CENTER FOR aUTonomy

The University of Texas at Austin

Modeling and Countering Misinformation in Adversarial Information Ecosystems

Arash Amini, Yigit Ege Bayiz, and Ufuk Topcu

AFOSR Center of Excellence in Assured Autonomy in Contested Environments

Information Enviroment

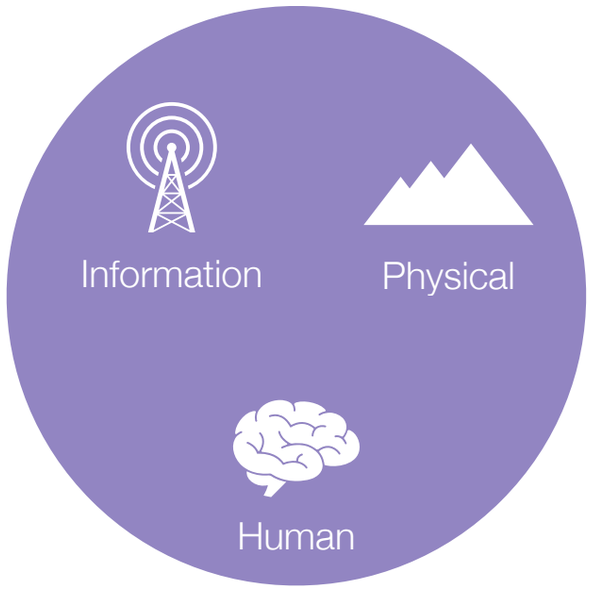
Defense for Special Operations and Low-Intensity Conflict, stated that adversarial use of *disinformation*, *misinformation*, and *propaganda* poses one of today's *greatest challenges* to the United States.

[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders.* Report to Congressional Committees.

Information Enviroment

CENTER FOR aUTonomy

Defense for Special Operations and Low-Intensity Conflict, stated that adversarial use of *disinformation*, *misinformation*, and *propaganda* poses one of today's *greatest challenges* to the United States.

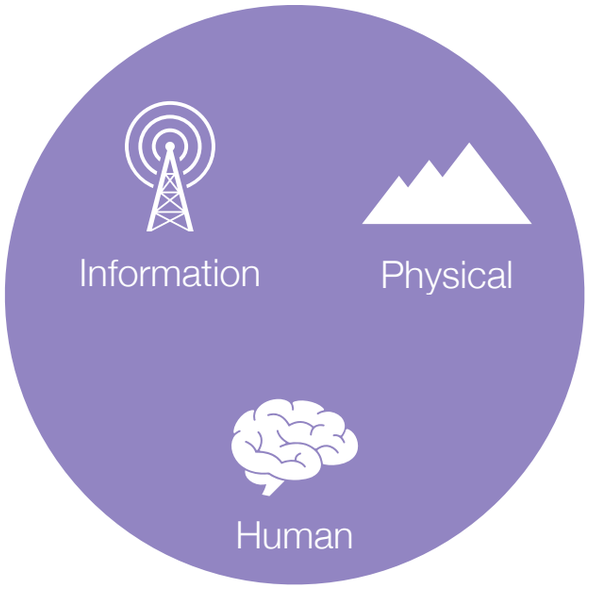


[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders.* Report to Congressional Committees.

Information Enviroment

CENTER FOR aUTonomy

Defense for Special Operations and Low-Intensity Conflict, stated that adversarial use of *disinformation*, *misinformation*, and *propaganda* poses one of today's *greatest challenges* to the United States.

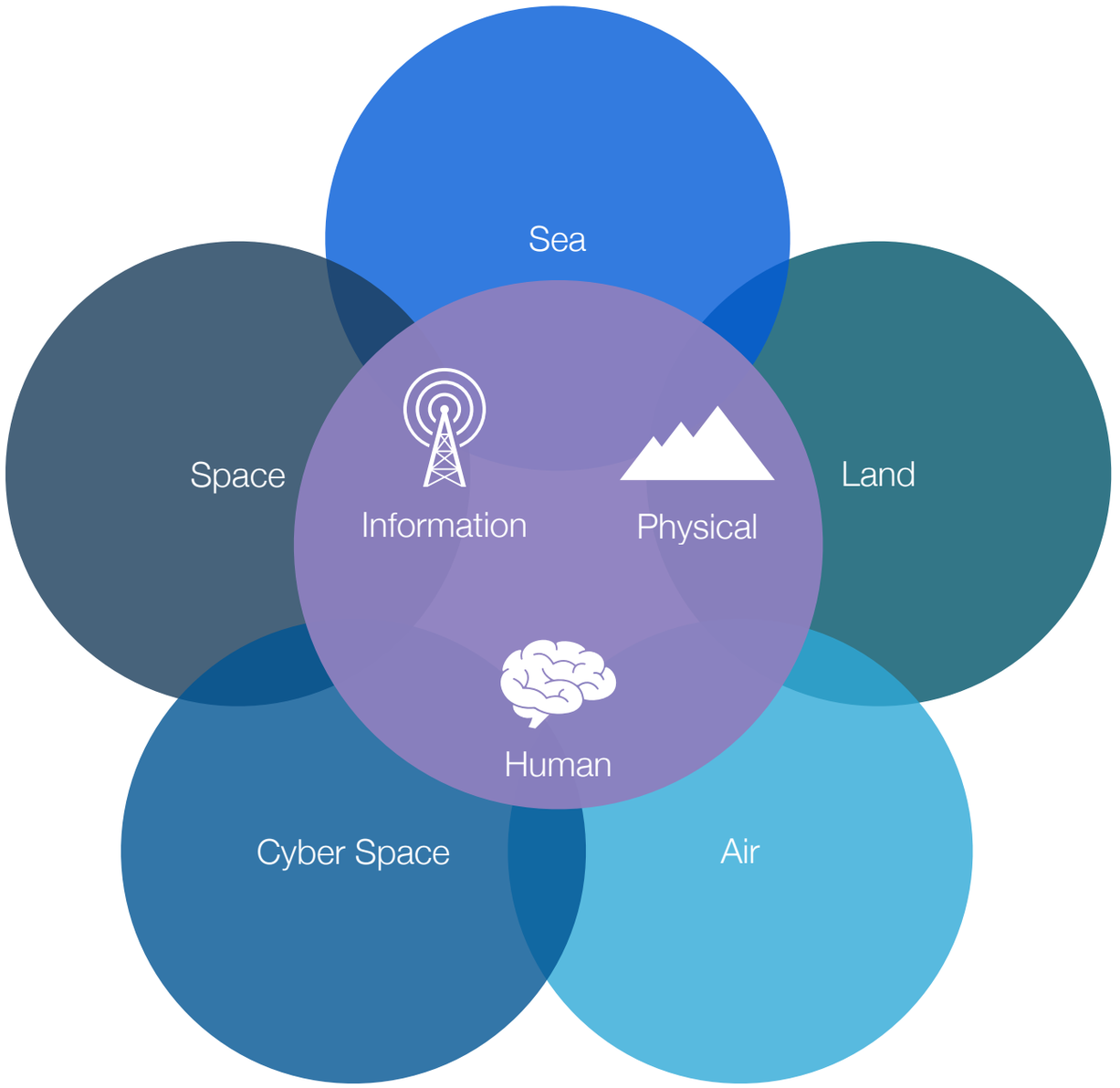


[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders.* Report to Congressional Committees.

Information Enviroment

CENTER FOR aUTonomy

Defense for Special Operations and Low-Intensity Conflict, stated that adversarial use of *disinformation*, *misinformation*, and *propaganda* poses one of today's *greatest challenges* to the United States.



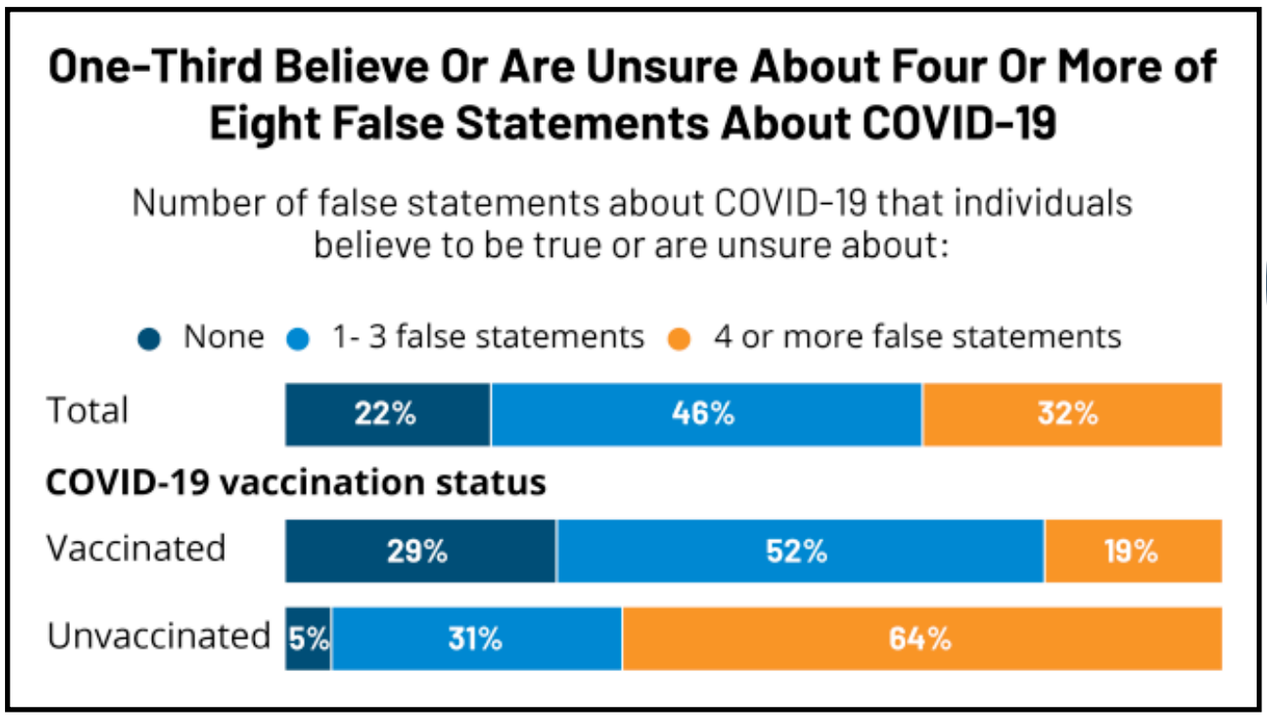
[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders.* Report to Congressional Committees.

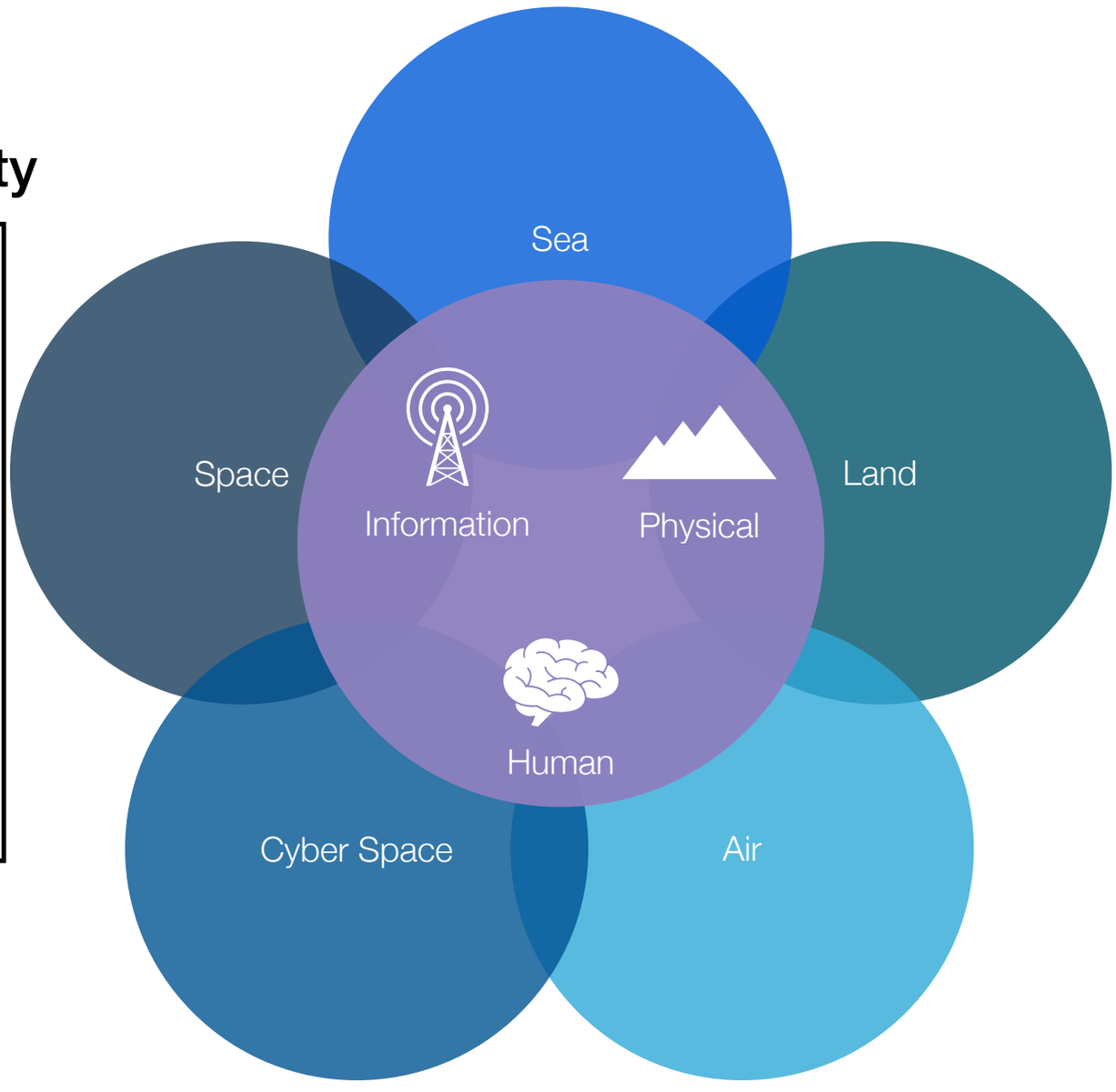
Information Enviroment

CENTER FOR aUTonomy

Defense for Special Operations and Low-Intensity Conflict, stated that adversarial use of *disinformation*, *misinformation*, and *propaganda* poses one of today's *greatest challenges* to the United States.

A real threat that goes beyond national security

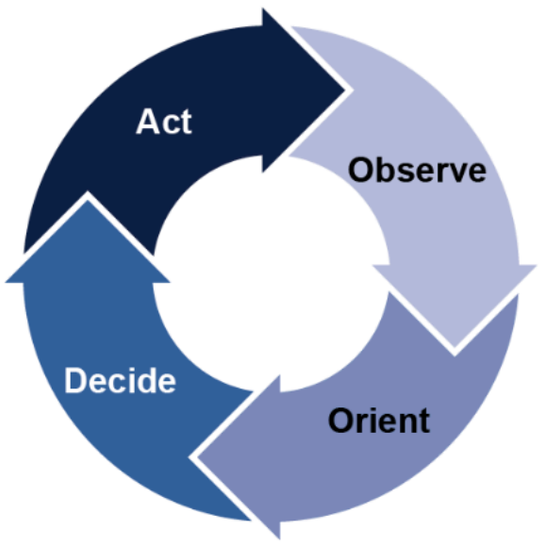




[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders.* Report to Congressional Committees.

How Disinformation Works

CENTER FOR aUTonomy

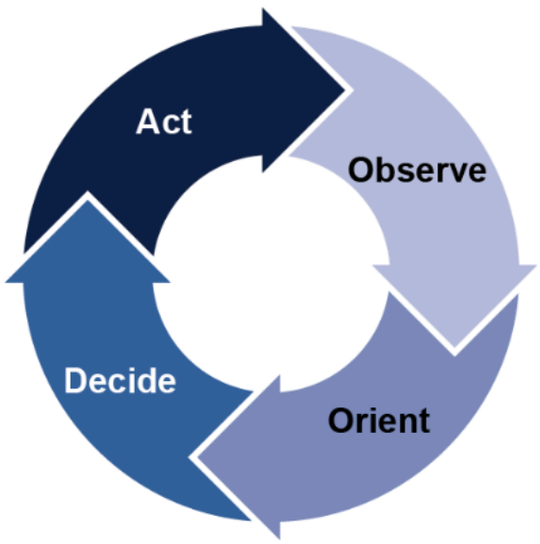


The OODA loop in theory appears as a consistent cycle of observe, orient, decide, and act. However, operational and adversarial inputs can distort this process in a number of ways.

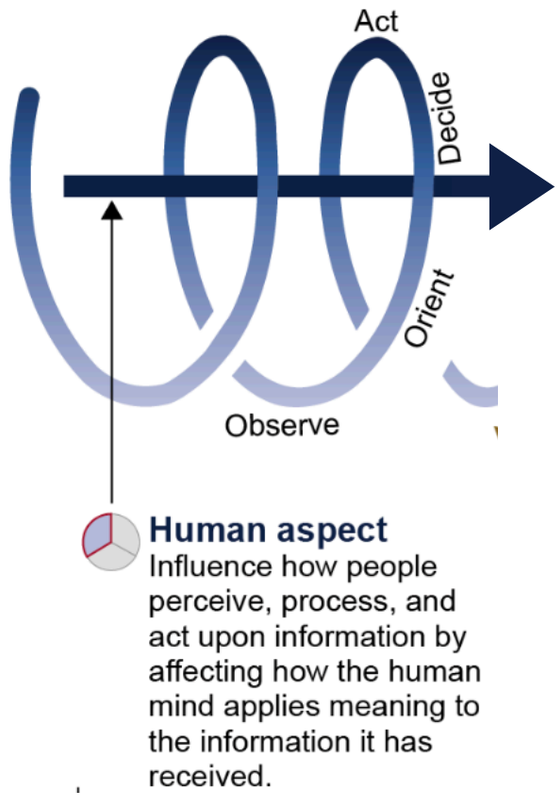
[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders*. Report to Congressional Committees.

How Disinformation Works

CENTER FOR aUTonomy



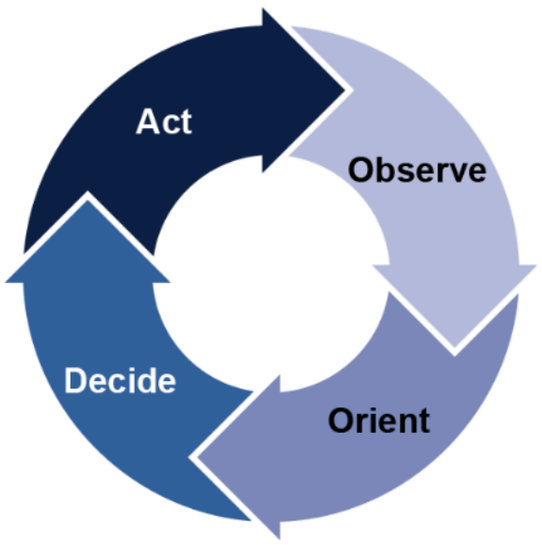
The OODA loop in theory appears as a consistent cycle of observe, orient, decide, and act. However, operational and adversarial inputs can distort this process in a number of ways.



