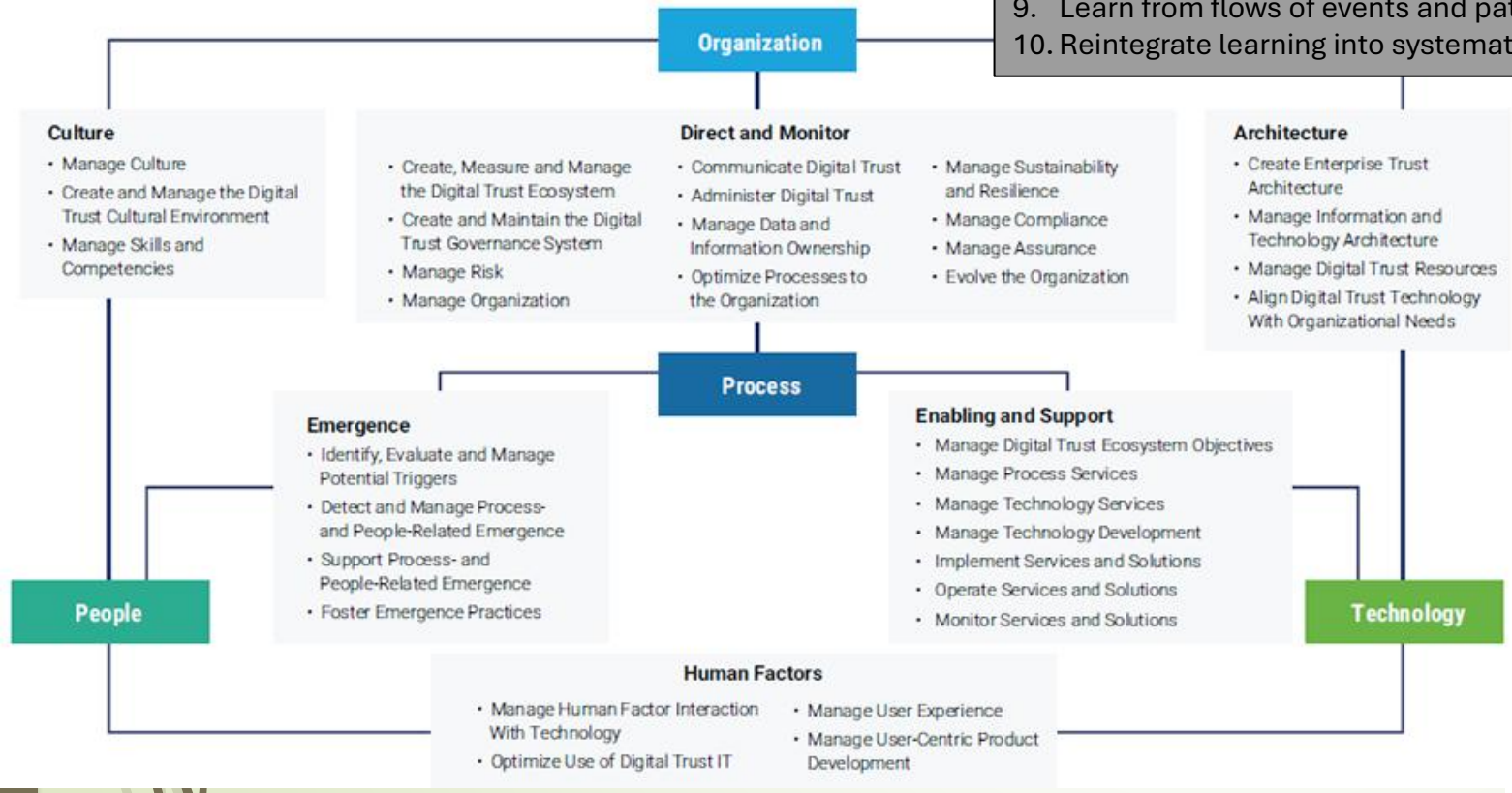
- Disinformation (false information disseminated to cause harm) takes many forms, e.g., fake news, audio-video impersonations, data falsification, etc.
- Elected officials are often the primary targets of disinformation campaigns, especially accusations of fraud or corruption, which can disrupt the public sector projects.
- However, when corruption allegations against politicians are proven true, they must be taken seriously and require immediate in-depth analysis.
- We argue that a model of "parliament accountability" is the best institutional arrangement to fight Disinformation and Corruption within an integrated strategy.
- We propose that parliaments develop innovative practices of creating an AI-Enabled "Chief Reality Officers" (CRO) as independent agency reporting to them directly.
- The mandate of a CRO would be to fight Disinformation and Corruption in partnership with businesses and governments.
- It would rely on AI to ensure the highet level of reliability and transparency in constant vigilance, due diligence, and collaboration in open-source intelligence (OSINT).
- As a framework to help guide these new CRO capabilities, we propose an extension to ISACA's Digital Trust Ecosystem Framework (DTEF), or the "Trust Framework".
- The framework can serve as the "glue" to bind all other frameworks covering the internal, external, ecosystem, as well as the public space of organizations.

