Labo d'applications en désinformation

Disinformation Applications Lab

Nous pouvons utiliser l'IA et vaincre la désinformation

Stéphane Gagnon, Ph.D.,
professeur agrégé, Université du Québec

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

Remerciements

Merci à toute l'équipe du *Laboratoire d'applications en désinformation* pour leurs contributions au projet.

Merci aux organismes ayant contribué au financement du projet

- *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)*

Publications antérieures

Une partie de cette présentation a été faite en octobre 2024 à Dublin au ISACA 2024 Europe Conference.

Un webinaire est disponible sur YouTube, organisé par Human-Centric Cybersecurity Partnership | Canada.

Disinformation and Corruption as Threats to Digital Trust

Un article a aussi été présenté en juin 2025, lequel est un survol moins technique.

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 à télécharger : https://doi.org/10.6084/m9.figshare.29264951.v1

Biographie

- Stéphane Gagnon, Ph. D.
- Professeur agrégé en gestion des technologies d'affaires (GTA) à l'Université du Québec au Canada
- Enseigne et supervise des étudiant.e.s, principalement aux doctorats en management de projet et en TI
- Obtention de son doctorat en administration des affaires en 2001 de l'UQAM (soutenance 9/11 à 14h)
- Publié ses recherches sur les applications en management de projet et des TI
- Dans plusieurs secteurs, notamment la finance, la santé et l'administration publique
- Recherches actuelles portent sur l'utilisation de l'IA pour lutter contre la désinformation et la corruption
- Focalise sur le développement de nouvelles méthodes de gouvernance pour la résilience face aux menaces hybrides, utilisant les LLM et Graphes de connaissances
- Objectif immédiat est d'identifier des partenaires EU pour élagrir let projet déjà axé Canada-USA

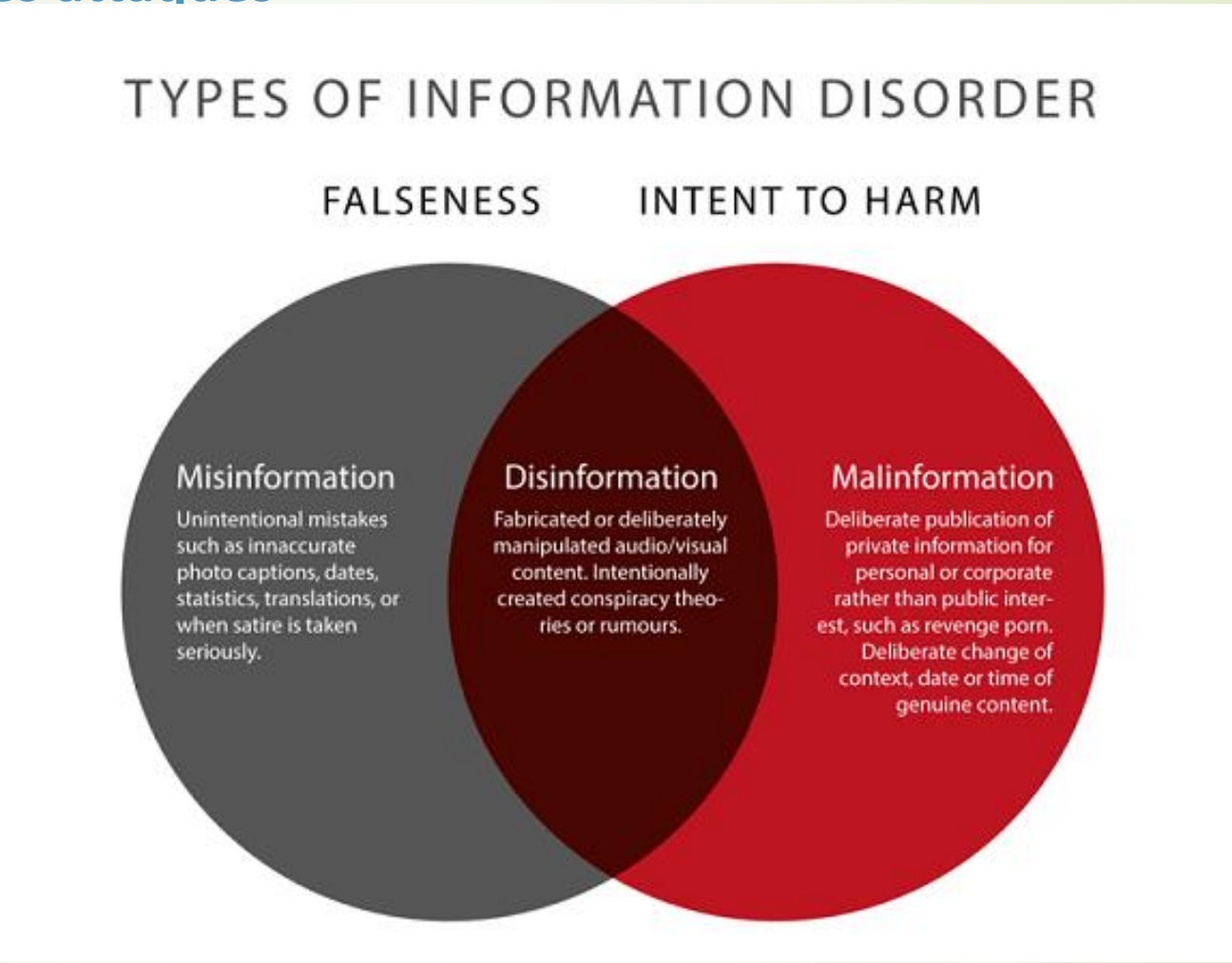
Plan

- Bienvenue
- Partie 1 : Cibler la désinformation
- Partie 2 : Réponses institutionnelles
- Partie 3 : Architecture globale
- Partie 4 : IA et graph de connaissances
- Partie 5 : Campagne de désinformation
- Partie 6 : Stratégie nationale
- Clôture