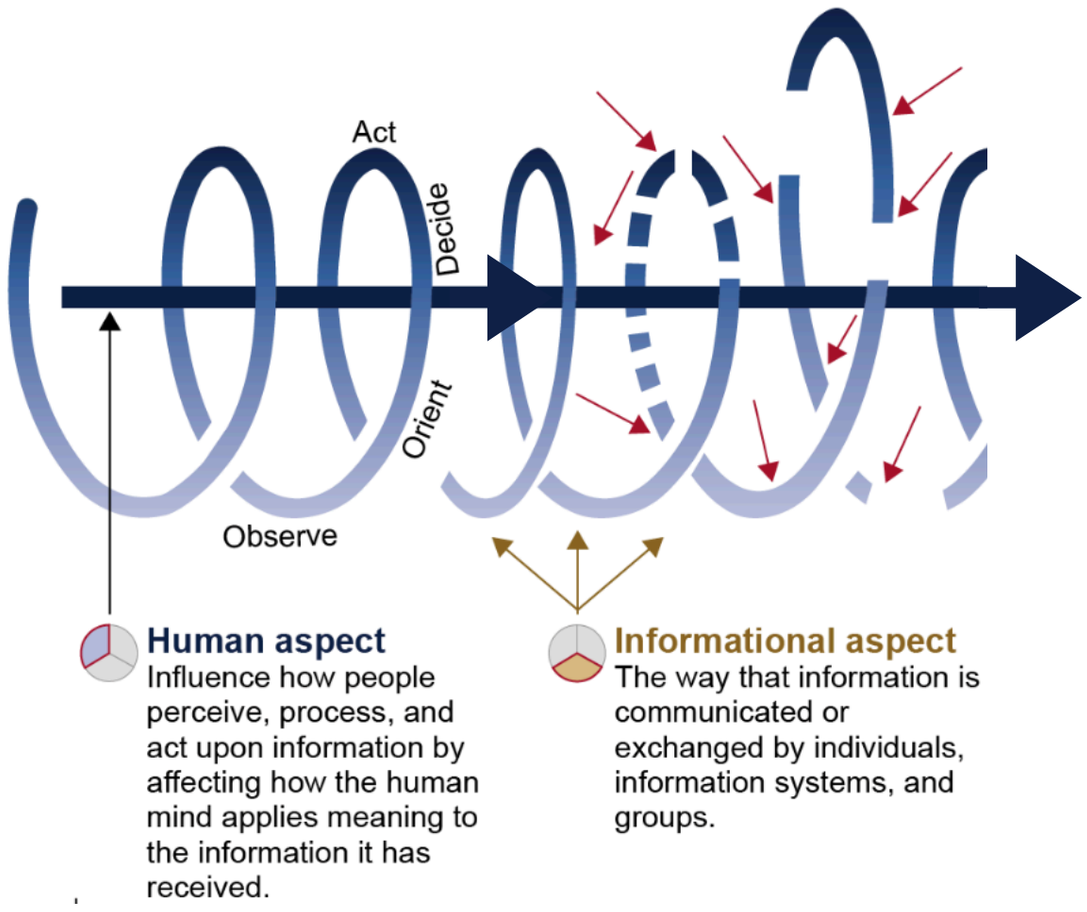
[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders*. Report to Congressional Committees.

How Disinformation Works

CENTER FOR aUTonomy



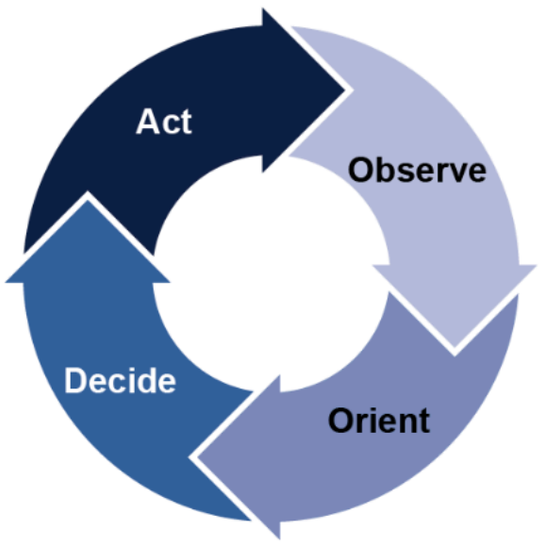
The OODA loop in theory appears as a consistent cycle of observe, orient, decide, and act. However, operational and adversarial inputs can distort this process in a number of ways.



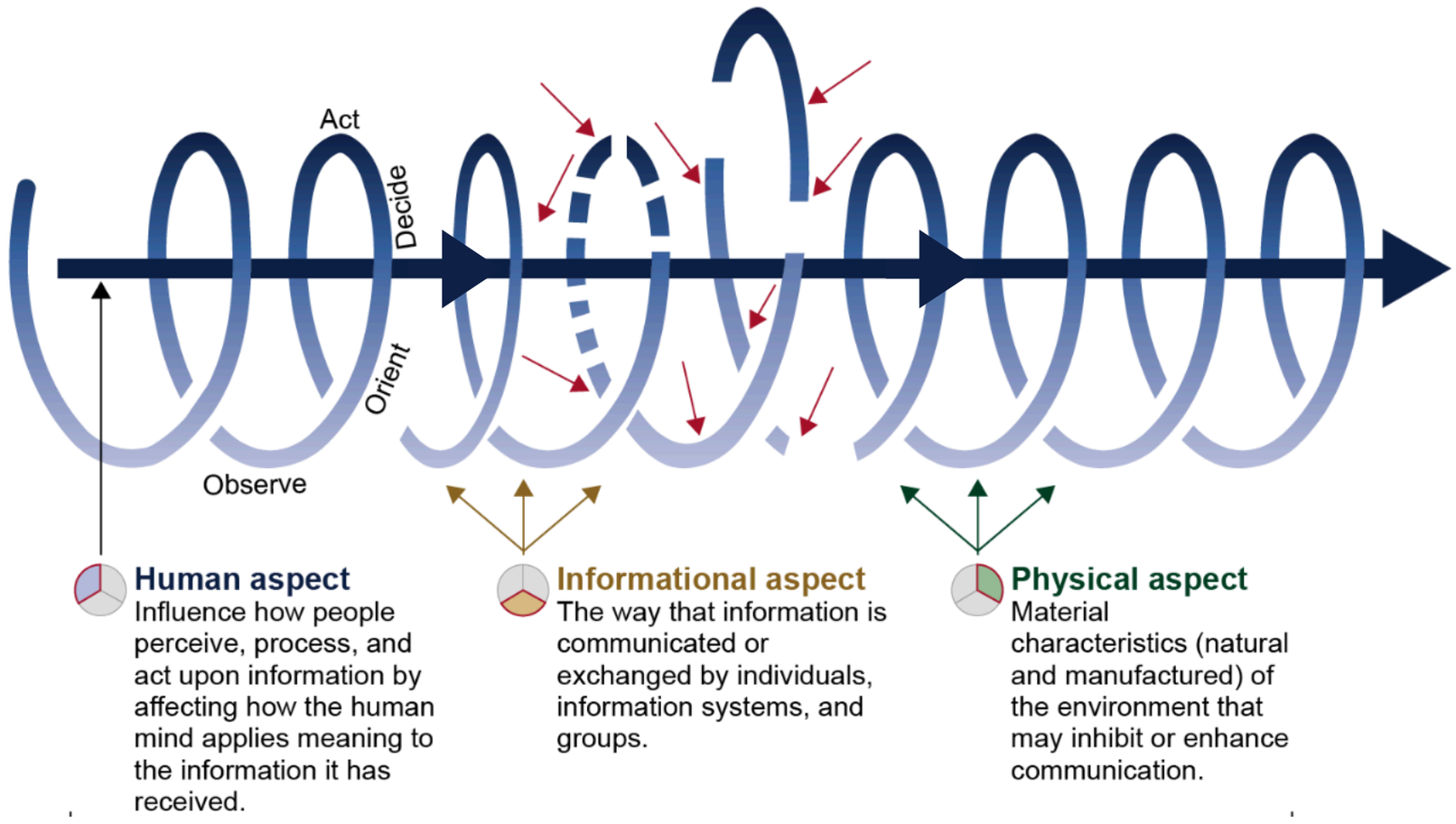
[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders*. Report to Congressional Committees.

How Disinformation Works

CENTER FOR aUTonomy



The OODA loop in theory appears as a consistent cycle of observe, orient, decide, and act. However, operational and adversarial inputs can distort this process in a number of ways.



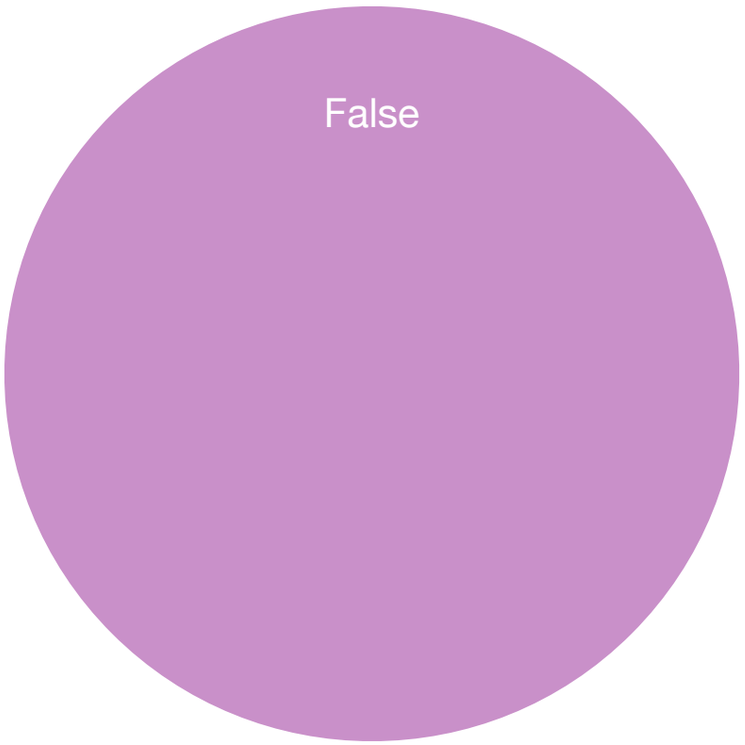
[1] Kirschbaum, Joseph W. *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders*. Report to Congressional Committees.