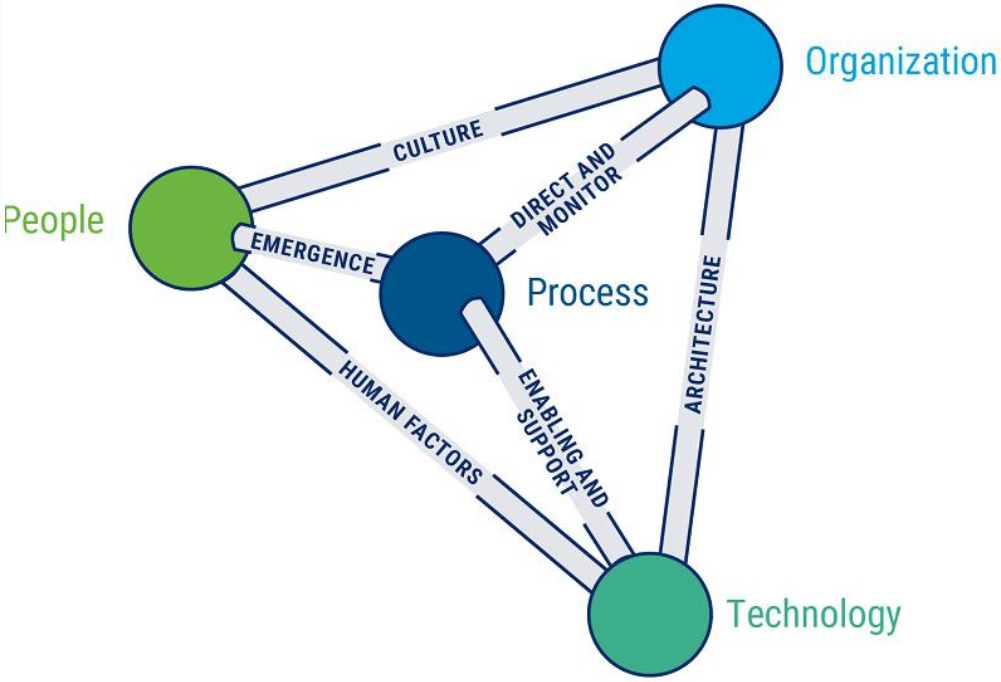
Digital Trust Components

Communit
y

Communicate and Control

1. Define threats, identify their sources
2. Share open-source intelligence within trusted ecosystems
3. Monitor threat emergence and diffusion
4. Correlate events and messages surrounding allegations
5. Identify disinformation campaigns patterns
6. Use AI to disambiguate false and true allegations
7. Link disinformation to countermeasures
8. Link true allegations to corruption judiciary and ethical proceedings
9. Learn from flows of events and patterns
10. Reintegrate learning into systematic monitoring capabilities





ISACA. (2024). *Using the Digital Trust Ecosystem Framework to Achieve Trustworthy AI*. https://www.isaca.org/resources/white-papers/2024/using-dtef-to-achieve-trustworthy-ai

Source: ISACA, (2022), *Digital Trust Ecosystem Framework* https://www.isaca.org/digital-trust

Thank you!

Stéphane Gagnon, Ph.D.

Associate Professor

Université du Québec en Outaouais (UQO)

https://gagnontech.org/bio/

admin@gagnontech.org

https://disinform.app