Partie 1 - Cibler la désinformation

Délimitation de l'échelle des attaques

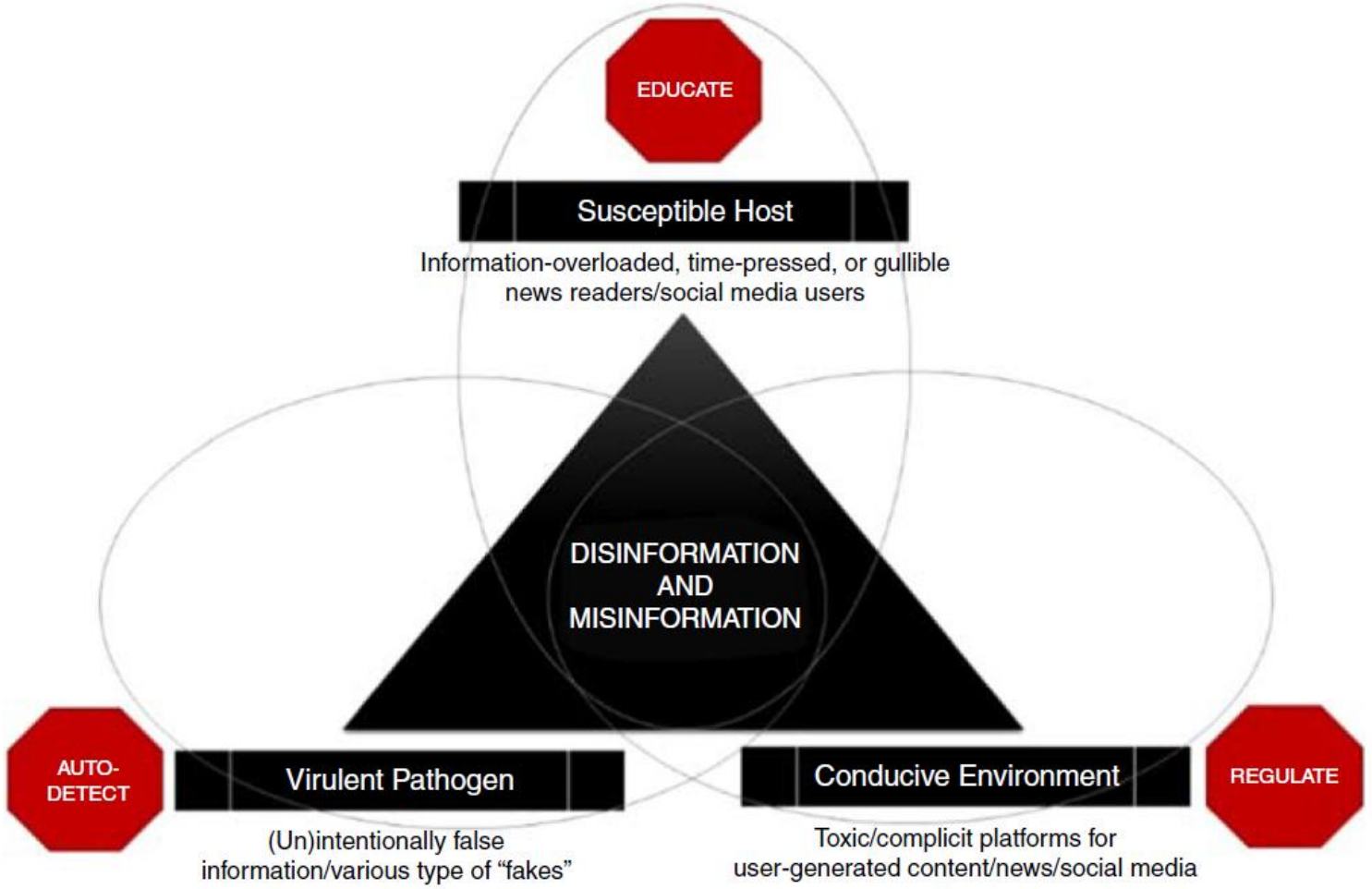
Mésinformation, malinformation



Wardle, C., & Derakhshan, H. (2017). Désordre informationnel : Vers un cadre interdisciplinaire pour la recherche et l'élaboration des politiques (vol. 27, pp. 1-107). Strasbourg : Conseil de l'Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

La désambiguïsation nécessite des preuves

La confiance ne vient qu'après avoir surmonté le doute et la peur



Source : Rubin, VL (2019). Désinformation et triangle de mésinformation : un modèle conceptuel pour l'épidémie de « fake news », ses facteurs causaux et ses interventions. *Journal of Documentation* , *75* (5), 1013–1034. https://doi.org/10.1108/JD-12-2018-0209

Partie 2 - Réponses institutionnelles

Comparaison internationale

Lutter contre la désinformation et l'ingérence étrangère dans les politiques américaines, canadiennes et européennes

- Les systèmes politiques américain, canadien et européen sont également la cible de la désinformation. Leurs politiques évoluent actuellement, et il est probable que leurs relations internationales mutuelles en soient affectées.
- En 2016, le FBI américain a conclu qu'une influence étrangère massive avait été exercée pour manipuler le sentiment électoral. Le Trésor américain a ensuite pris des sanctions contre des pays et des entités, et le ministère de la Justice américain a également mené plusieurs enquêtes et poursuites. En février 2025, le nouveau procureur général des États-Unis a mis fin à toute action dans ce domaine.
- En janvier 2025, le Parlement du Canada a déposé un important rapport de la Commission sur l'ingérence étrangère dans les institutions canadiennes . En avril, le gouvernement a été réélu, grandement aidé par diverses menaces à la souveraineté canadienne. Le nouveau cabinet n'a pas encore adopté de nouvelles politiques pour relever ces défis.
- Parallèlement, en 2020 et 2022, le Parlement européen a créé des commissions pour lutter contre les campagnes de désinformation. En 2023, des recommandations ont été formulées en vue des élections de 2024. Certains pays, comme la France, ont choisi de se doter d'une agence nationale, Viginum, pour soutenir les efforts de l'UE dans la lutte contre la désinformation.
- Cette présentation explorera comment les politiques contrastées de ces trois démocraties ont produit des résultats très différents. Nous ouvrirons des discussions sur l'avenir des démocraties américaine, canadienne et européenne, et sur les moyens de rétablir la confiance et la convergence diplomatique.

Rapport final sur l'ingérence étrangère – janvier 2025

Enquête publique sur l'ingérence étrangère dans les processus électoraux fédéraux et les institutions démocratiques

1) Intelligence
2) Le conseiller à la sécurité nationale et au renseignement du Premier ministre
3) Clarifier les rôles de coordination
4) Stratégie d'ingérence étrangère
5) Stratégie de communication
6) Connaissance de l'environnement national de l'information en ligne
7) Le protocole public relatif aux incidents électoraux critiques et le panel des cinq
8) Groupe de travail sur les menaces en matière de sécurité et de renseignement contre les élections
9) Instaurer la confiance avec le public et les parties prenantes
10) Devoir d'avertir
11) parlementaires
12) partis politiques
13) Ambassades et consulats étrangers
14) Déclaration internationale
15) Coopération intergouvernementale
16) La GRC
17) Le défi de la transformation du renseignement en preuve
18) Interdictions
19) Financement politique par des tiers
20) Pénalités
21) Naviguer dans l'environnement informationnel
22) Développer l'éducation numérique et médiatique
23) Protéger et promouvoir l'intégrité des informations en ligne

Source : commissioningerenceetrangere.ca

foreigninterferencecommission.ca

Rapport final sur l'ingérence étrangère – janvier 2025

Recommandations du commissaire – Quelques exemples

11. Le gouvernement devrait envisager de créer une entité gouvernementale chargée de surveiller l'environnement national d'information en ligne à source ouverte afin de détecter la désinformation et la mésinformation qui pourraient avoir une incidence sur les processus démocratiques canadiens.