Definition of Disinformation

CENTER FOR aUTonomy

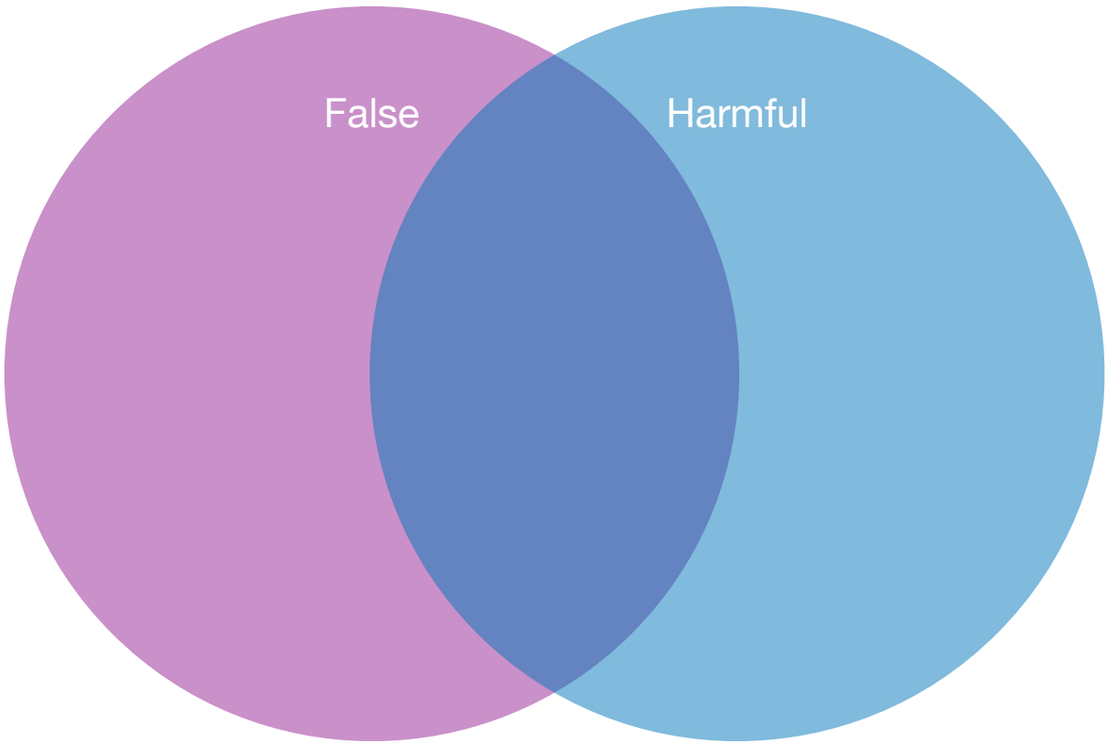
Definition of Disinformation

CENTER FOR aUTonomy



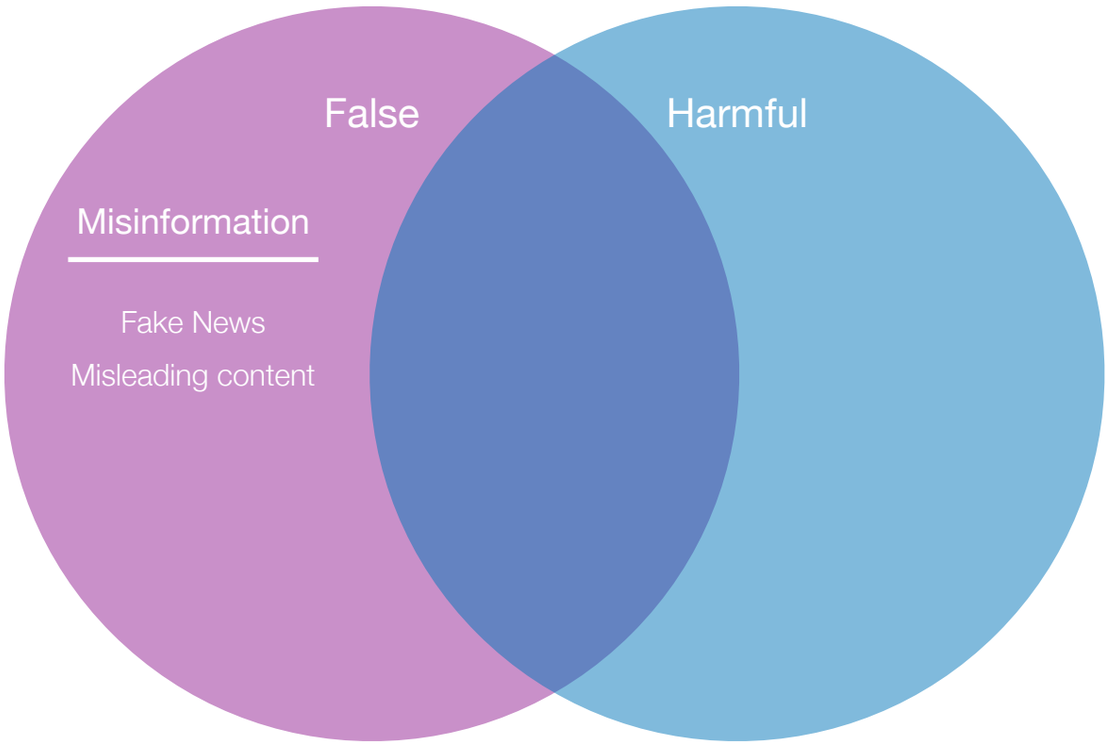
Definition of Disinformation

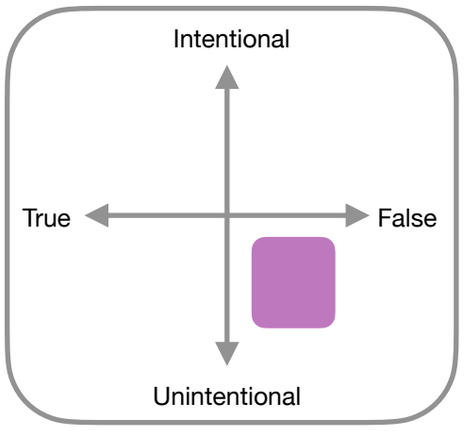
CENTER FOR aUTonomy



Definition of Disinformation

CENTER FOR aUTonomy

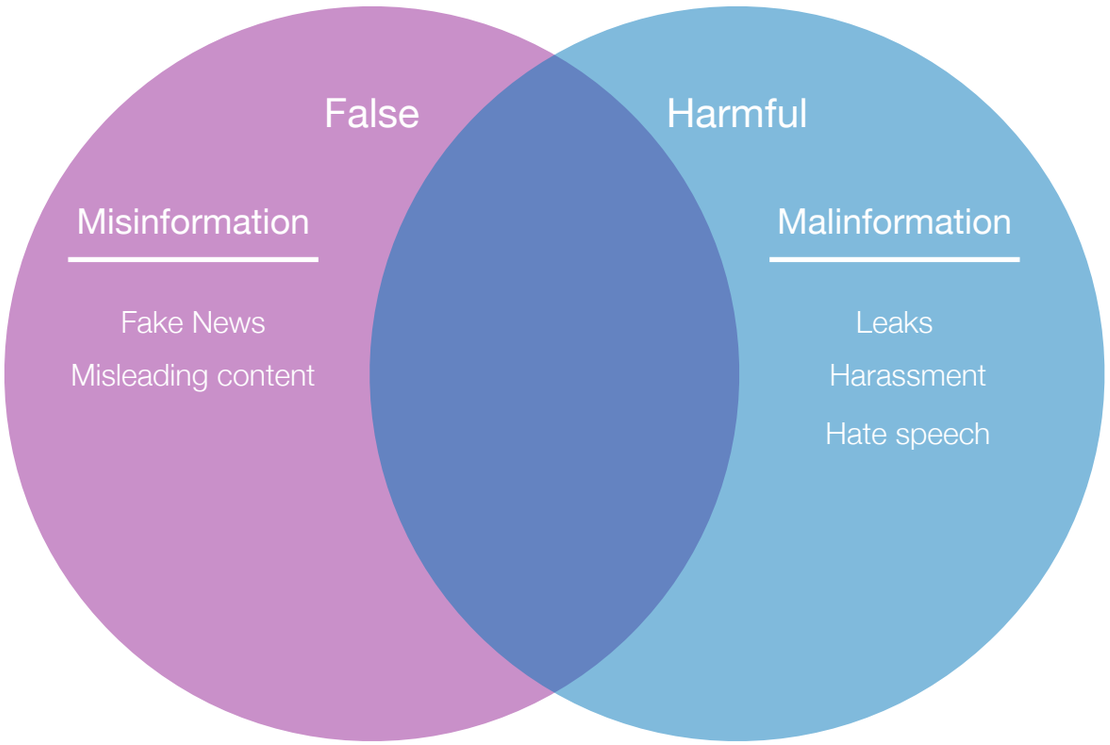


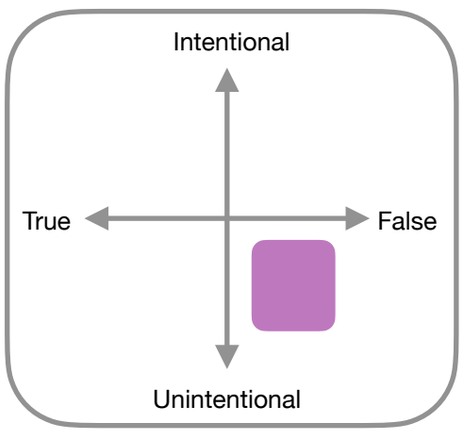


Misinformation

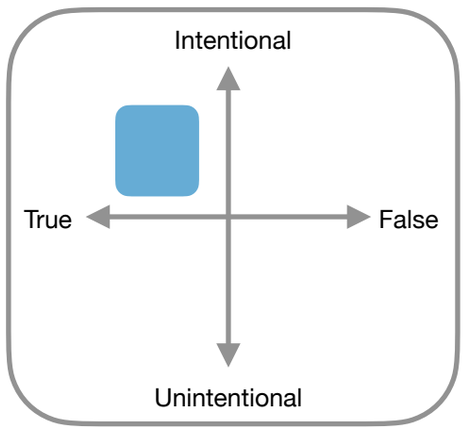
Definition of Disinformation

CENTER FOR aUTonomy





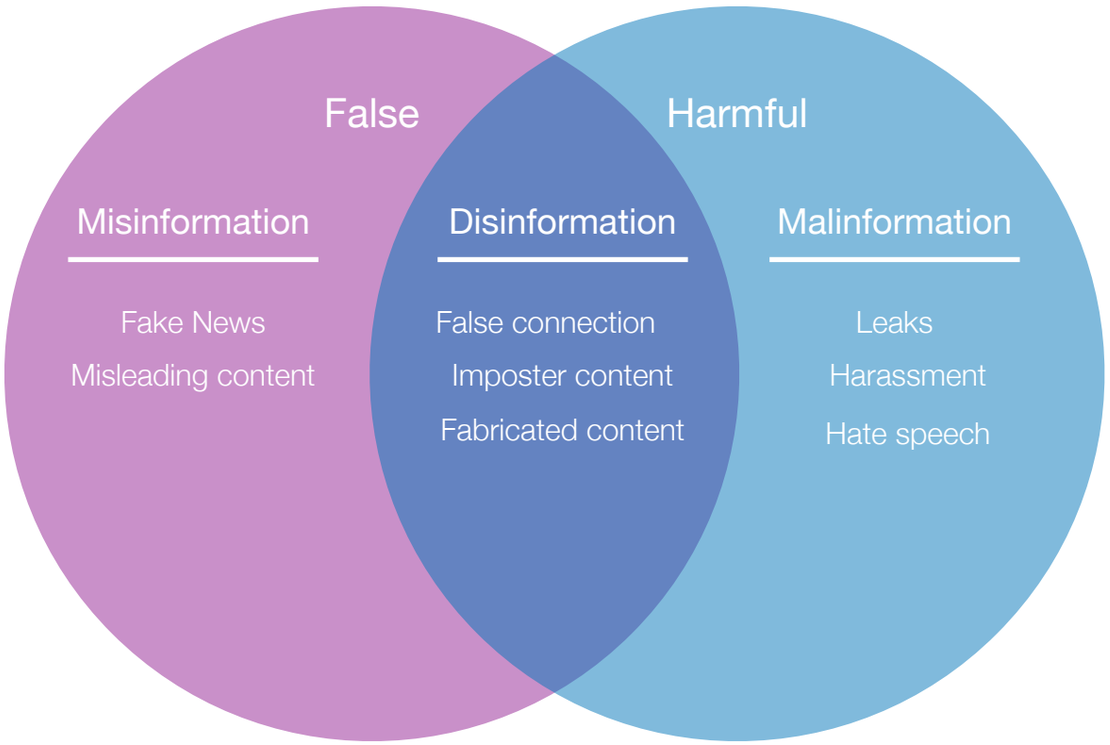
Misinformation

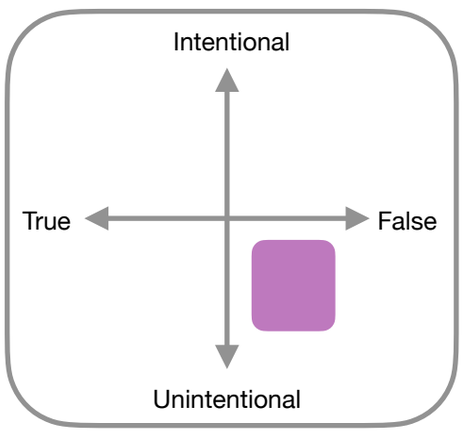


Malinformation

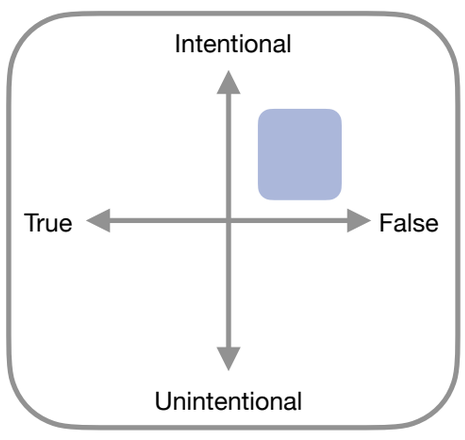
Definition of Disinformation

CENTER FOR aUTonomy

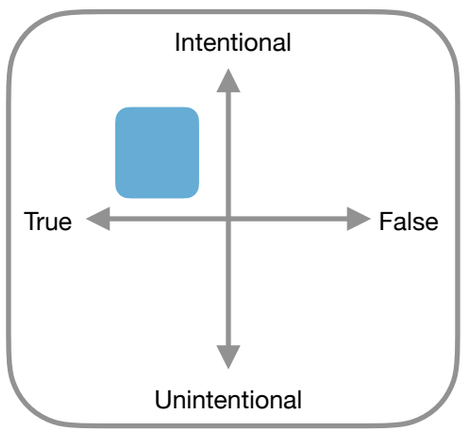




Misinformation



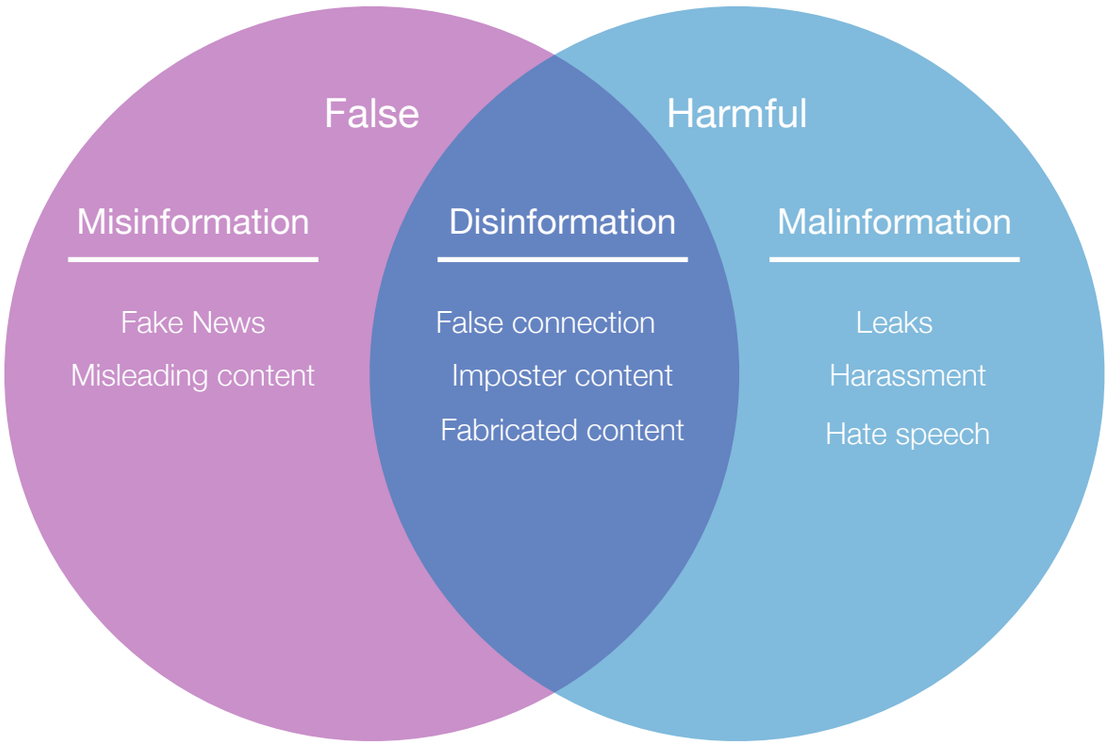
Disinformation

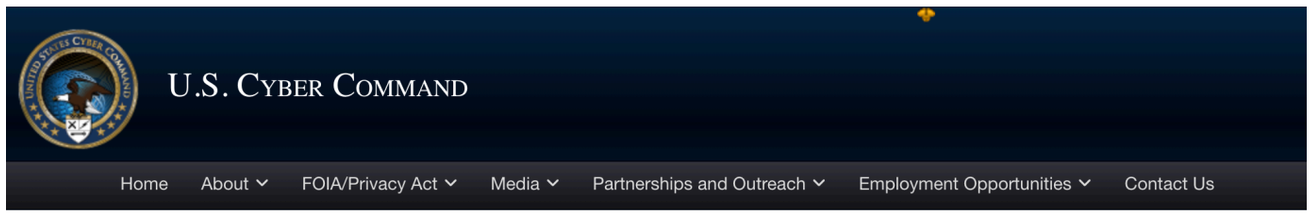


Malinformation

Definition of Disinformation

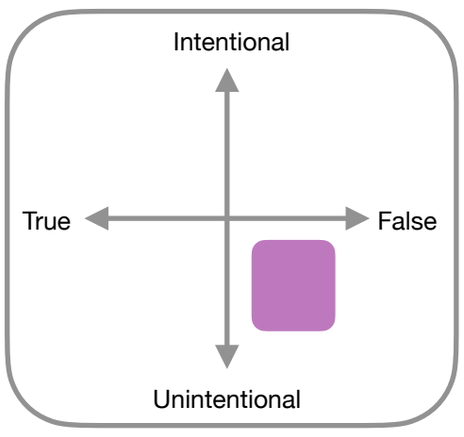
CENTER FOR aUTonomy



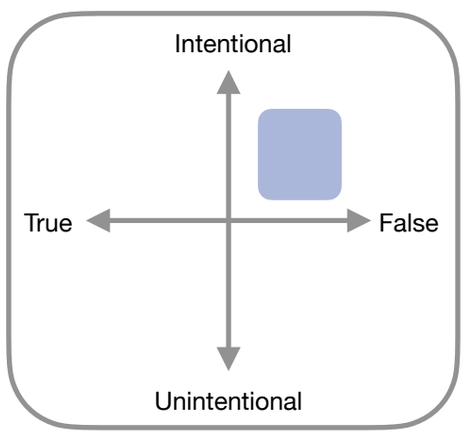


NEWS | Sept. 3, 2024

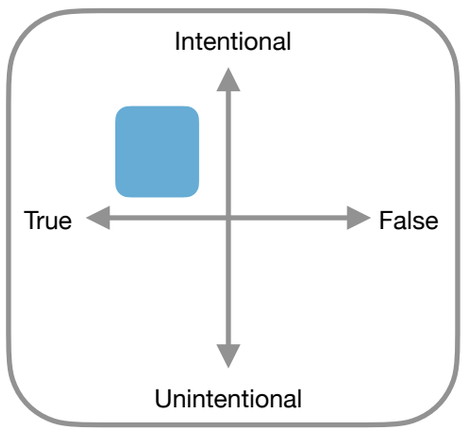
Russian Disinformation Campaign "DoppelGänger" Unmasked: A Web of Deception



Misinformation



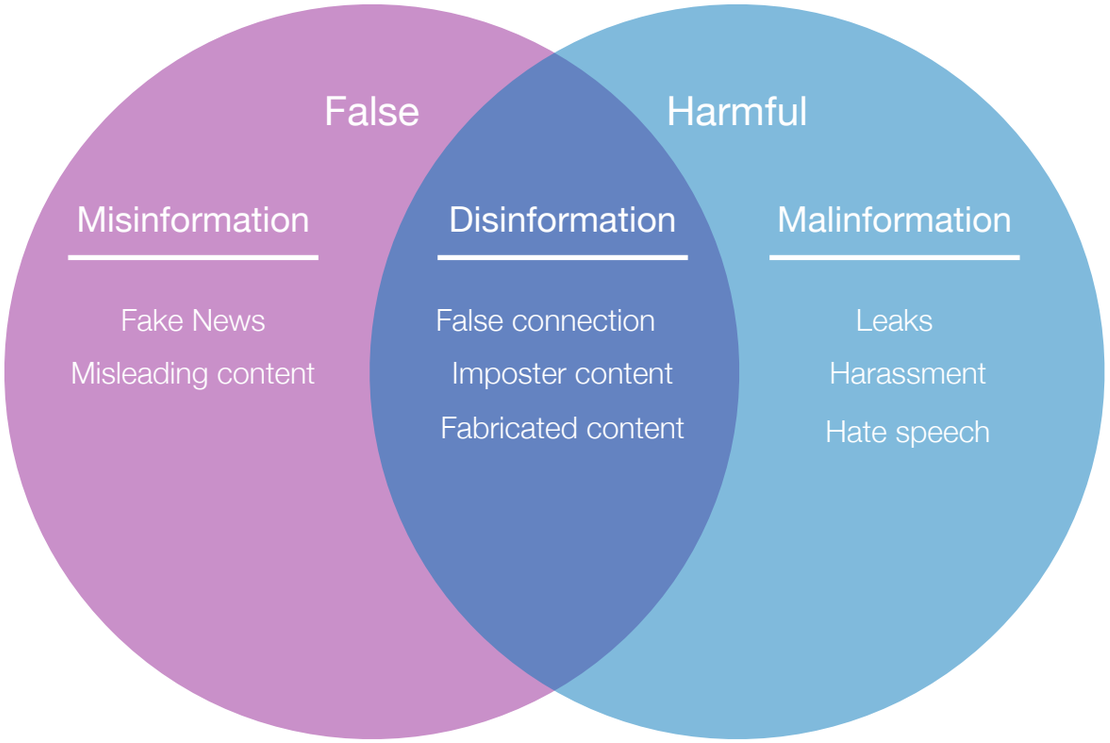
Disinformation

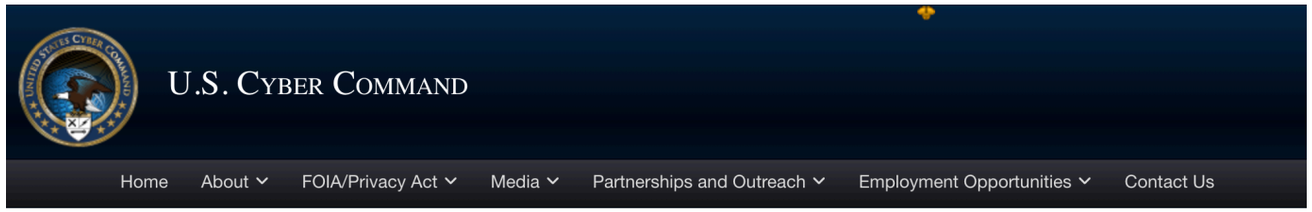


Malinformation

Definition of Disinformation

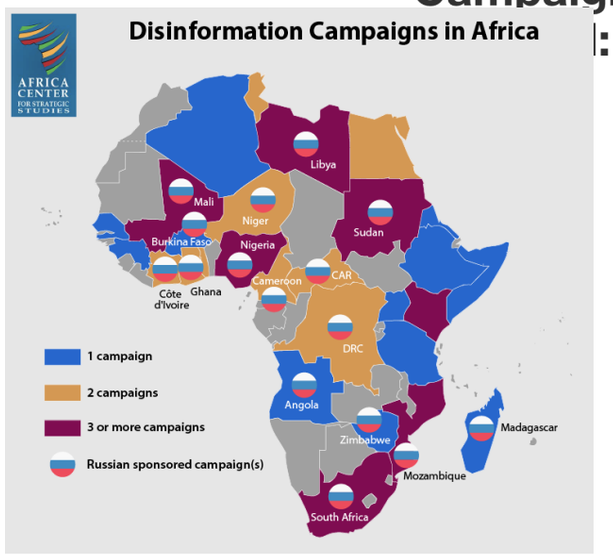
CENTER FOR aUTonomy

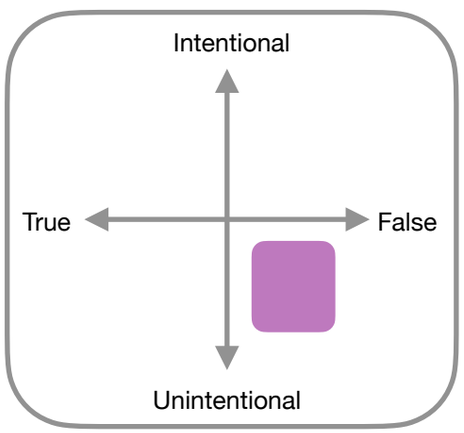




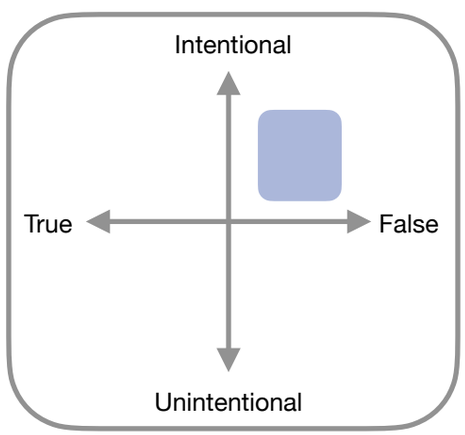
NEWS | Sept. 3, 2024

Russian Disinformation Campaign "DoppelGänger" : A Web of Deception

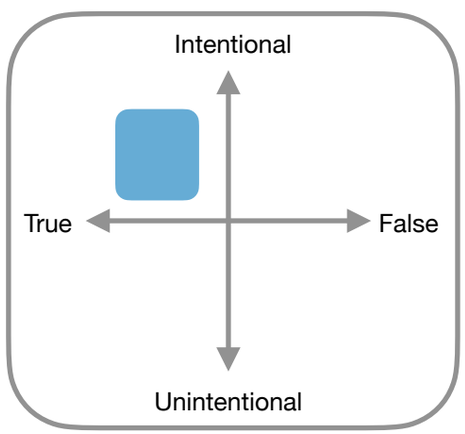




Misinformation



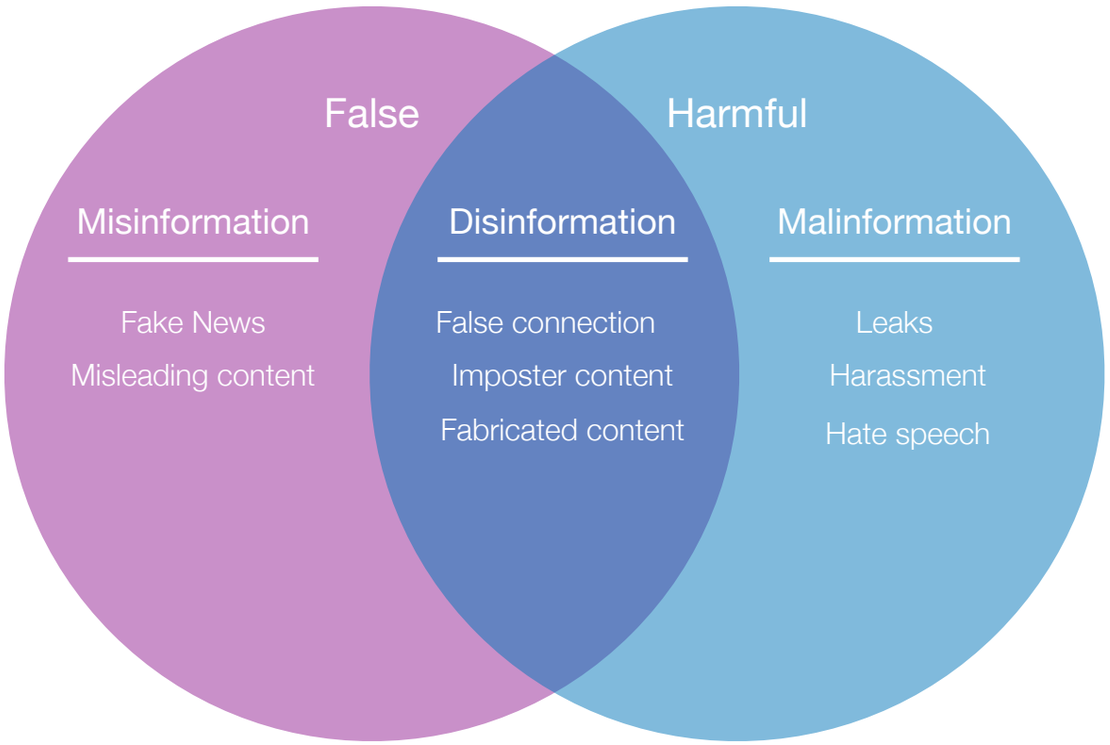
Disinformation

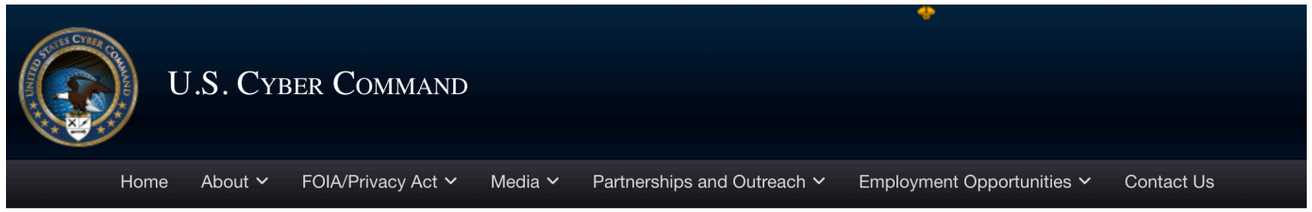


Malinformation

Definition of Disinformation

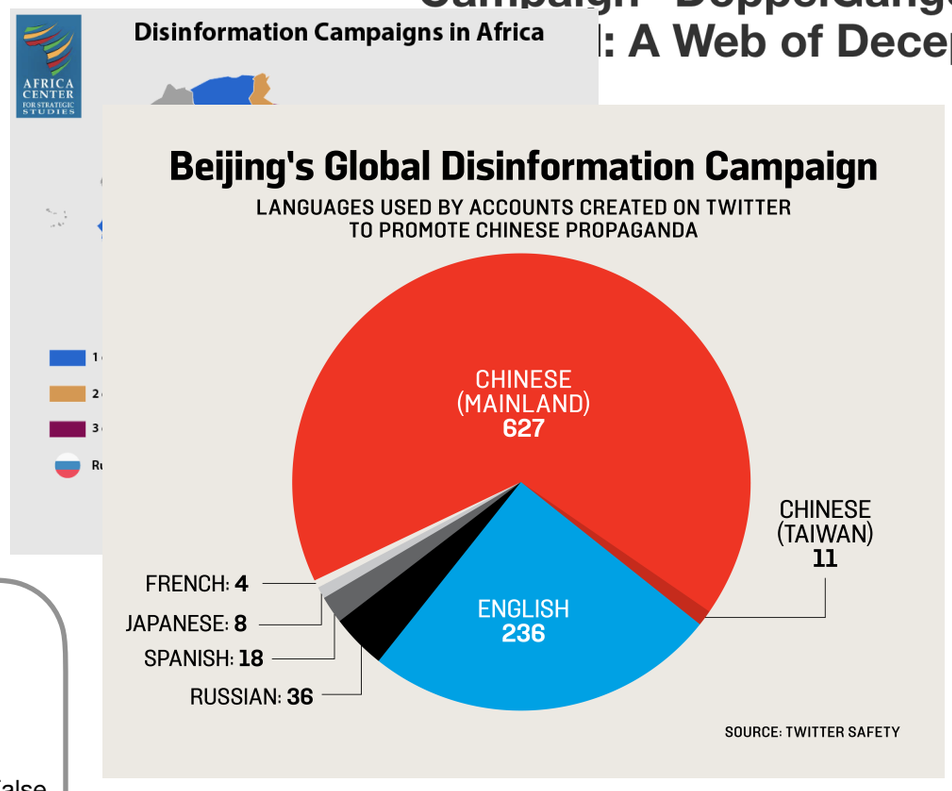
CENTER FOR aUTonomy

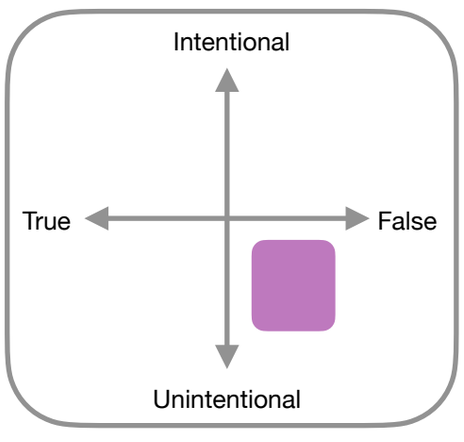




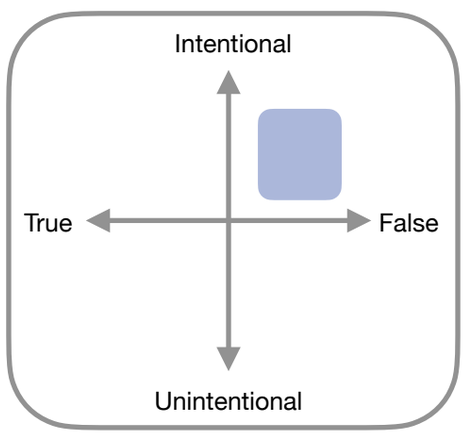
NEWS | Sept. 3, 2024

Russian Disinformation Campaign "DoppelGänger" : A Web of Deception

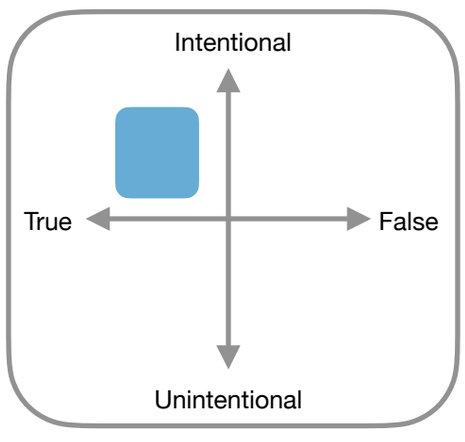




Misinformation



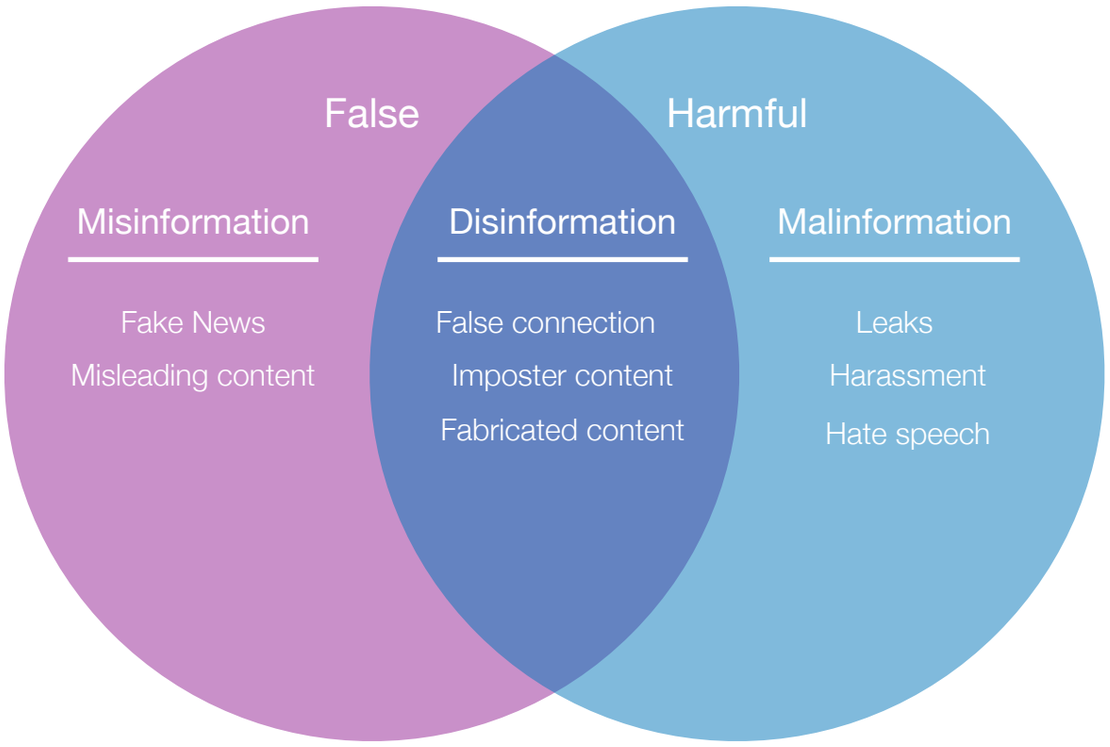
Disinformation

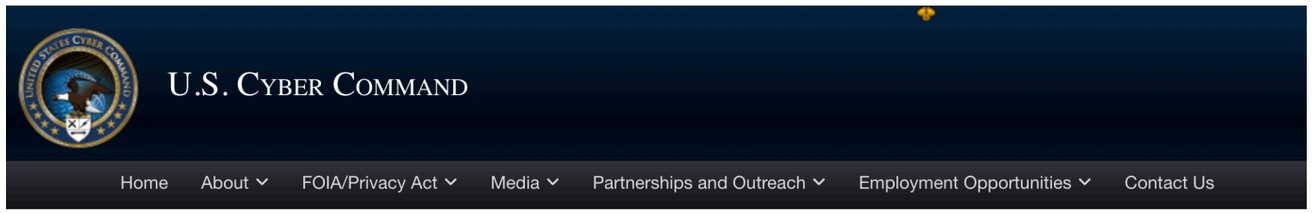


Malinformation

Definition of Disinformation

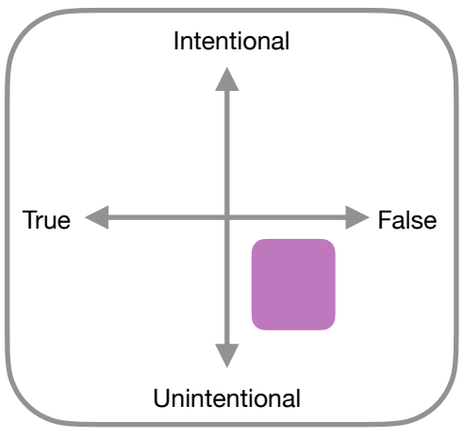
CENTER FOR aUTonomy



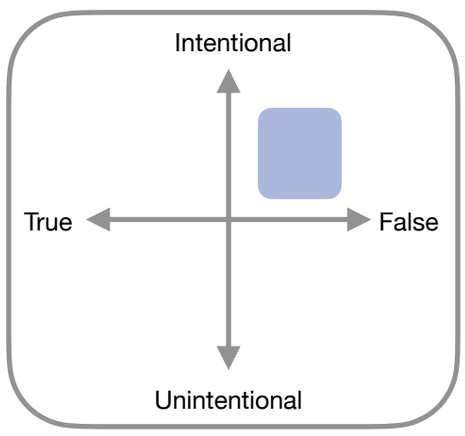


NEWS | Sept. 3, 2024

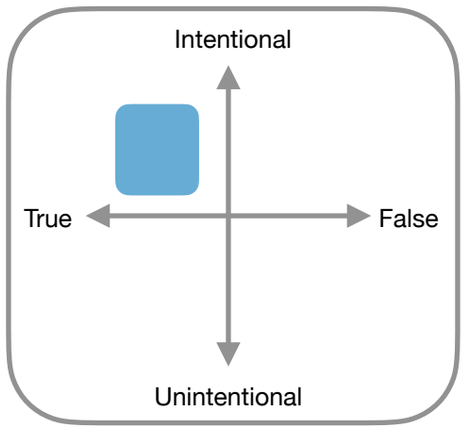
Russian Disinformation Campaign "DoppelGänger" : A Web of Deception