17. Il devrait y avoir un point de contact ou une ligne directe unique, très visible et facilement accessible pour signaler les ingérences étrangères au gouvernement, qui est chargé de contacter l'agence ou le ministère approprié.

25. Les députés, les sénateurs et leur personnel devraient être encouragés à vérifier si les personnes avec lesquelles ils interagissent sont inscrites au Registre de l'influence étrangère et de la transparence.

32. Le gouvernement devrait examiner s'il serait approprié de créer un système de financement public pour les partis politiques.

44. Le gouvernement devrait poursuivre les discussions avec les organisations médiatiques et le public sur la modernisation du financement des médias et des modèles économiques afin de soutenir les médias professionnels, y compris les médias en langues locales et étrangères, tout en préservant l'indépendance et la neutralité des médias.

Source : commissioningerenceetrangere.ca

foreigninterferencecommission.ca

Bonnes pratiques citoyennes pour la résilience – avril 2025

Rapport de recherche Media Smarts – Basé sur un sondage mené auprès de plus de 1 000 citoyens au Canada

1) **Difficulté de discernement :** la plupart des participants ont eu du mal à distinguer les informations vraies des informations fausses, se fiant souvent à l'intuition ou aux suppositions.

2) **Fiabilité de la source :** les participants étaient plus susceptibles de faire confiance aux informations si elles provenaient de publications connues, d'experts ou d'amis fiables.

3) **Manque de connaissance des outils de vérification des faits :** De nombreux participants pensaient que les outils de vérification des faits étaient difficiles à trouver, la plupart ne connaissant pas des outils relativement populaires comme Snopes.

4) **Paradoxe de la désinformation :** bien qu'ils pensaient être doués pour repérer la désinformation, les participants se sont sentis dépassés par le processus de vérification des faits et la majorité a eu du mal à déterminer si l'information était vraie ou non.

5) **Désinformation visuelle (comme les deepfakes) :** alors qu'un peu moins de la moitié des participants ont déclaré qu'ils pensaient pouvoir identifier les images générées par l'IA en ligne, beaucoup ont eu du mal à le faire lors des exercices, confondant les fausses images avec les vraies.

6) **Adultes plus âgés (55 ans et plus) :** Les adultes plus âgés étaient plus susceptibles de croire aux fausses informations et avaient moins confiance en leur capacité à identifier les faux contenus par rapport aux participants plus jeunes.

7) **Habitudes de partage :** la plupart des participants ne partageaient pas régulièrement du contenu en ligne, mais ceux qui le faisaient ont déclaré vérifier son exactitude avant de le partager.

8) ****Efficacité des interventions vidéo BTF** :** Les participants qui ont regardé une vidéo BTF étaient légèrement moins susceptibles de partager une fausse image, et les participants qui ont regardé la vidéo BTF sur la façon de vérifier les faits étaient légèrement plus susceptibles de « rechercher » des informations pour déterminer leur exactitude.

Source : Motifs et méthodes : Renforcer la résilience face à la désinformation en ligne au Canada | HabiloMédias

Motivations et méthodes : Renforcer la résilience face à la désinformation en ligne au Canada | HabiloMédias

Bonnes pratiques citoyennes pour la résilience – avril 2025

Rapport de recherche Habilo Media – Recommandations

1) **Traitez la désinformation visuelle comme une chose distincte :** concentrez-vous sur la désinformation visuelle séparément de la désinformation textuelle.

2) **Message positif :** rassurez les individus en leur disant qu'ils n'ont pas besoin d'être des experts pour identifier la désinformation visuelle, tout en reconnaissant que cela peut sembler accablant.

3) **Évitez les « hacks » :** ne vous fiez pas à des méthodes telles que le zoom sur les images ou la mesure de la quantité de clignements dans une vidéo, car ces « indications » peuvent devenir obsolètes.

4) **Accessibilité des vidéos :** gardez les vidéos de moins de 60 secondes, utilisez un langage clair et concentrez-vous sur un message clé, en particulier pour les personnes âgées.

5) **Scénarios pertinents :** Partagez des histoires personnelles et des exemples réels pour aider les individus, en particulier les personnes âgées, à se sentir rassurés et non seuls dans le processus de vérification des faits.

6) ****Encouragez l'humilité intellectuelle** :** incitez les gens à reconsidérer leur capacité à distinguer les informations vraies des informations fausses et à reconnaître leurs propres préjugés et limites.

7) **Abordez le paradoxe de la désinformation :** reconnaissez que même si les gens pensent souvent être doués pour dire la vérité en ligne, ils expriment néanmoins un sentiment de dépassement et des connaissances limitées lorsqu'il s'agit de vérifier les faits.

8) **Enseignez le tri des informations :** insistez sur le fait que toutes les informations trouvées en ligne ne nécessitent pas forcément une vérification. Il est préférable de prioriser les informations à vérifier en fonction de leur pertinence, de leur importance et de leur urgence.

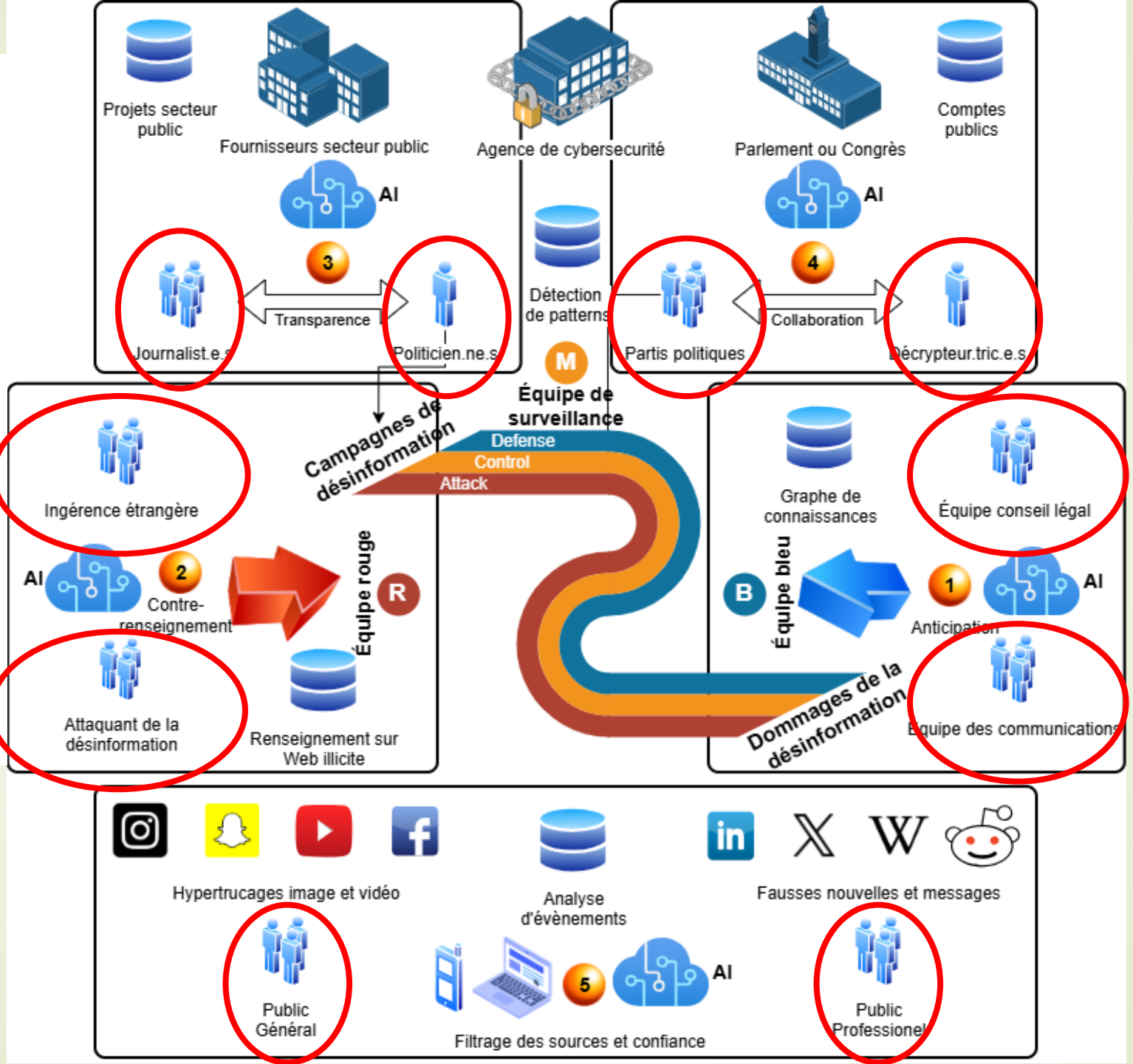
Source : Motifs et méthodes : Renforcer la résilience face à la désinformation en ligne au Canada | HabiloMédias