Misinformation



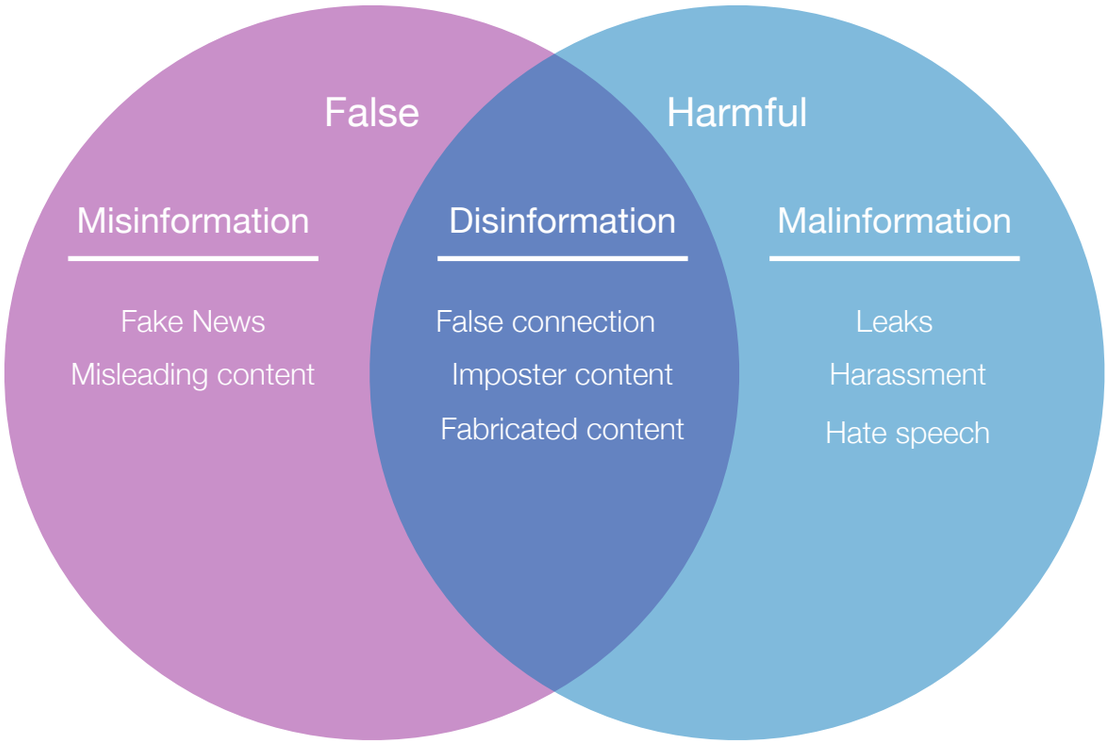
Disinformation

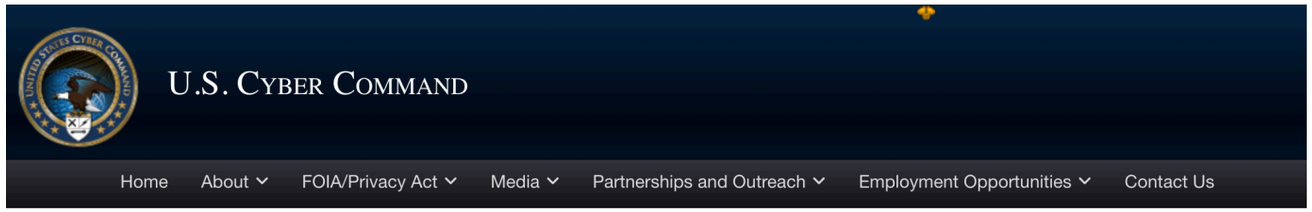


Malinformation

Definition of Disinformation

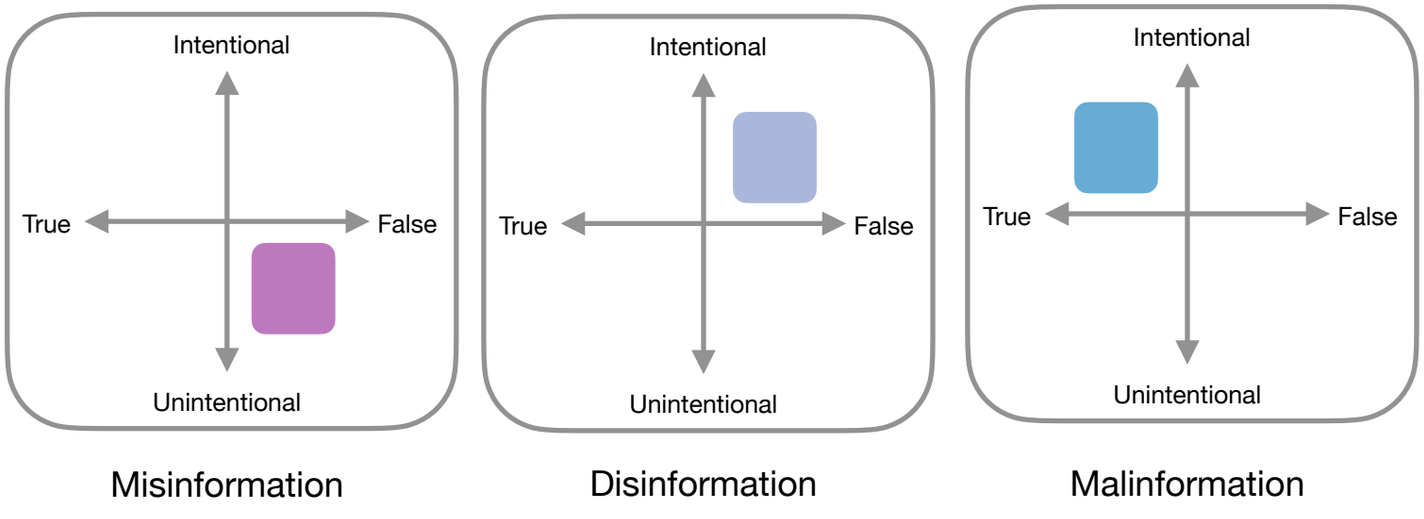
CENTER FOR aUTonomy





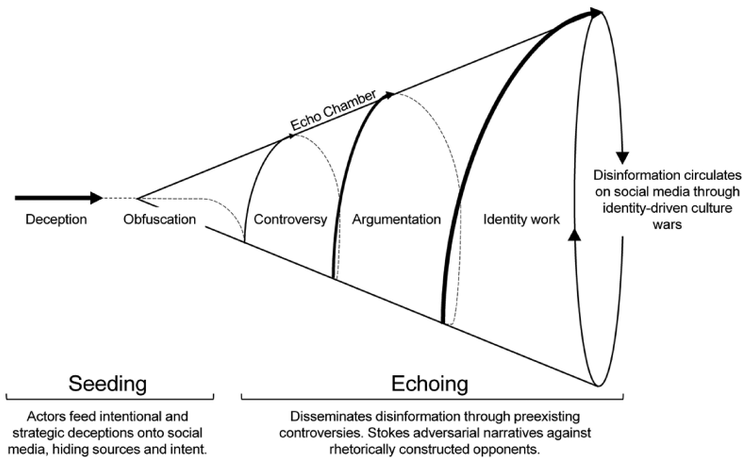
NEWS | Sept. 3, 2024

Russian Disinformation Campaign "DoppelGänger" : A Web of Deception



Disinformation Evolution

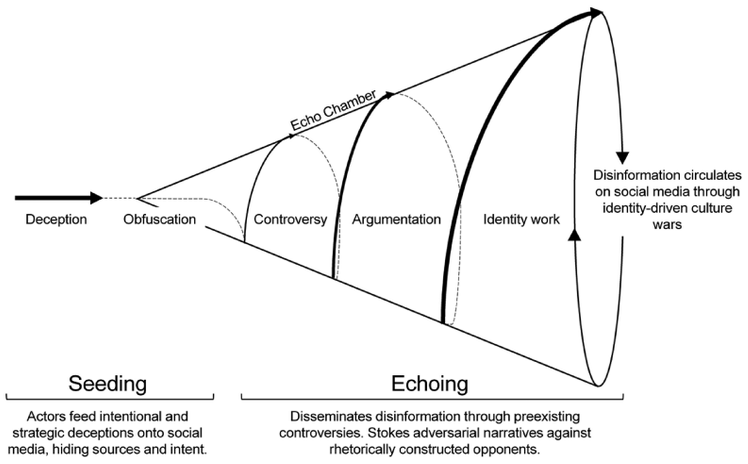
CENTER FOR aUTonomy

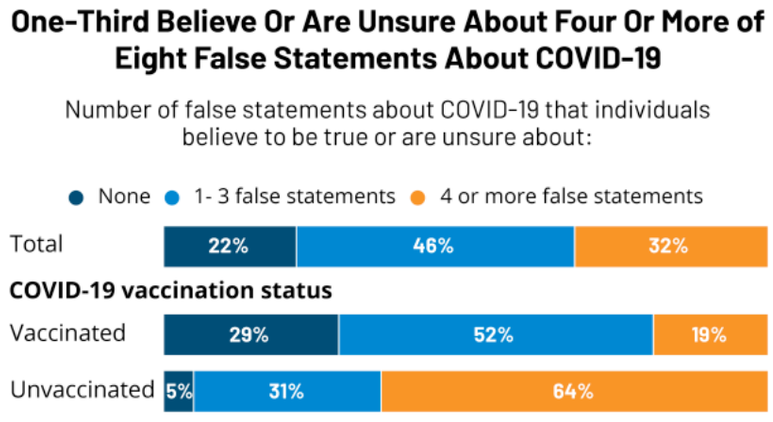


Initiation

Disinformation Evolution

CENTER FOR aUTonomy

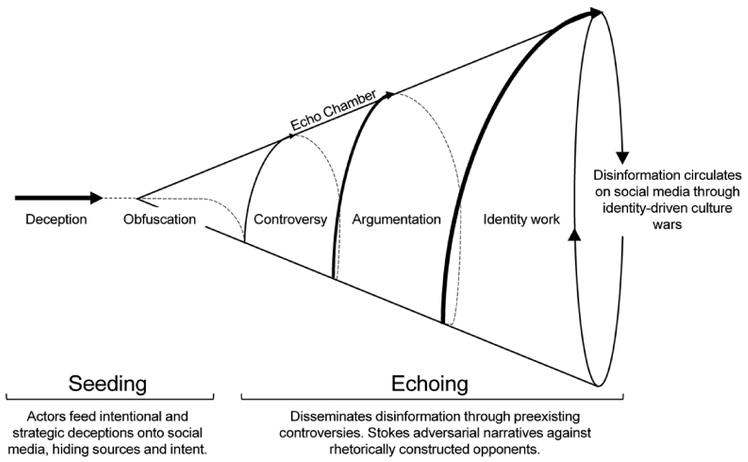


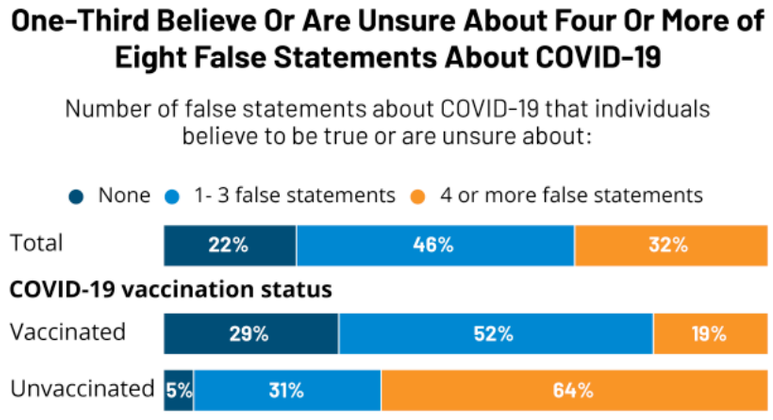


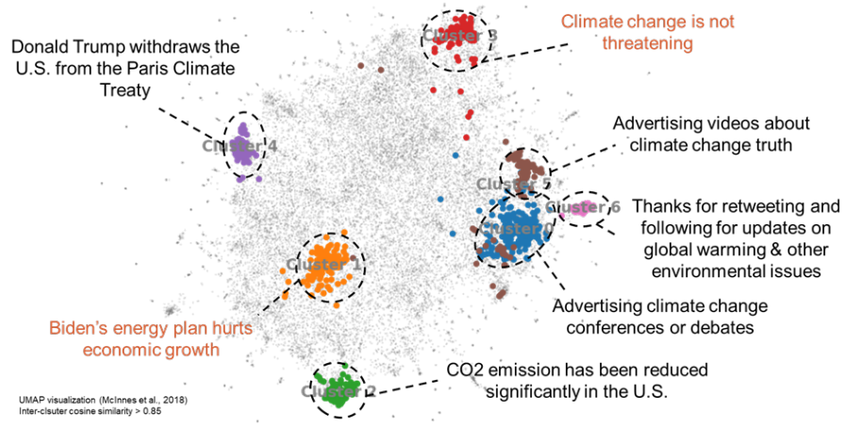
Initiation **Internalization**

Disinformation Evolution

CENTER FOR aUTonomy

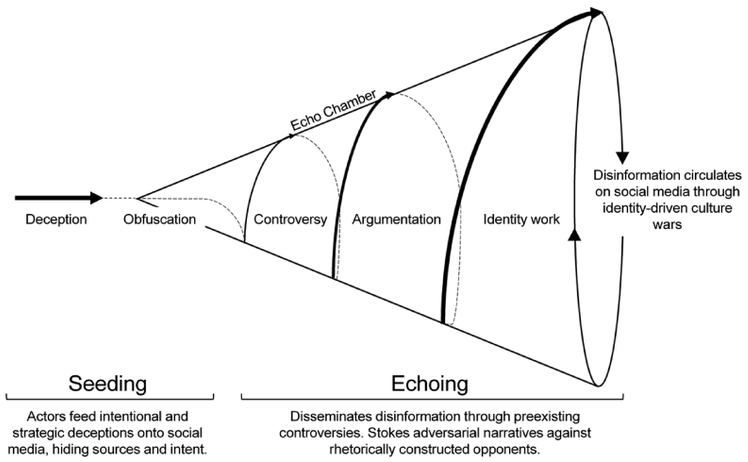


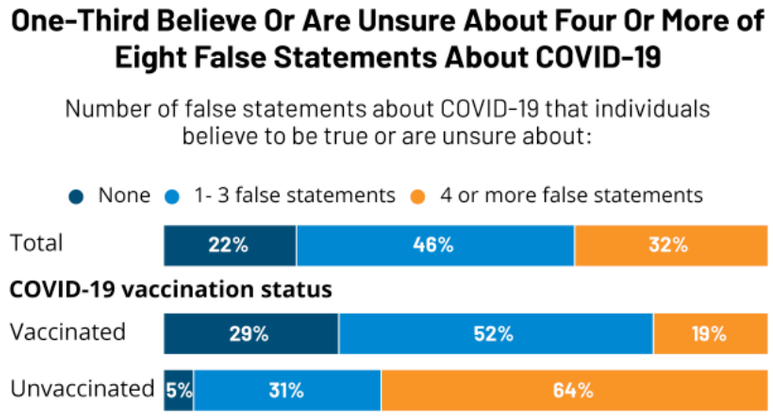


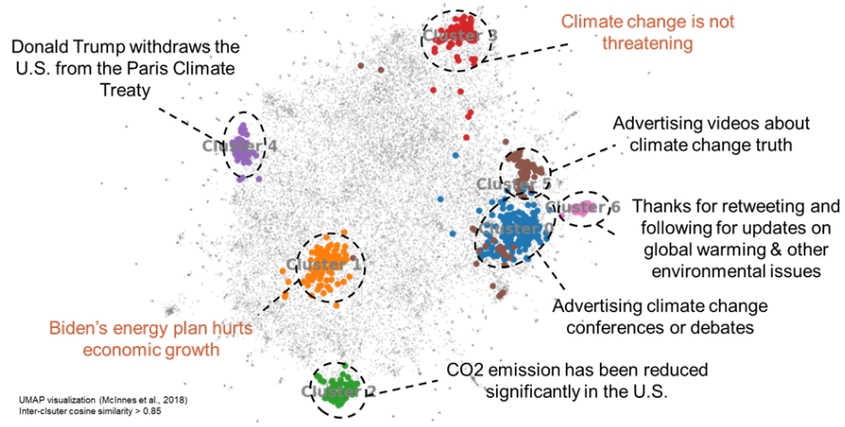


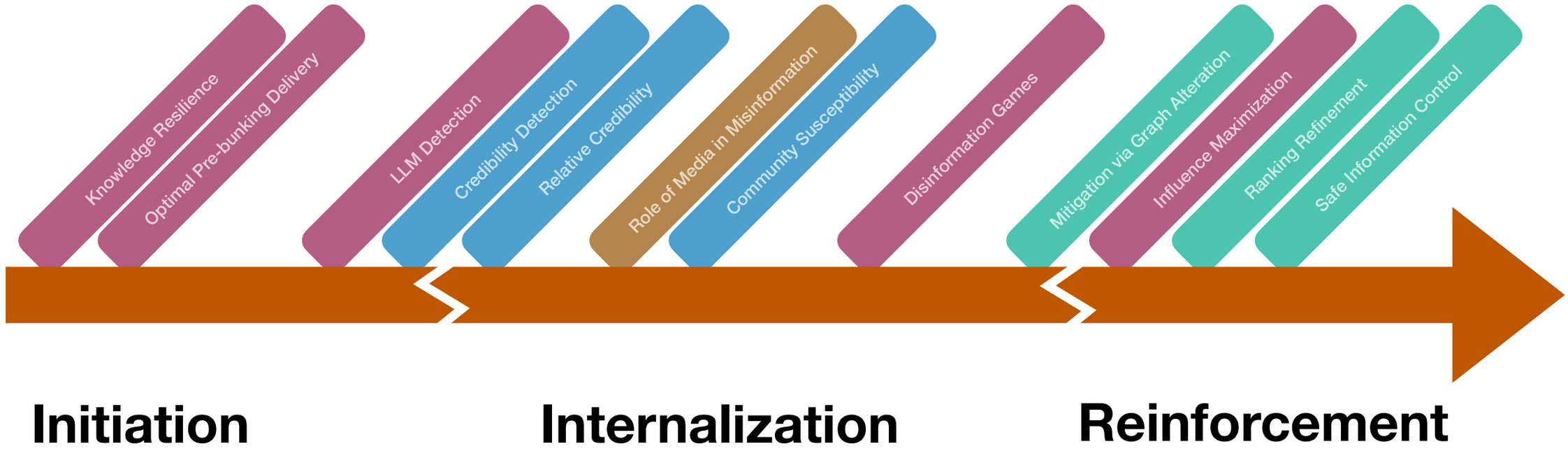
Disinformation Evolution

CENTER FOR aUTonomy



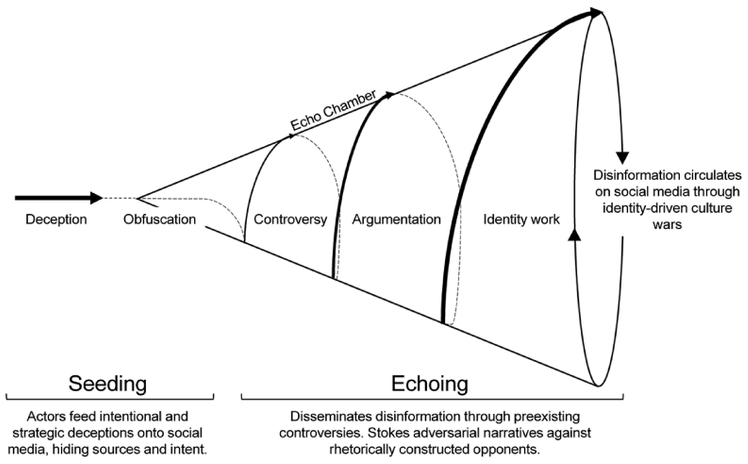


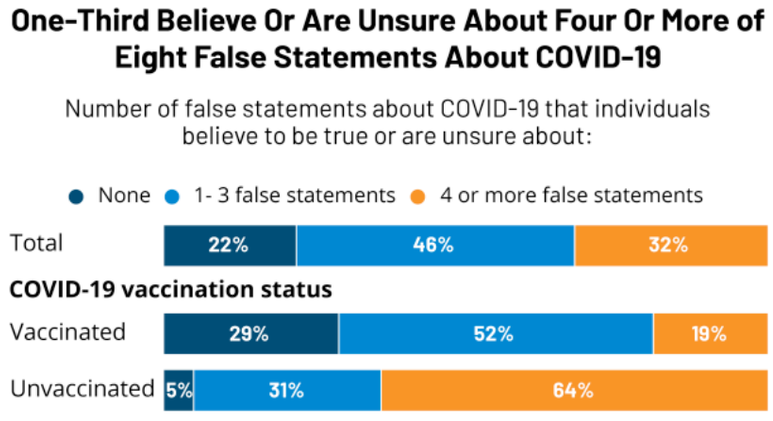


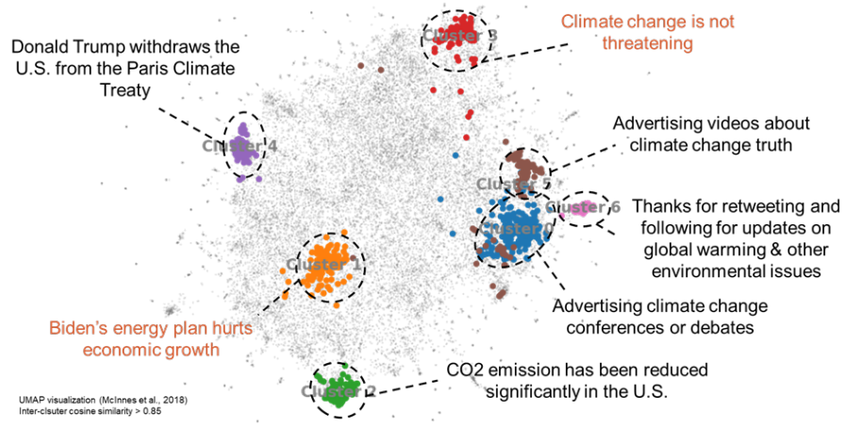


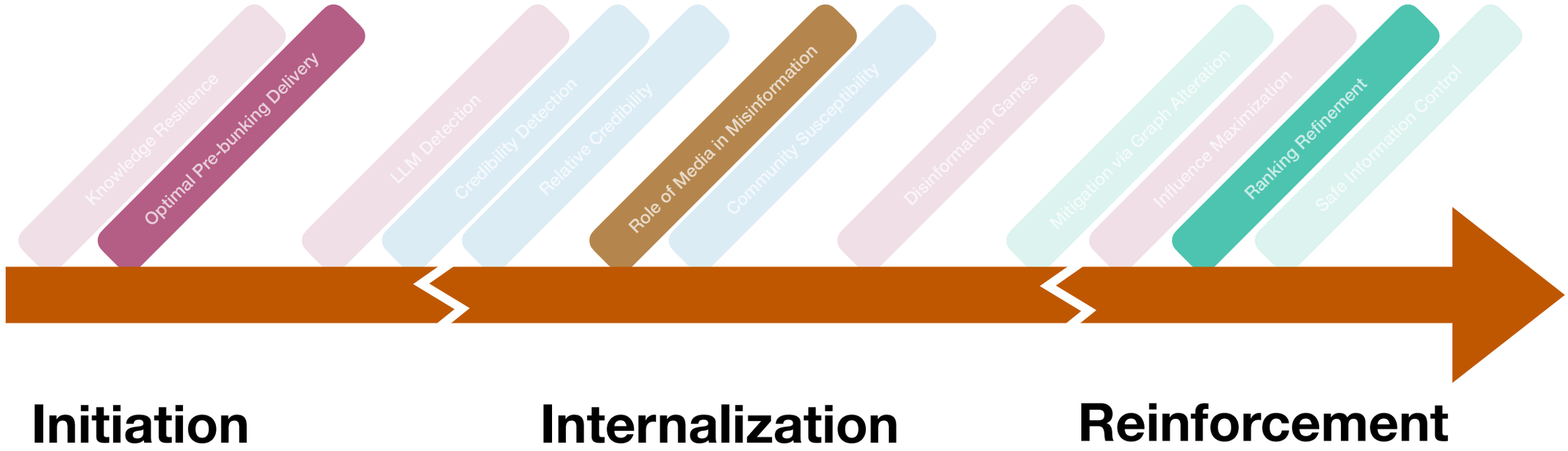
Disinformation Evolution

CENTER FOR aUTonomy





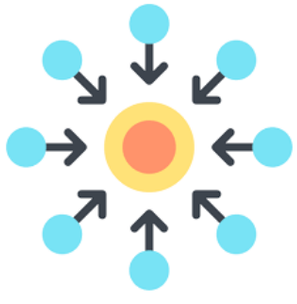




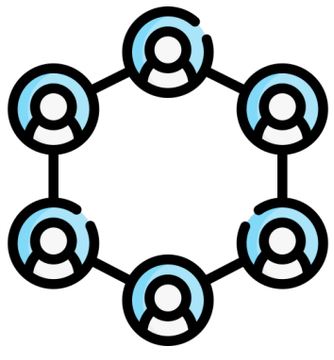
Challenges and Contributions

CENTER FOR aUTonomy

- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



Totalitarian



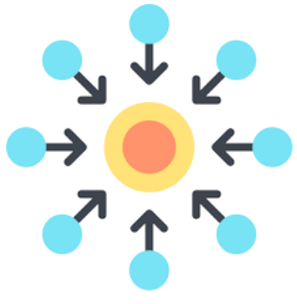
Democratic

[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

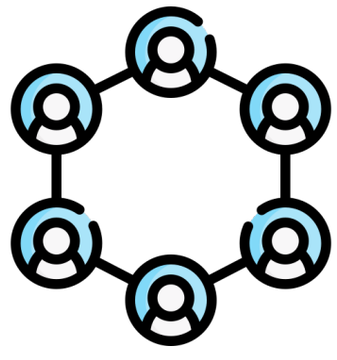
Challenges and Contributions

CENTER FOR aUTonomy

- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



Totalitarian



Democratic

Feasible Solutions

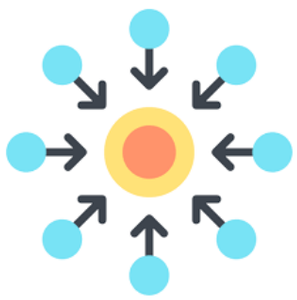
- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

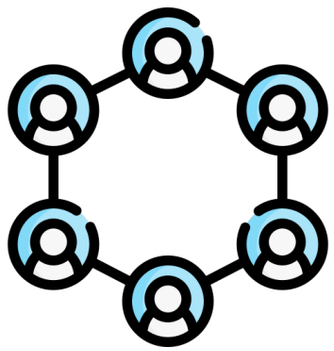
Challenges and Contributions

CENTER FOR aUTonomy

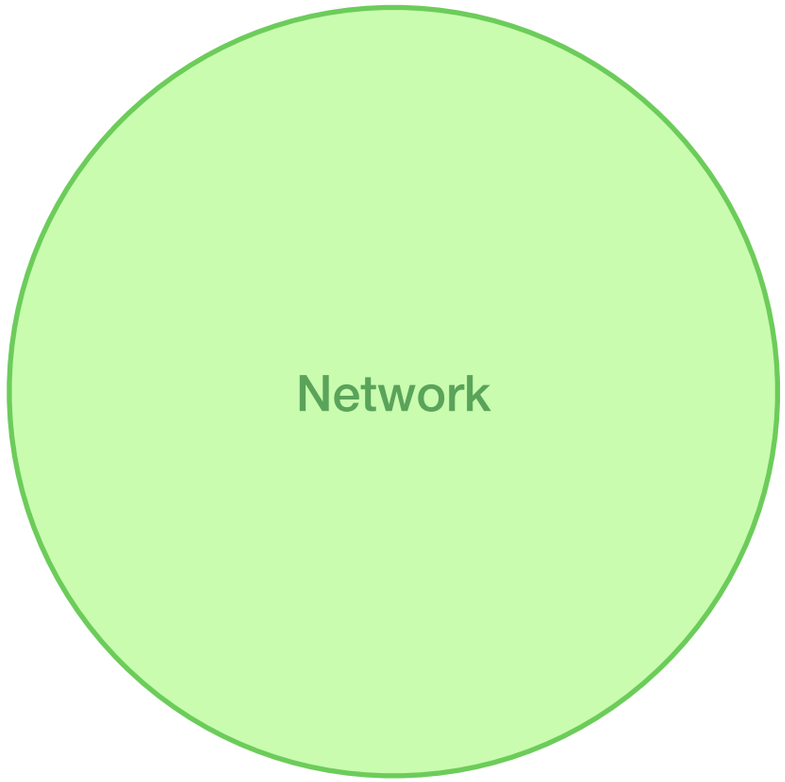
- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



Totalitarian



Democratic



Feasible Solutions

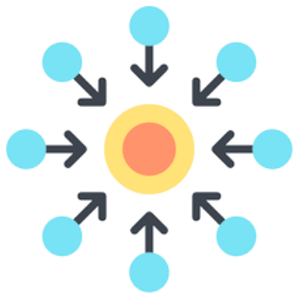
- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