Motivations et méthodes : Renforcer la résilience face à la désinformation en ligne au Canada | HabiloMédias

Partie 3 - Architecture globale

Allégations de désinformation et de corruption

***Attaque** - **Défense** - **Contrôle** des flux au sein de l'écosystème numérique*



https://disinform.app/fr/a-propos/

Partie 4 - IA et graph de connaissances

Intégrer la désinformation dans les ontologies de cyberdéfense

Exemple : Ontologie des exercices de cyberdéfense (CDX)

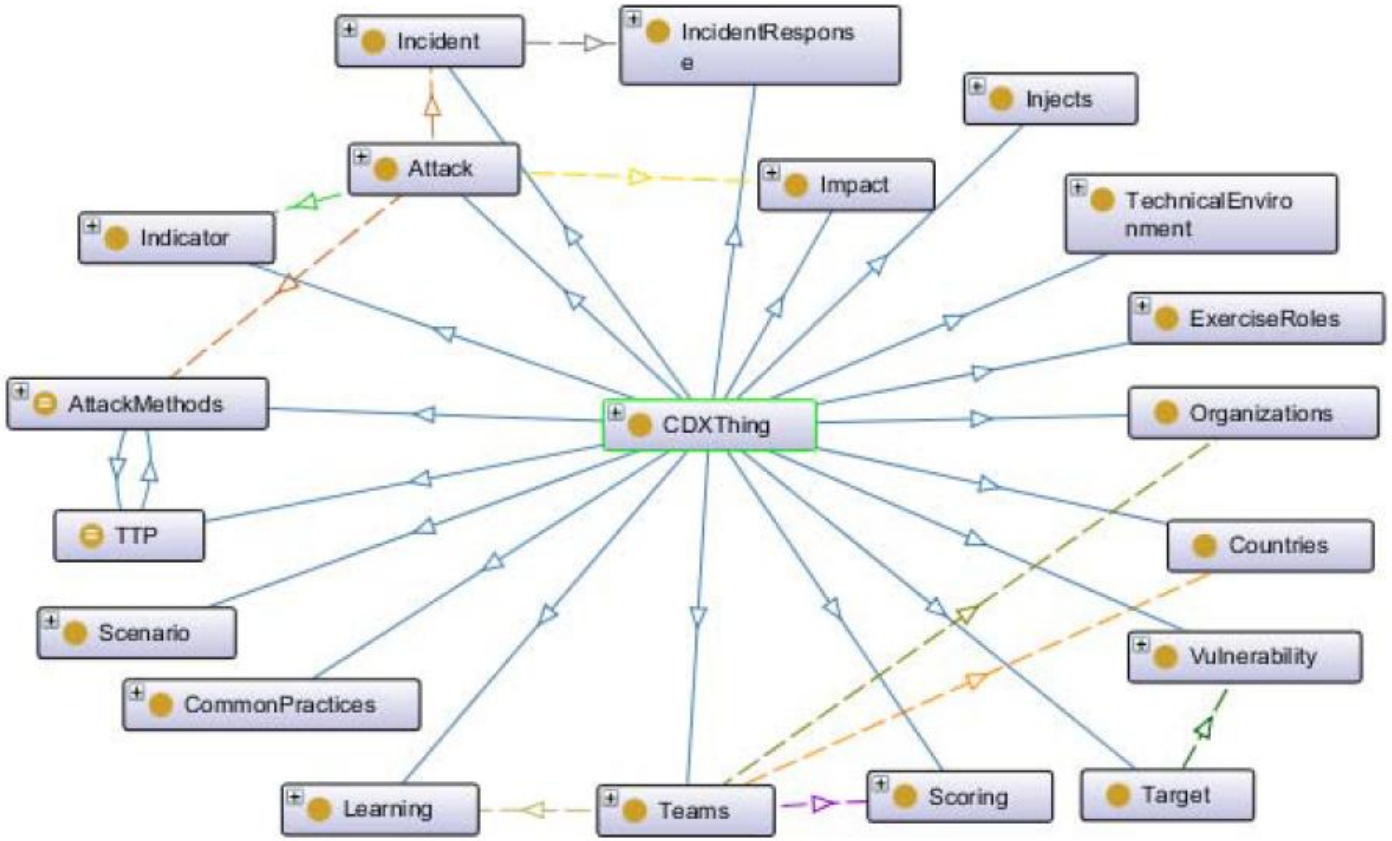


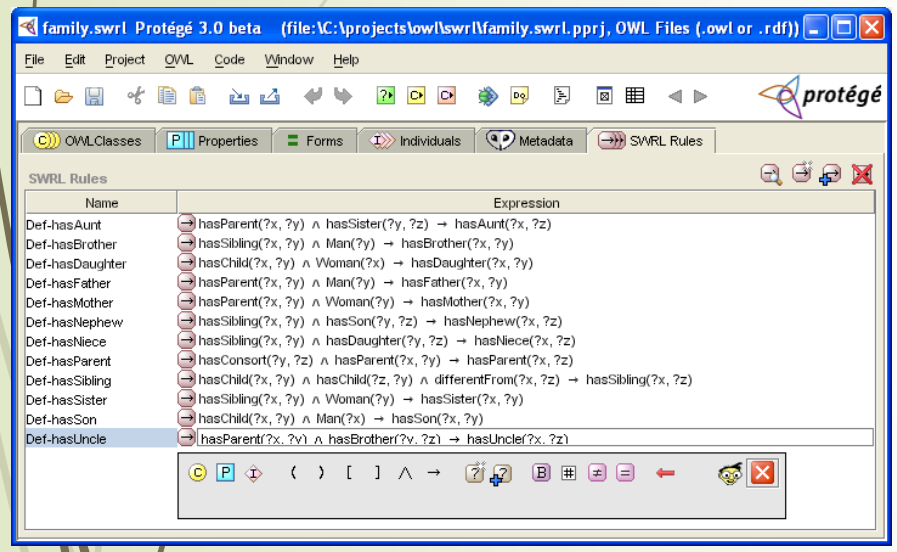
Figure 3. Meta-level CDX Ontology

Source : Babayeva, G., Maennel, K. et Maennel. OM (2022). Création d'une ontologie pour les exercices de