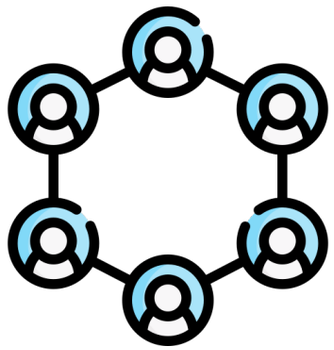
Challenges and Contributions

CENTER FOR aUTonomy

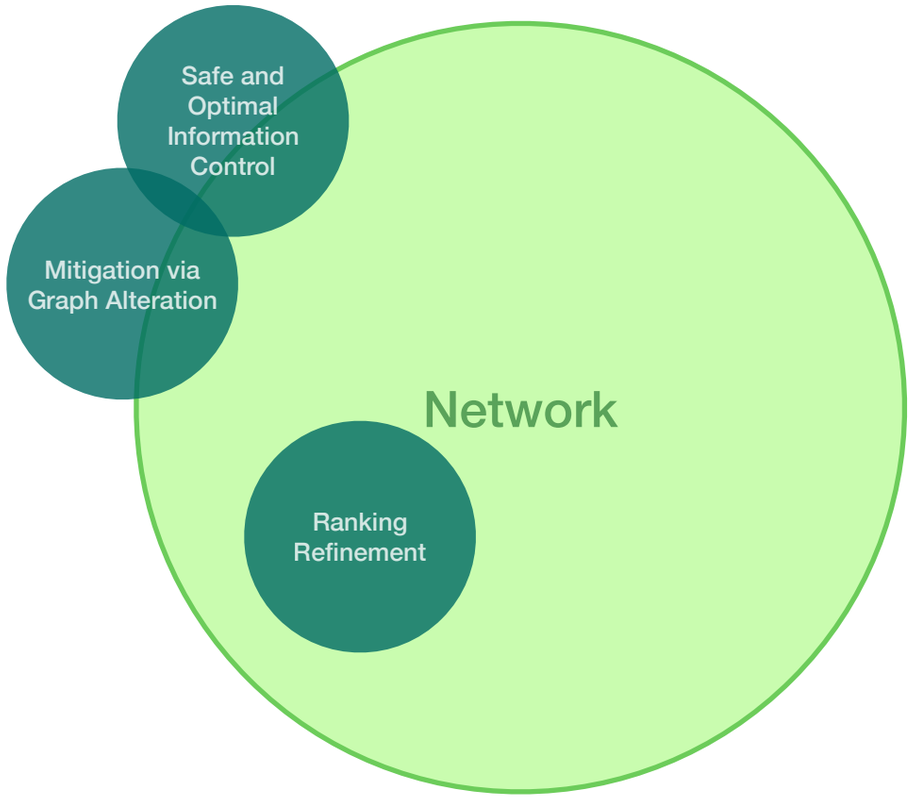
- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



Totalitarian



Democratic



Feasible Solutions

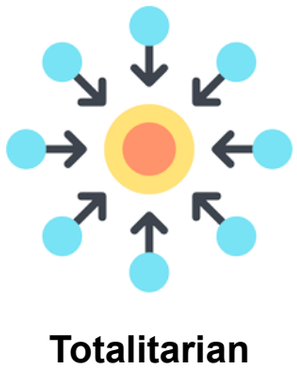
- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

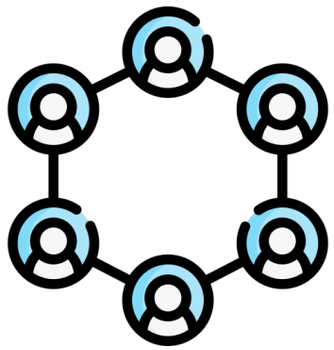
Challenges and Contributions

CENTER FOR aUTonomy

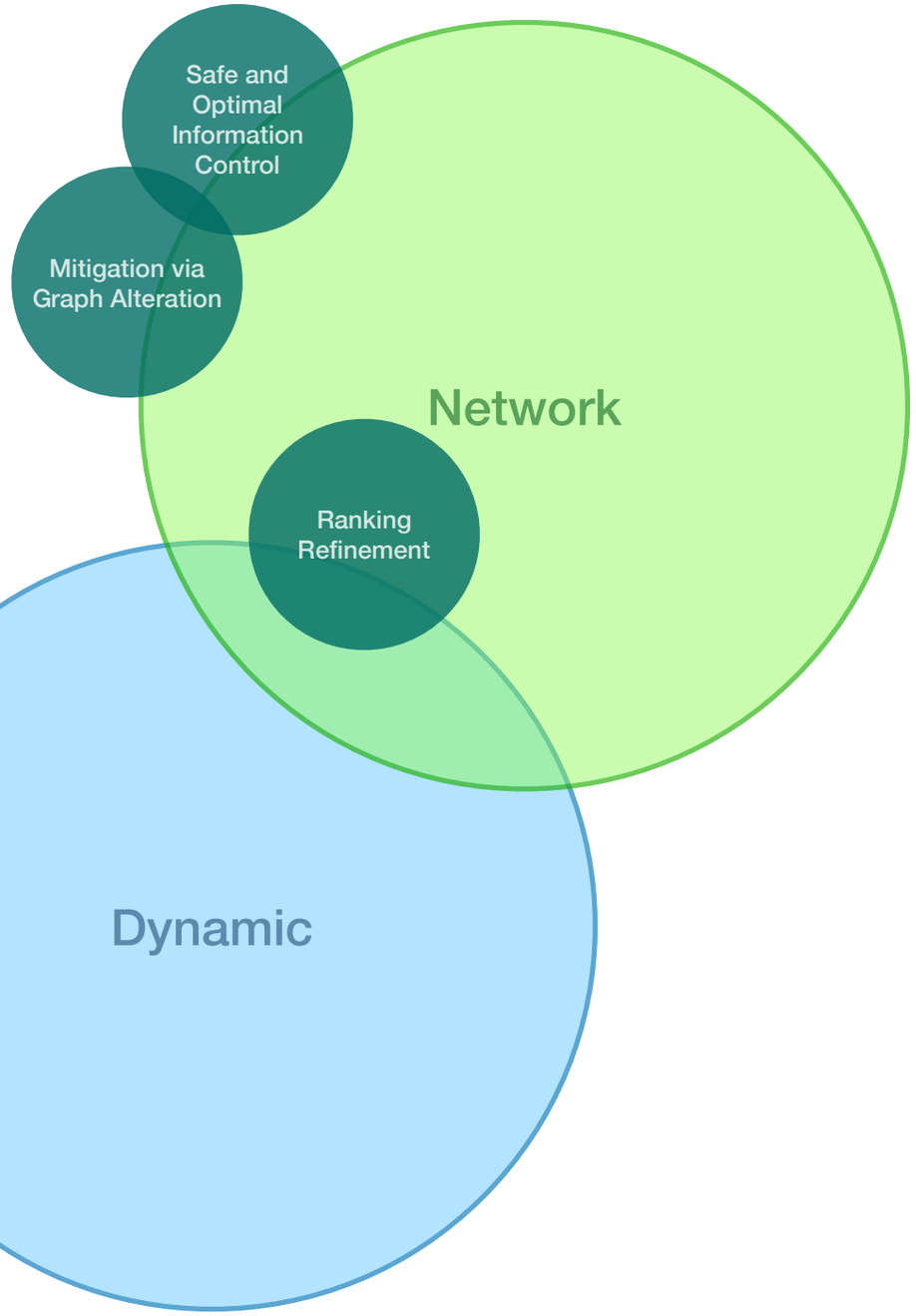
- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



Totalitarian



Democratic



Feasible Solutions

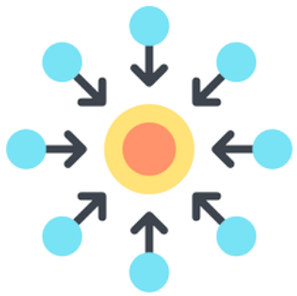
- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

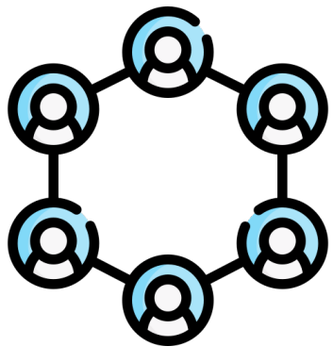
Challenges and Contributions

CENTER FOR aUTonomy

- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



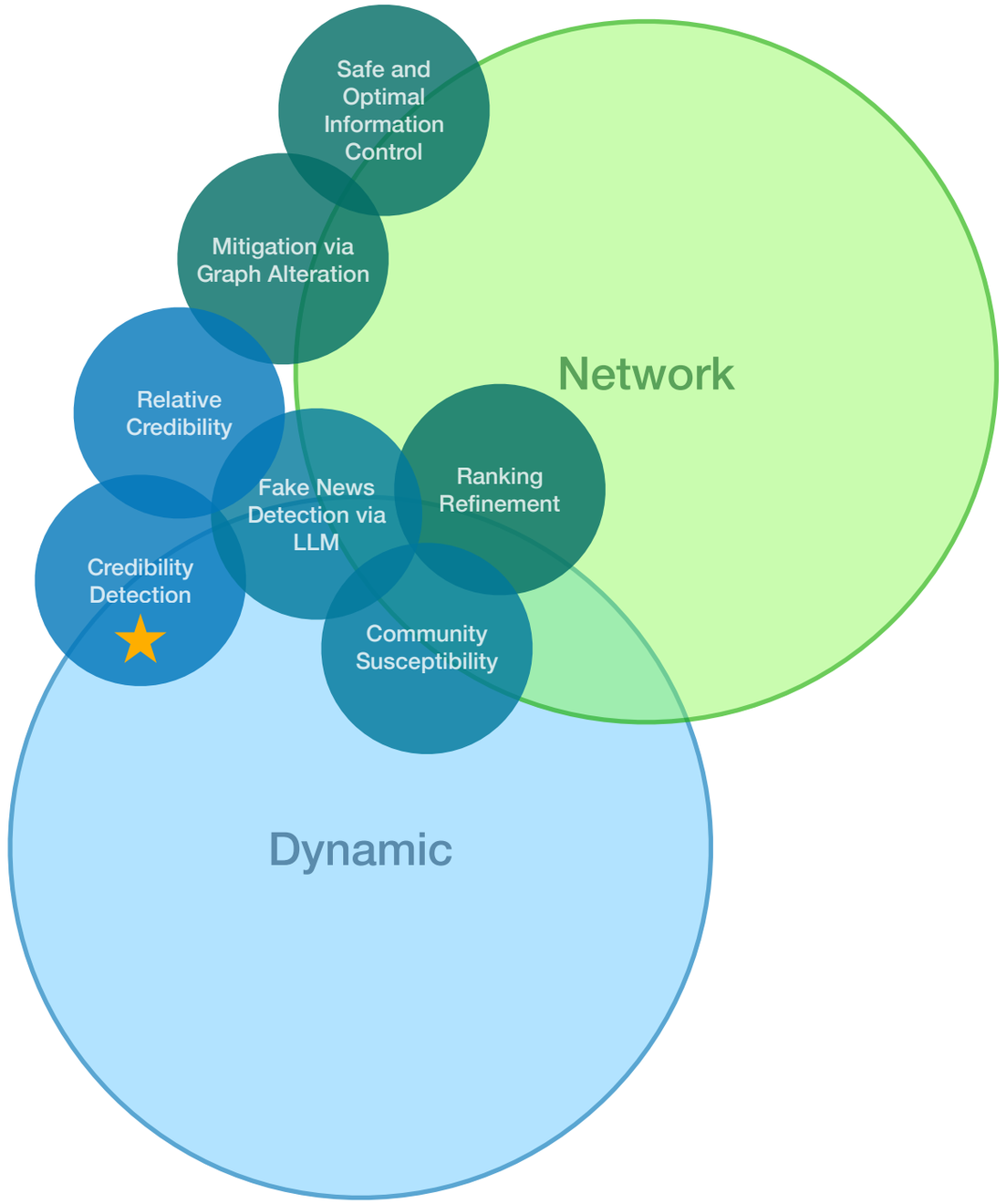
Totalitarian



Democratic

Feasible Solutions

- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

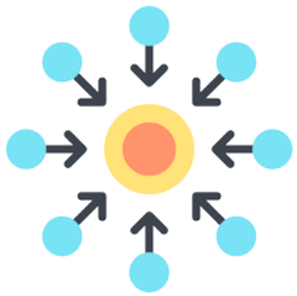


[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

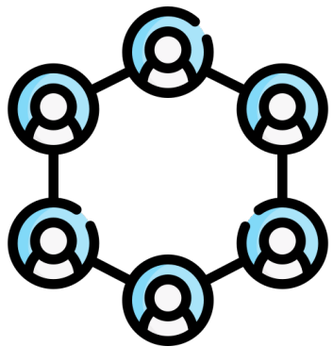
Challenges and Contributions

CENTER FOR aUTonomy

- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



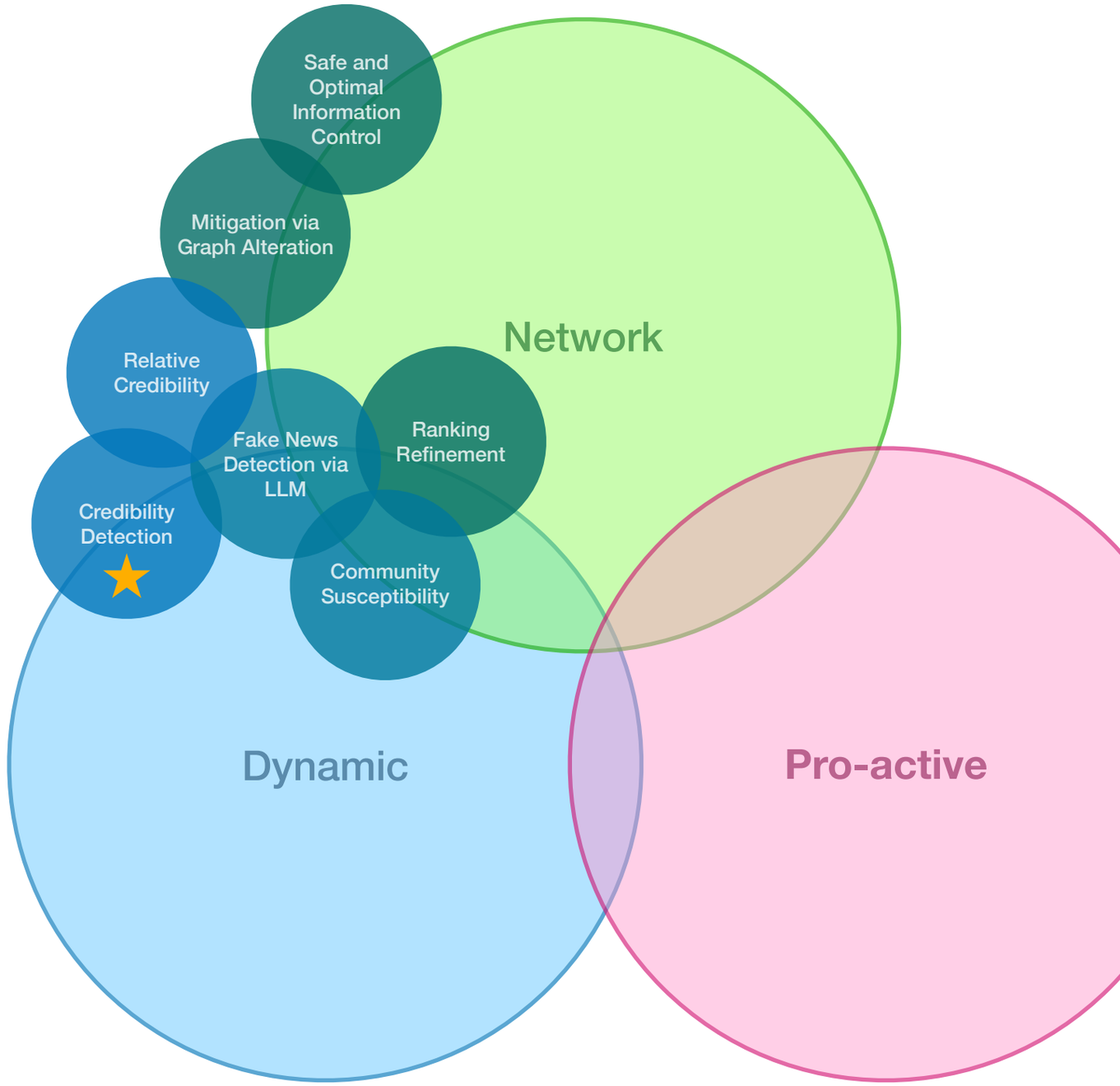
Totalitarian



Democratic

Feasible Solutions

- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

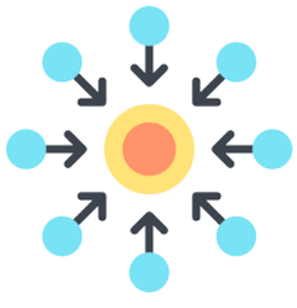


[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

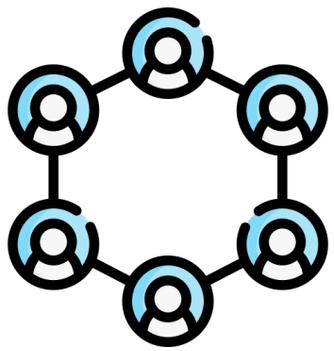
Challenges and Contributions

CENTER FOR aUTonomy

- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



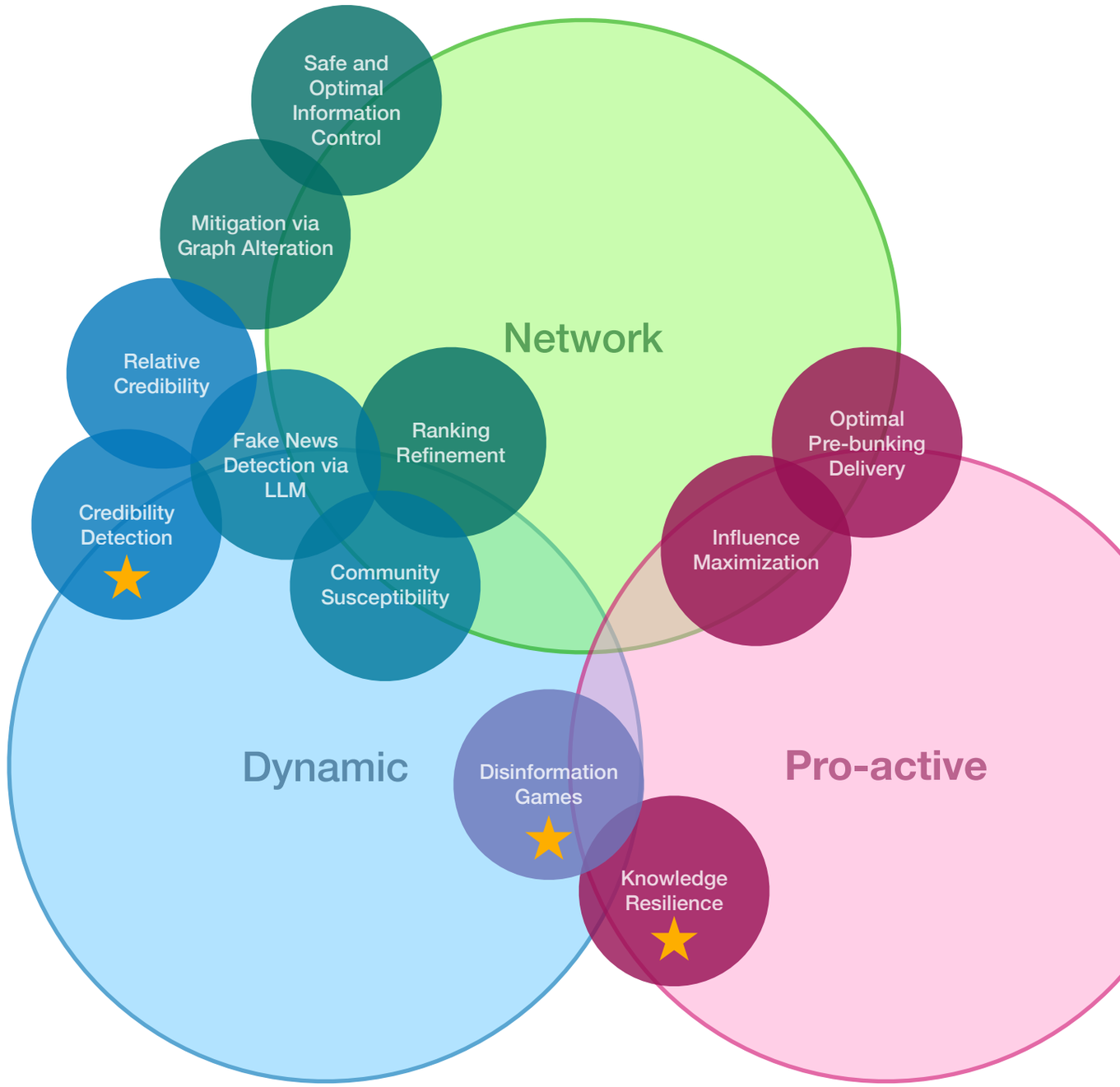
Totalitarian



Democratic

Feasible Solutions

- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders

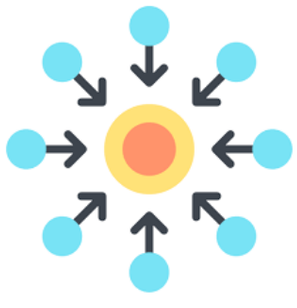


[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

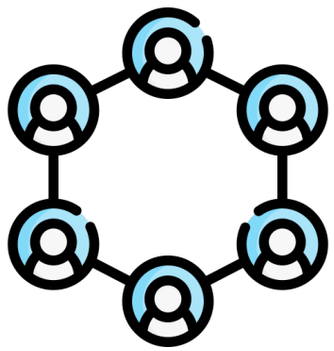
Challenges and Contributions

CENTER FOR aUTonomy

- **Totalitarian:** centralized information management
- **Democratic:** distributed information management



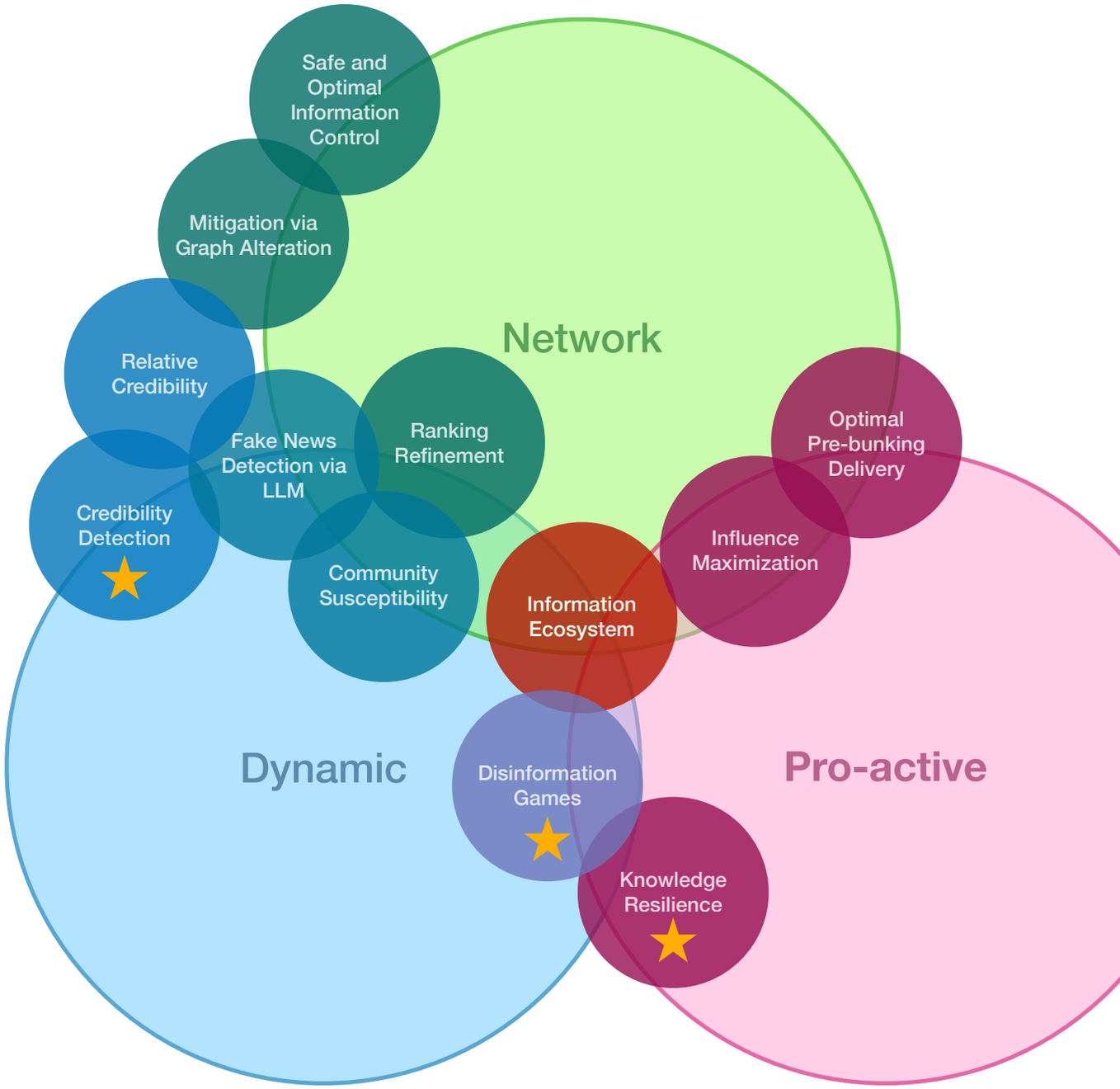
Totalitarian



Democratic

Feasible Solutions

- Democratic
- Resilient
- Safe
- Applicable
- Considers all stakeholders



[1] Harari, Yuval Noah. *Nexus: A brief history of information networks from the Stone Age to AI.* Signal, 2024.

Modeling Information Environment

CENTER FOR aUTonomy

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Environment

News Source Credibility

- What is the chance of a news source sharing misinformation?

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Environment

News Source Credibility

- What is the chance of a news source sharing misinformation?

User Susceptibility

- Chance of a user getting influence under exposure to information from a low credibility source

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Environment

CENTER FOR aUTonomy

News Source Credibility

- What is the chance of a news source sharing misinformation?

User Susceptibility

- Chance of a user getting influence under exposure to information from a low credibility source

Sharing misinformation: gain attention loose integrity

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Environment

CENTER FOR aUTonomy

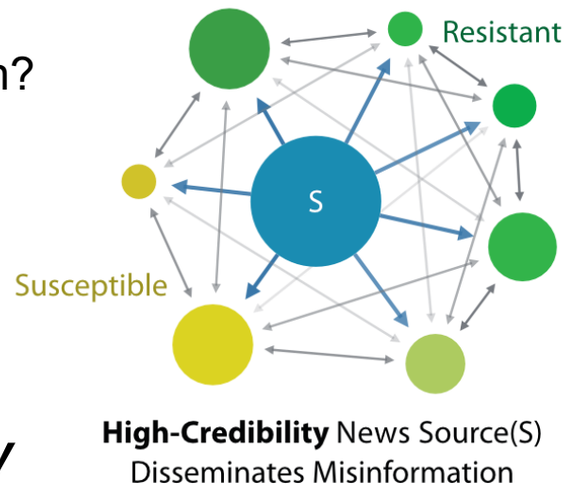
News Source Credibility

- What is the chance of a news source sharing misinformation?

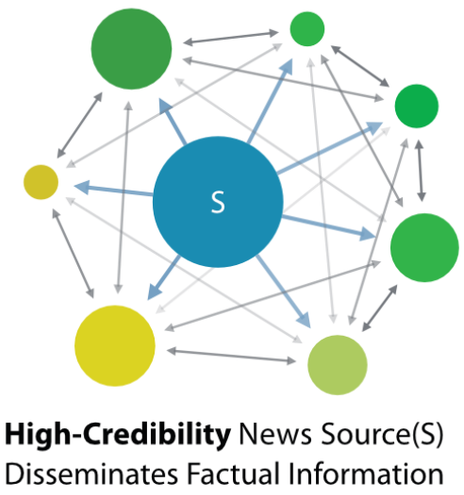
User Susceptibility

- Chance of a user getting influence under exposure to information from a low credibility source

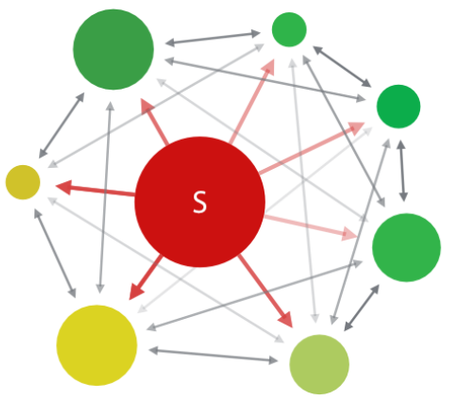
Sharing misinformation: gain attention loose integrity



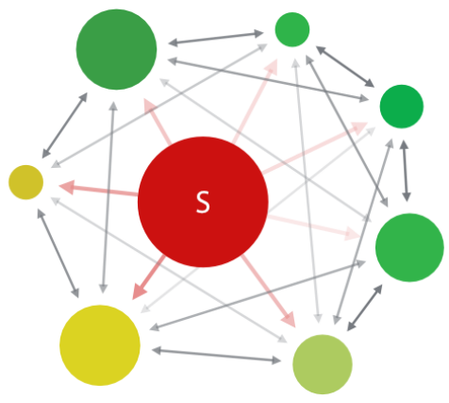
High-Credibility News Source(S) Disseminates Misinformation



High-Credibility News Source(S) Disseminates Factual Information



Low-Credibility News Source(S) Disseminates Misinformation



Low-Credibility News Source(S) Disseminates Factual Information

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Environment

CENTER FOR aUTonomy

News Source Credibility

- What is the chance of a news source sharing misinformation?

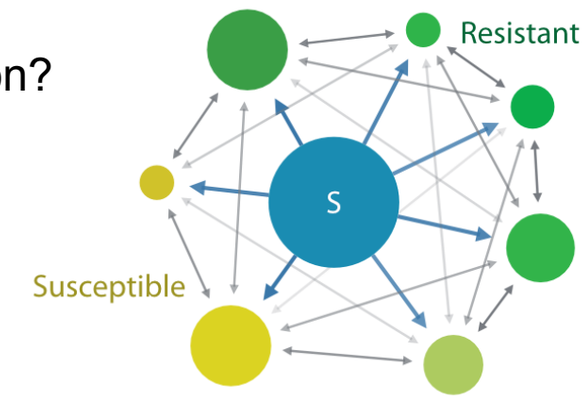
User Susceptibility

- Chance of a user getting influence under exposure to information from a low credibility source

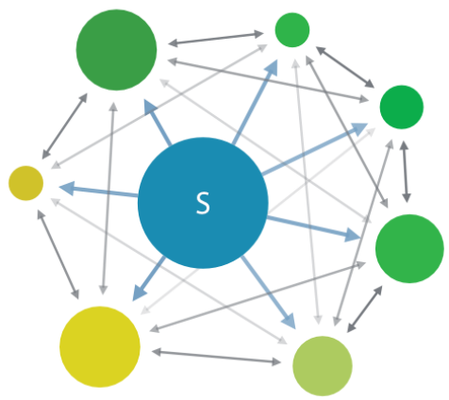
Sharing misinformation: gain attention loose integrity

How users beliefs change under exposure to misinformation?

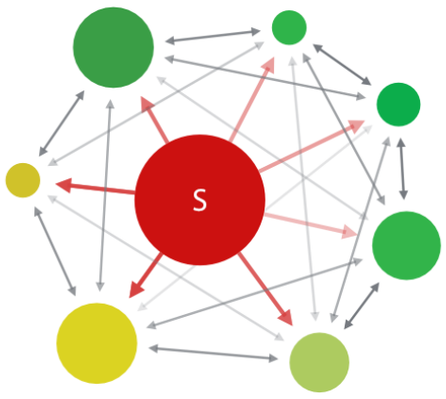
$$x_{t+1}^i = \underbrace{\frac{a}{A_t^i}\sum_{j=1}^{N}\phi(|x_t^i - x_t^j|)(x_t^j - x_t^i)}_{\text{Social influence}} + \underbrace{\frac{b}{B_t^i}\sum_{m=1}^{M}\psi(|x_t^i - y^m|, c_t^m, a_t^m, s^i)(y^m - x_t^i)}_{\text{Media influence}} + \sigma w_t^i,$$



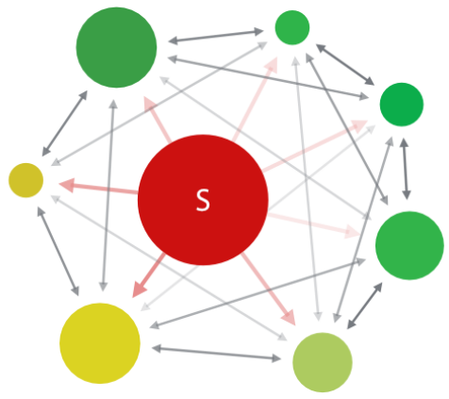
High-Credibility News Source(S) Disseminates Misinformation



High-Credibility News Source(S) Disseminates Factual Information



Low-Credibility News Source(S) Disseminates Misinformation



Low-Credibility News Source(S) Disseminates Factual Information

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Enviroment

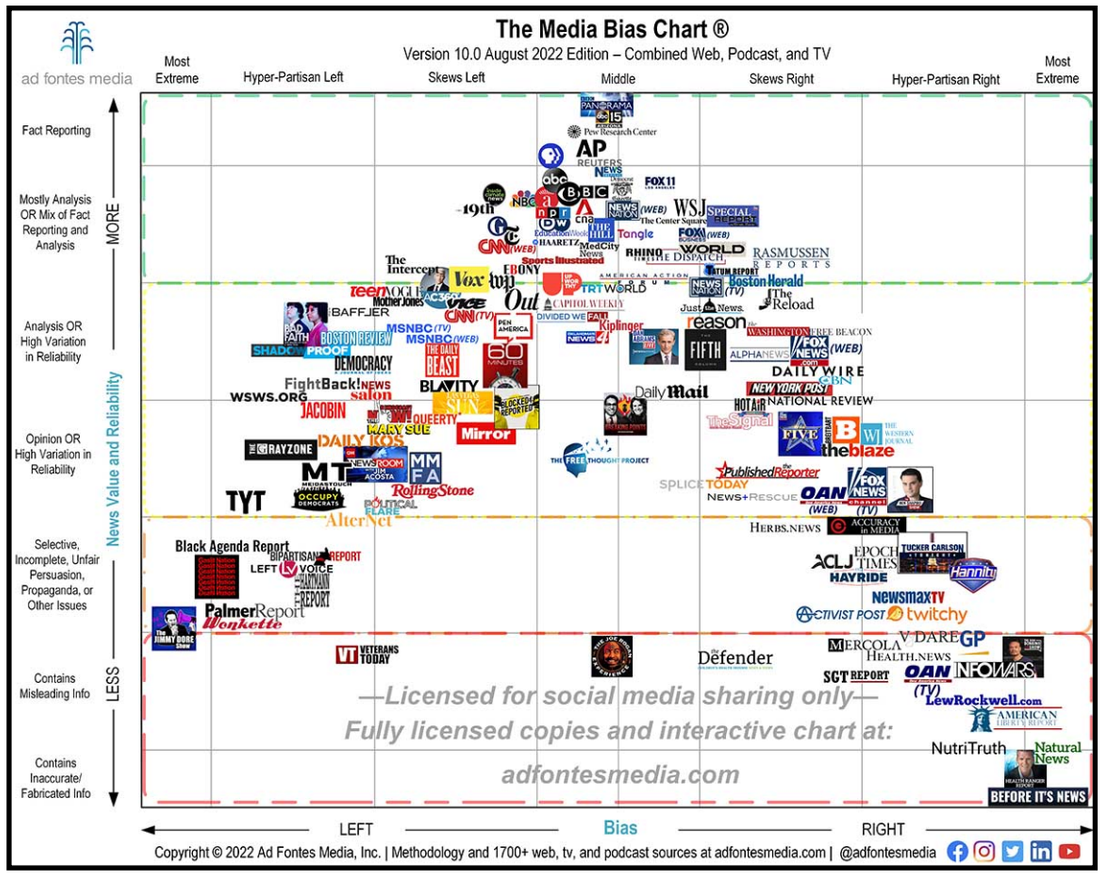
CENTER FOR aUTonomy

[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Enviroment

CENTER FOR aUTonomy

Source Credibility

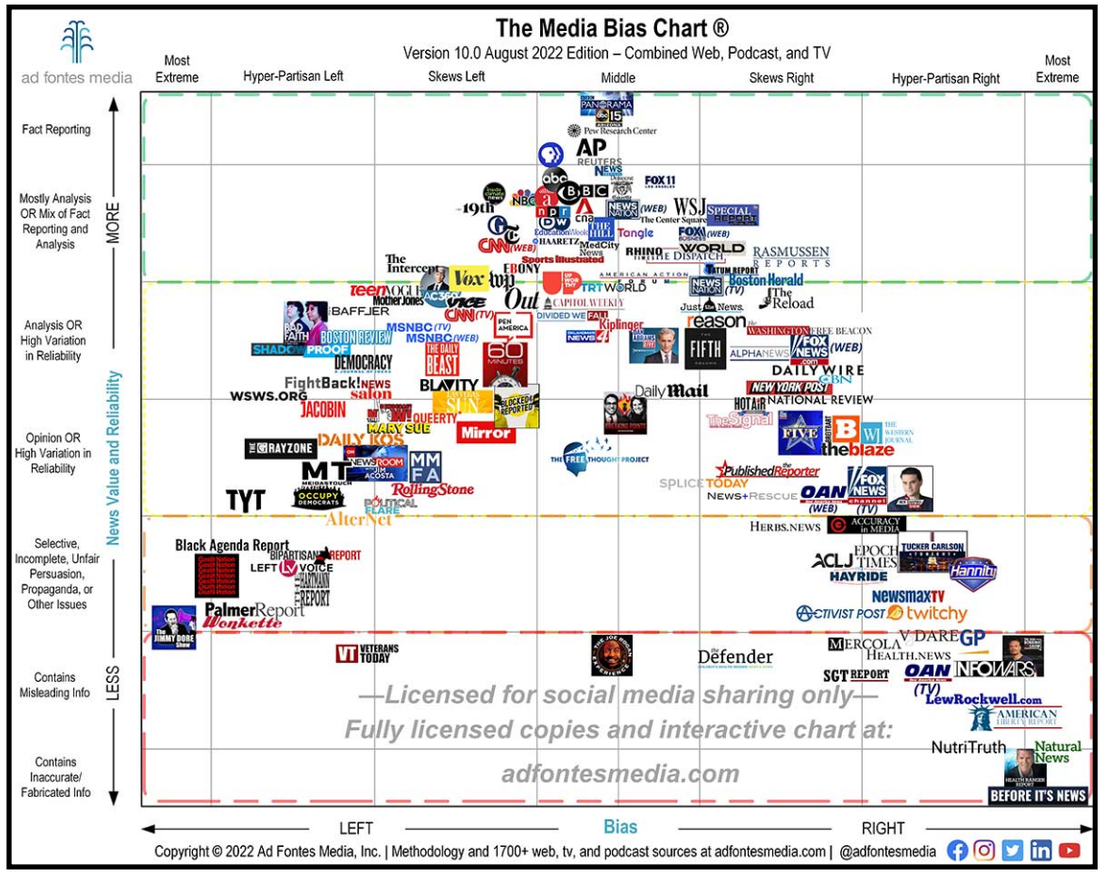


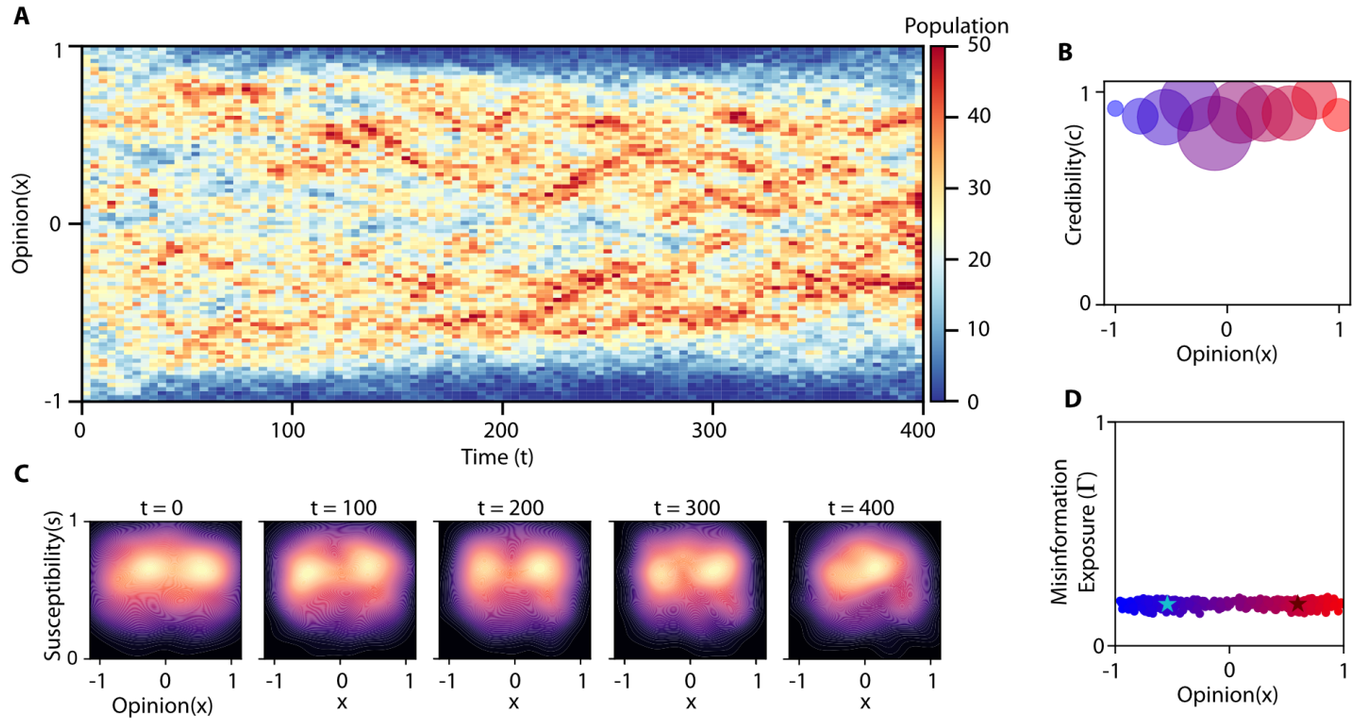
[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Enviroment