cyberdéfense. *Ateliers du Symposium européen IEEE 2022 sur la sécurité et la confidentialité (EuroS&PW)* , p. 423–432. https://doi.org/10.1109/EuroSPW55150.2022.00050

Ontologie et règles au sein de Graphique de connaissances (KG)

Protégé

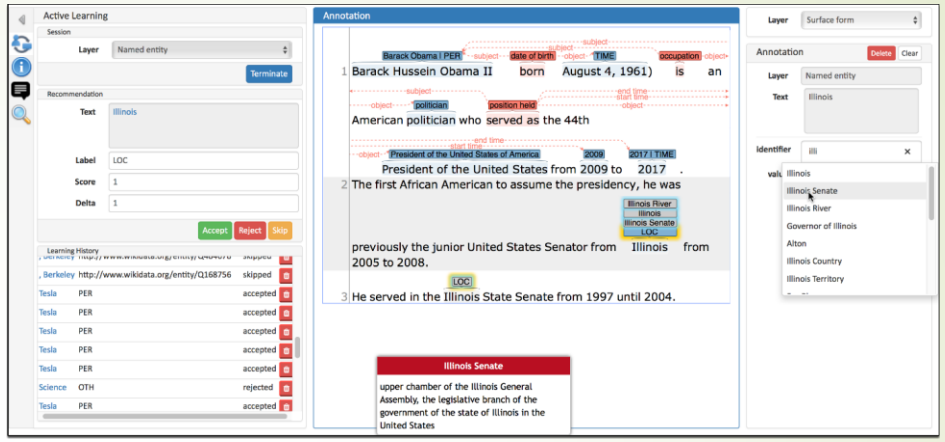
Ontologie et annotation



https://protege.stanford.edu

Intégration d'ontologies et de règles

Création

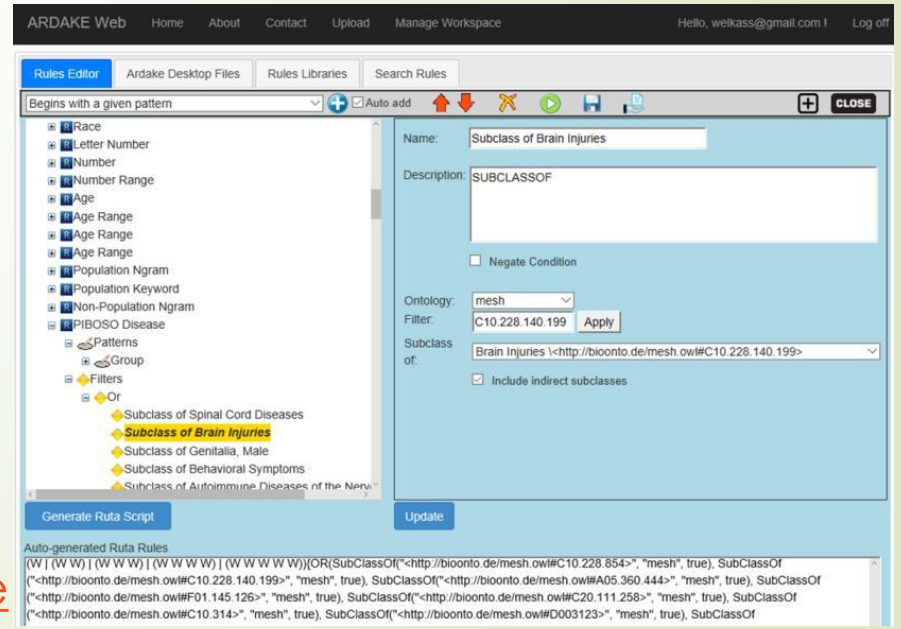


https://inception-project.github.io

Extraction de règles

ARDAKE

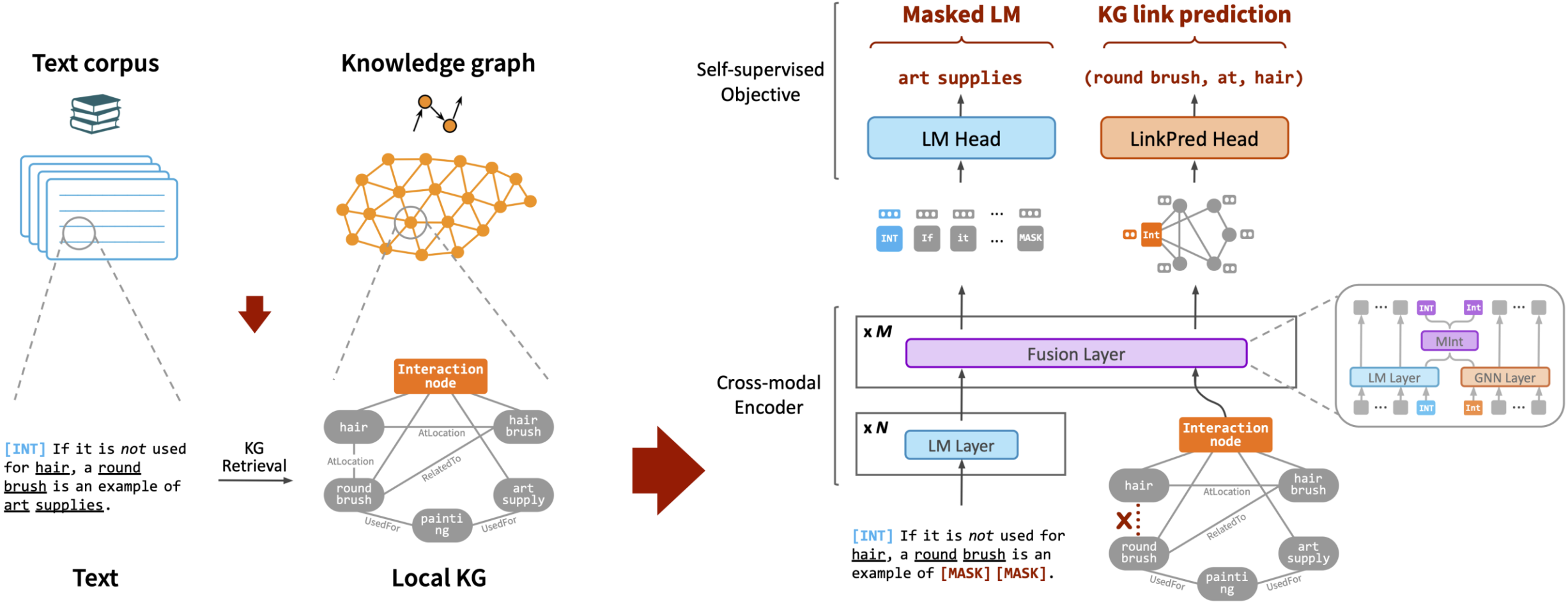
Développement d'ontologies, annotation sémantique et plateformes d'annotation pilotées par des règles pour la désambiguïsation de texte et l'extraction de storyboards



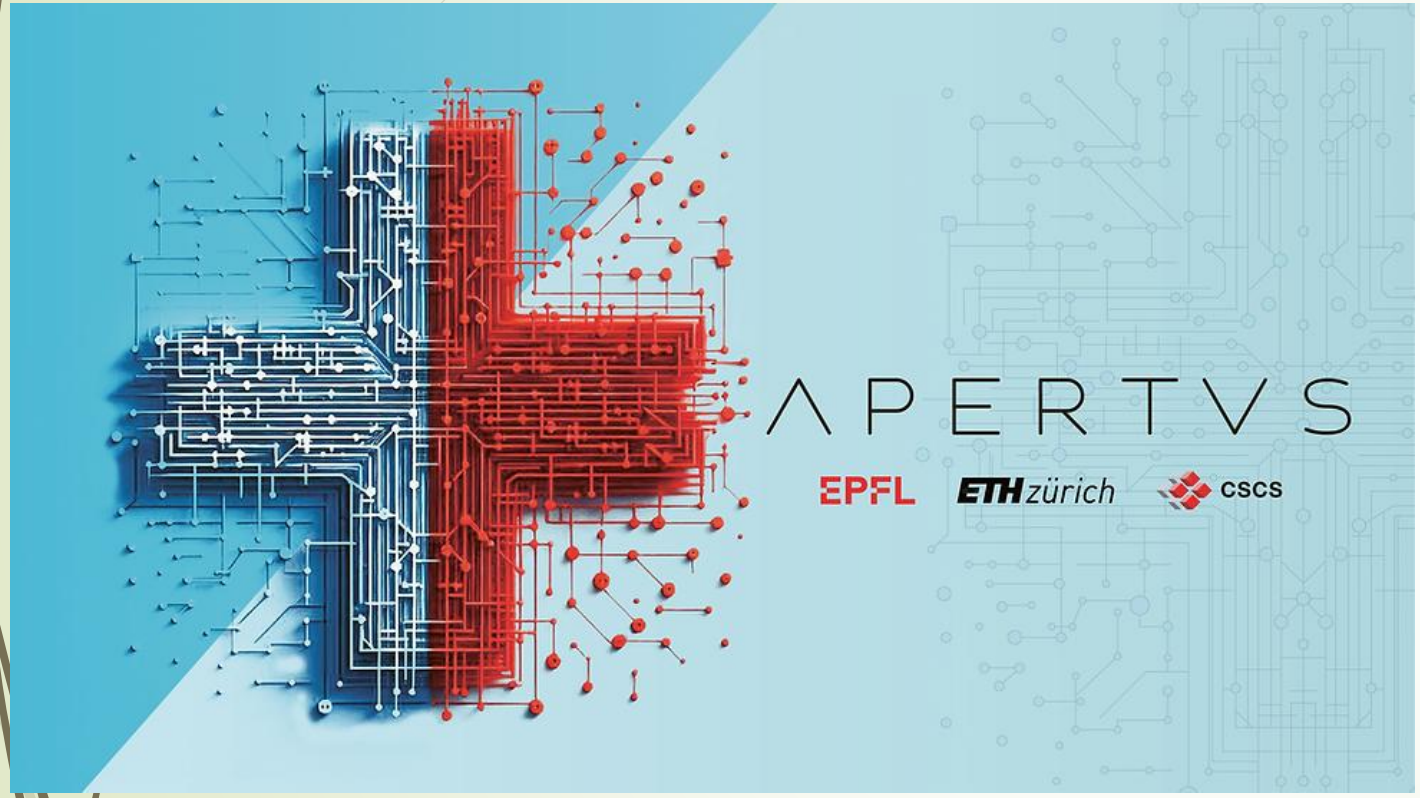
https://gagnontech.org/ardake

Ontologie et KG GPT AI intégrées

© Michihiro Yasunaga, et al., (2022), « DRAGON : Pré-entraînement bidirectionnel approfondi des graphes de connaissances linguistiques » (Conférence NeurIPS), https://github.com/michiyasunaga/dragon



Apertus comme plateforme de développement

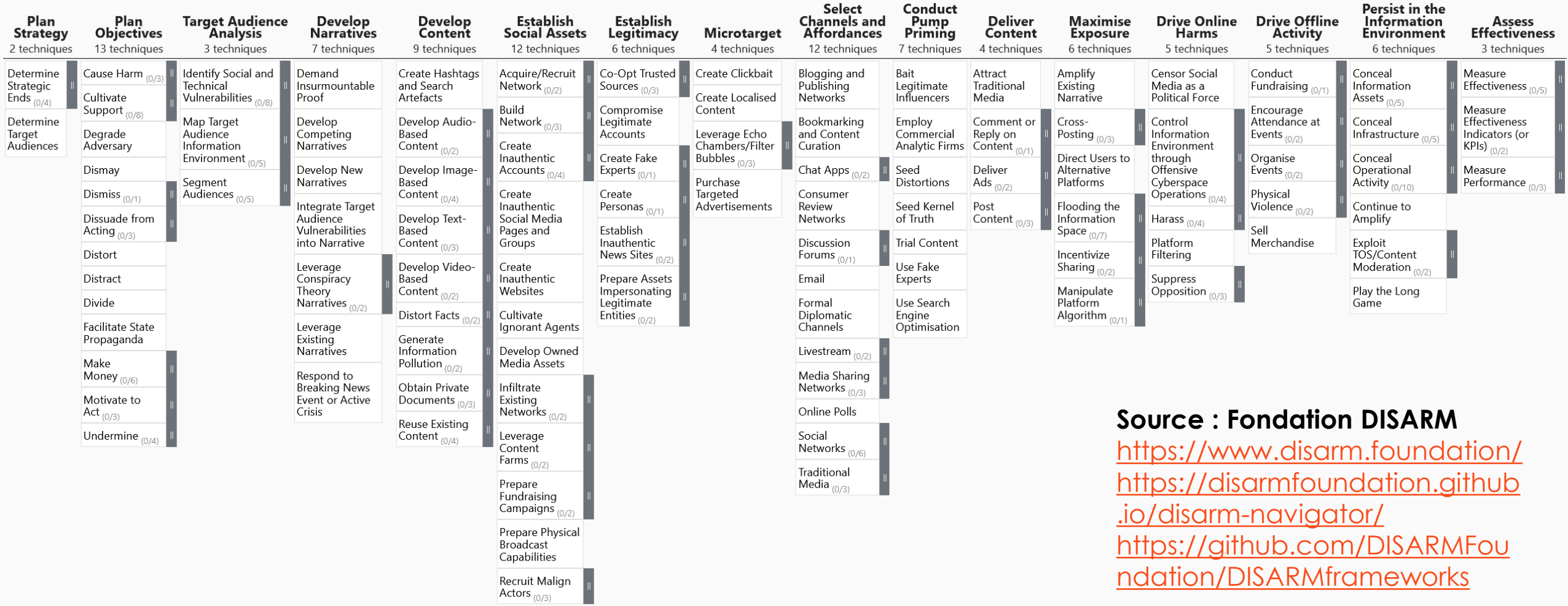


© Alejandro Hernández-Cano, et al., (2025), "Apertus: Democratizing Open and Compliant LLMs for Global Language Environments", Swiss National Supercomputing Centre (CSCS), https://www.swiss-ai.org/apertus, https://arxiv.org/abs/2509.14233

Partie 5 - Campagne de désinformation

Modélisation des campagnes de désinformation

Anticiper les prochaines étapes pour renforcer la désambiguïsation



Source : Fondation DISARM

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

Cadre DISARM

Modèle de référence pour la lutte contre les « campagnes » de désinformation

1. Phases : regroupements de tactiques de niveau supérieur, créés pour vérifier que nous n'avons rien manqué
2. Tactiques : étapes que la personne qui organise un incident de désinformation est susceptible d'utiliser
3. Techniques : activités qui peuvent être vues à chaque étape
4. Tâches : les actions à accomplir à chaque étape. En pablospeak, les tâches sont des actions que l'on fait, les techniques sont la manière de les réaliser.
5. Compteurs : contre-mesures pour DÉSARMER les tactiques, techniques et procédures (TTP).
6. Types d'acteurs : ressources nécessaires pour exécuter les contre-mesures
7. Types de réponses : les catégories de cours d'action que nous avons utilisées pour créer des compteurs
8. Métatechniques : un groupement de niveau supérieur pour les contre-mesures
9. Incidents : descriptions d'incidents utilisées pour créer les cadres DISARM

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

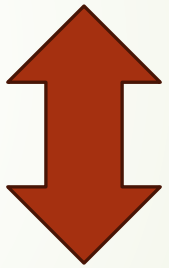
Cadre DISARM

Équipes rouges contre bleues

Tactique
TA01 Plan Stratégie
Objectifs du plan TA02
Microcible TA05
TA06 Développer du contenu
TA07 Sélectionner les canaux et les affordances
TA08 Effectuer l'amorçage de la pompe
TA09 Diffuser du contenu
Activité hors ligne TA10 Drive
TA11 Persister dans l'environnement informationnel
TA12 Évaluer l'efficacité
TA13 Analyse du public cible
TA14 Développer des récits
TA15 Établir des actifs sociaux
TA16 Établir la légitimité
TA17 Maximiser l'exposition
TA18 Conduire en ligne nuit

Équipe rouge
https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/disarm_red_framework.md

Tâches et techniques



Contre-mesures et manuel de jeu

Équipe bleue
https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/disarm_blue_framework.md

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

Cadre DISARM

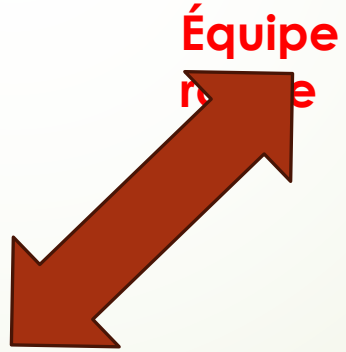
Techniques d'attaque vs contre-attaques : *exemple* - Tactique TA02 : Objectifs du plan

Résumé : Fixer des objectifs clairement définis, mesurables et atteignables. Dans certains cas, l'atteinte des objectifs lie l'exécution des tâches tactiques à l'atteinte de l'état final stratégique souhaité. Dans d'autres cas, en l'absence d'état final stratégique clairement défini, l'objectif tactique peut être autonome. L'énoncé de l'objectif ne doit pas

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

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



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

Partie 6 - Stratégie nationale

Stratégie nationale – « Chef.fe de la réalité » pour chaque institution

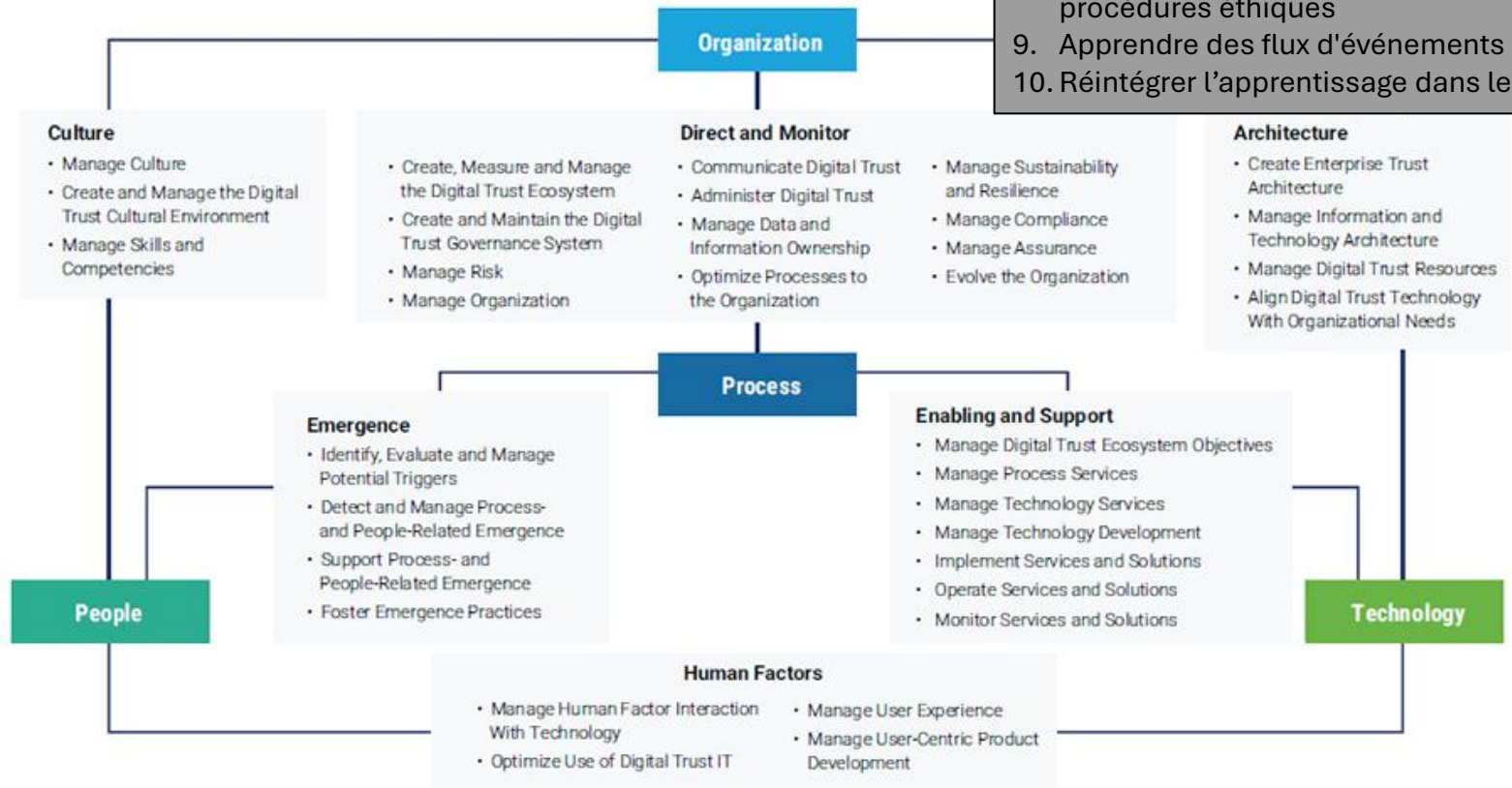
- La désinformation (fausses informations diffusées dans le but de nuire) prend de nombreuses formes, par exemple les fausses nouvelles, les usurpations d'identité audio-vidéo, la falsification de données, etc.
- Les élus sont souvent les principales cibles des campagnes de désinformation, notamment des accusations de fraude ou de corruption, qui peuvent perturber les projets du secteur public.
- Toutefois, lorsque les allégations de corruption contre des hommes politiques s'avèrent vraies, elles doivent être prises au sérieux et nécessitent une analyse approfondie immédiate.
- Nous soutenons qu'un modèle de « responsabilité parlementaire » constitue le meilleur dispositif institutionnel pour lutter contre la désinformation et la corruption dans le cadre d'une stratégie intégrée.
- Nous proposons que les parlements développent des pratiques innovantes consistant à créer un.e « Chef.fe de la réalité », ou « Chief Reality Officer » (CRO), doté d'une intelligence artificielle, en tant qu'agence indépendante qui leur rend directement compte.
- Le mandat d'un CRO serait de lutter contre la désinformation et la corruption en partenariat avec les entreprises et les gouvernements.
- Elle s'appuierait sur l'IA pour garantir le plus haut niveau de fiabilité et de transparence dans une vigilance constante, une diligence raisonnable et une collaboration en matière de renseignement open source (OSINT).
- En tant que cadre pour aider à guider ces nouvelles capacités CRO, nous proposons une extension du Digital Trust Ecosystem Framework (DTEF) de l'ISACA, ou le « Trust Framework ».
- Le cadre peut servir de « colle » pour lier tous les autres cadres couvrant l'écosystème interne, externe, ainsi que l'espace public des organisations.

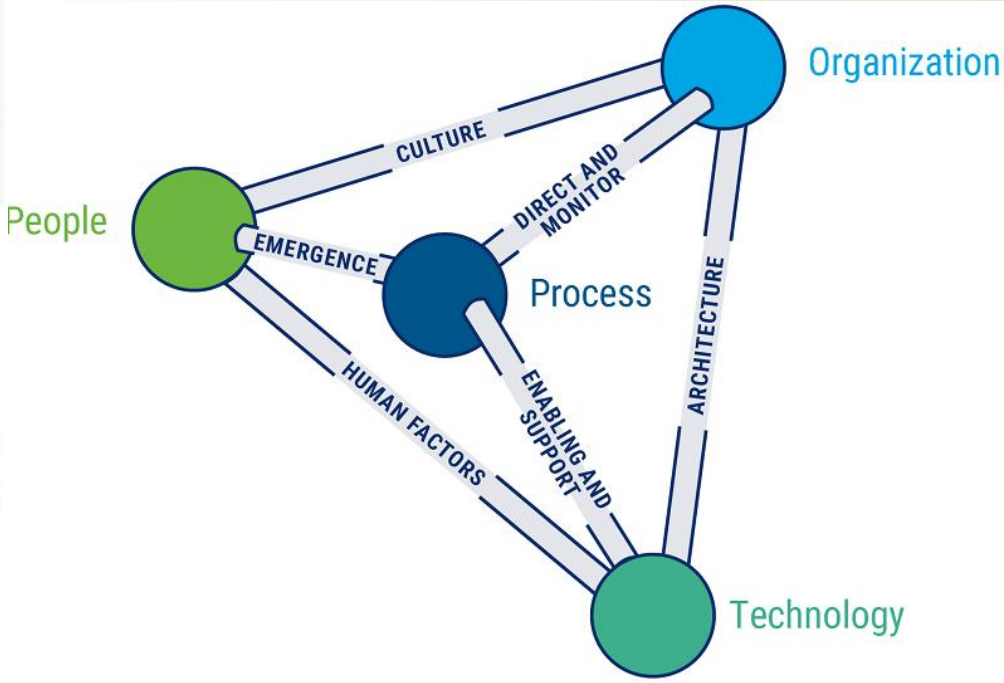
Composants de la confiance nume

Commun auté

Communiquer et contrôler

1. Définir les menaces, identifier leurs sources
2. Partager des renseignements open source au sein d'écosystèmes de confiance
3. Surveiller l'émergence et la diffusion des menaces
4. Corréler les événements et les messages entourant les allégations
5. Identifier les modèles de campagnes de désinformation
6. Utiliser l'IA pour distinguer les allégations fausses et vraies
7. Lier la désinformation aux contre-mesures
8. Relier les véritables allégations à la corruption du système judiciaire et aux procédures éthiques
9. Apprendre des flux d'événements et de modèles
10. Réintégrer l'apprentissage dans les capacités de surveillance systématique





ISACA. (2024). *Utiliser le cadre de l'écosystème de confiance numérique pour une IA digne de confiance* . https://www.isaca.org/resources/white-papers/2024/using-dtef-to-achieve-trustworthy-ai

Source : ISACA, (2022), *Cadre de l'écosystème de confiance numérique* https://www.isaca.org/digital-trust

Merci!

Stéphane Gagnon, Ph.D.

Professeur agrégé

Université du Québec en Outaouais (UQO)

https://gagnontech.org/bio/

admin@gagnontech.org

https://disinform.app