CENTER FOR aUTonomy

Source Credibility



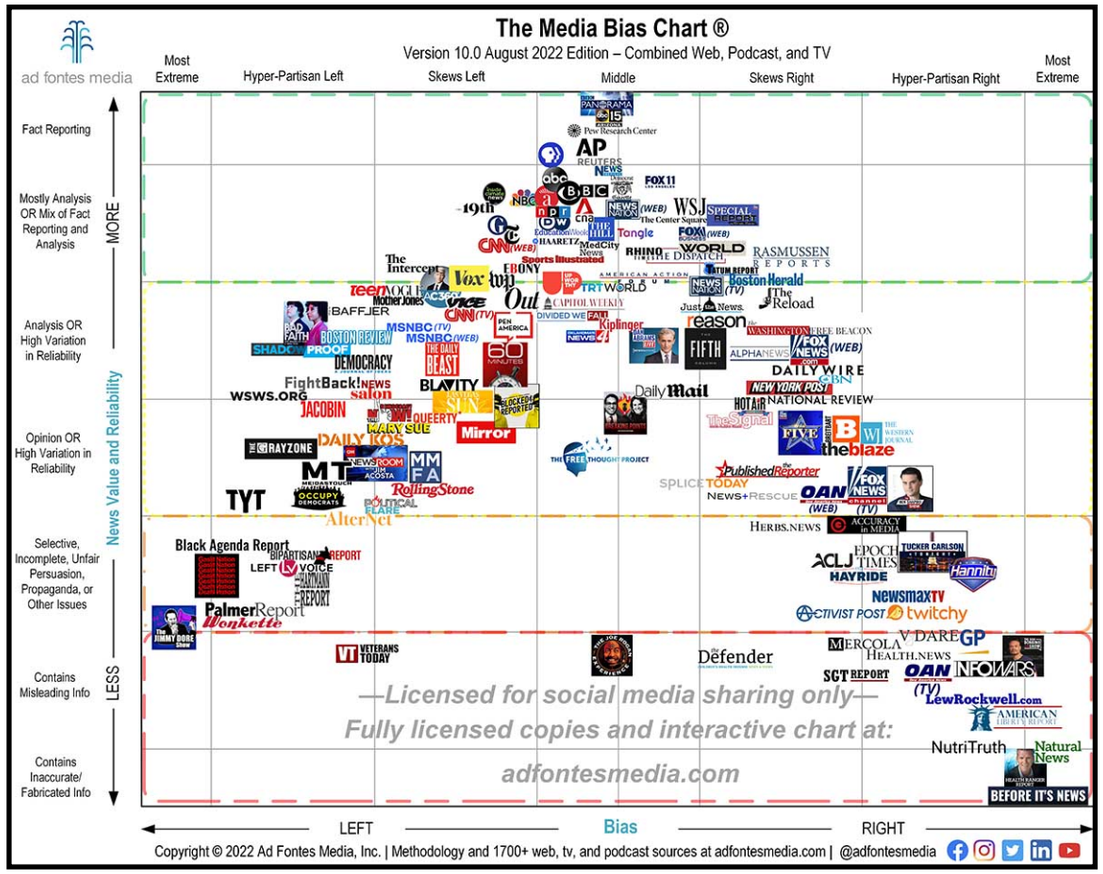


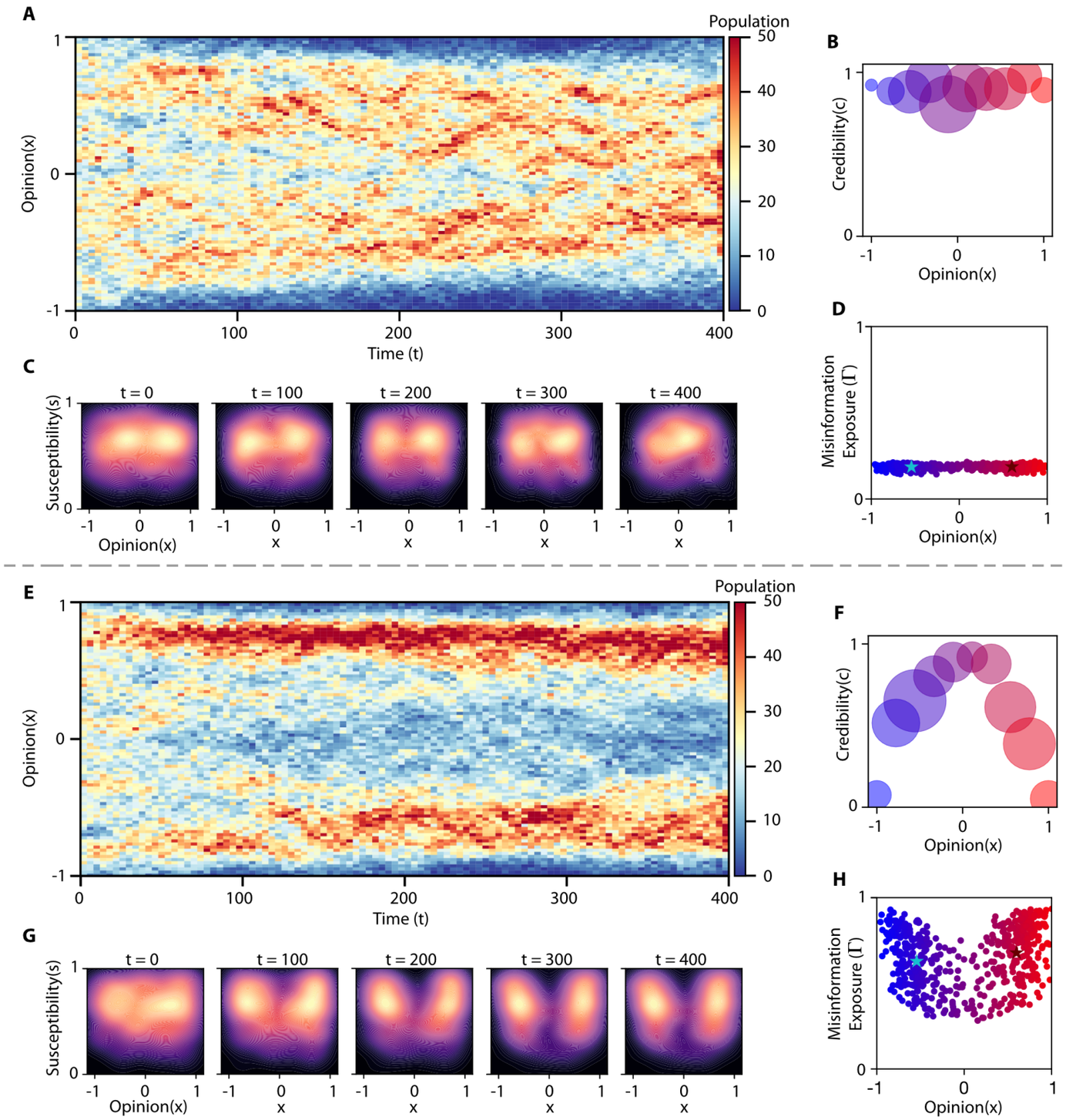
[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Modeling Information Enviroment

CENTER FOR aUTonomy

Source Credibility



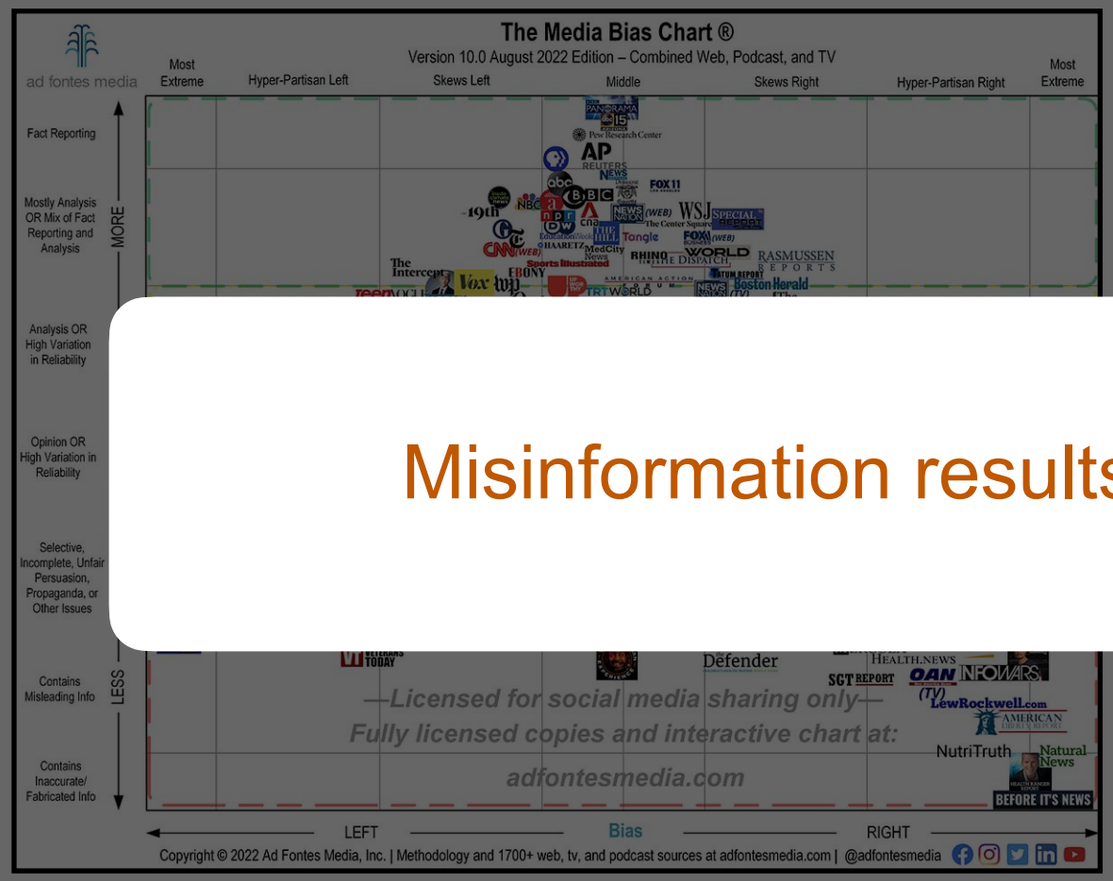


[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

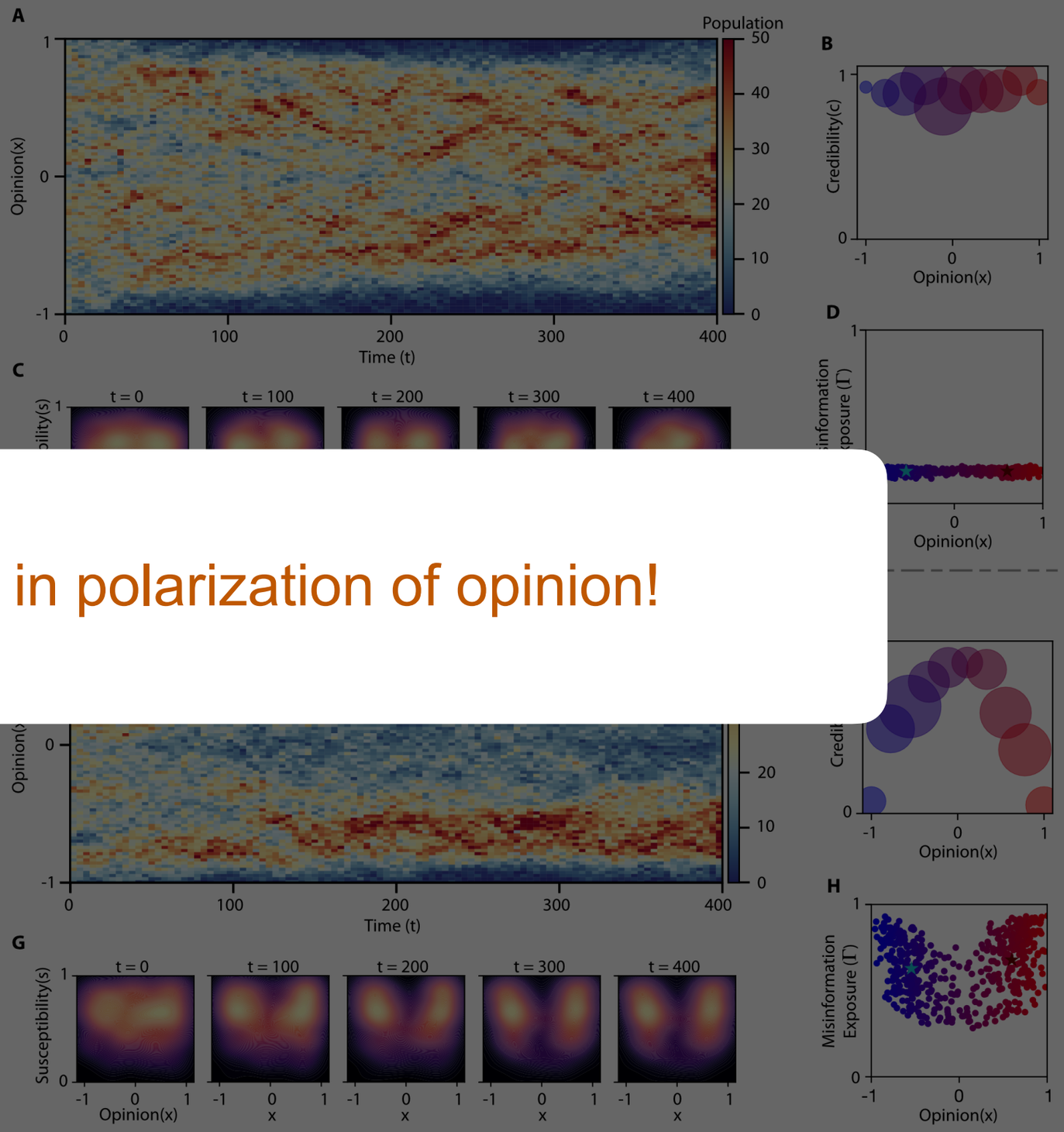
Modeling Information Enviroment

CENTER FOR aUTonomy

Source Credibility



Misinformation results in polarization of opinion!



[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

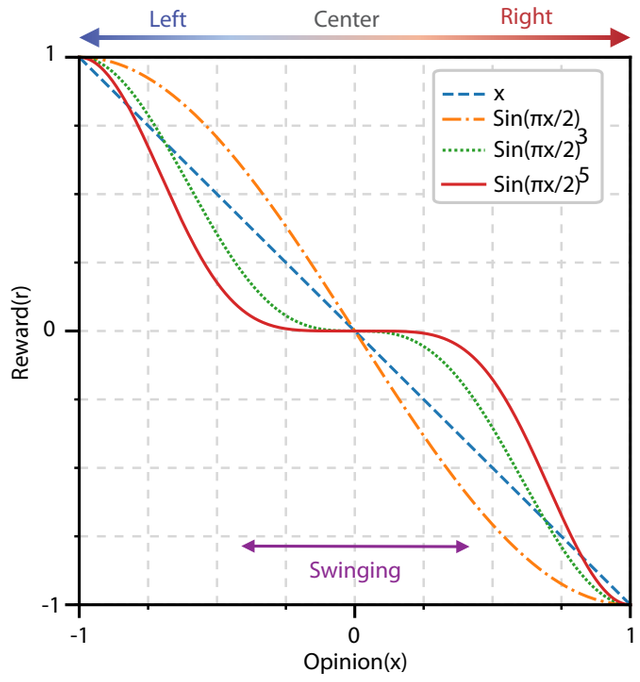
Gain Influence Loose Credibility

CENTER FOR aUTonomy

$$\max_{\pi^L} \min_{\pi^R} \qquad J := \mathbb{E}\Big[\sum_{k=1}^{\infty} \gamma^k r(x_k)\Big]$$

$$x_{t+1}^i = \underbrace{\frac{a}{A_t^i}\sum_{j=1}^{N}\phi(|x_t^i - x_t^j|)(x_t^j - x_t^i)}_{\text{Social influence}} + \underbrace{\frac{b}{B_t^i}\sum_{m=1}^{M}\psi(|x_t^i - y^m|, c_t^m, a_t^m, s^i)(y^m - x_t^i)}_{\text{Media influence}} + \sigma w_t^i,$$

$$c_{t+1}^m = \lambda c_t^m + (1-\lambda)a_t^m,$$



[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

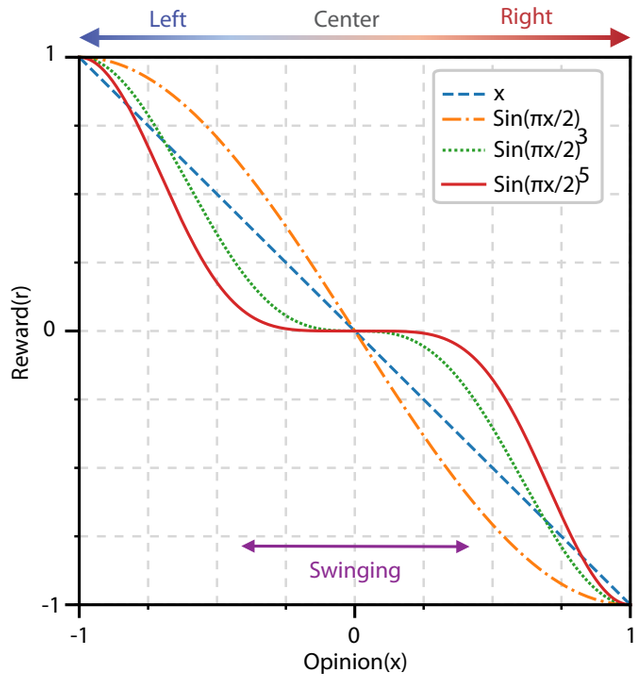
Gain Influence Loose Credibility

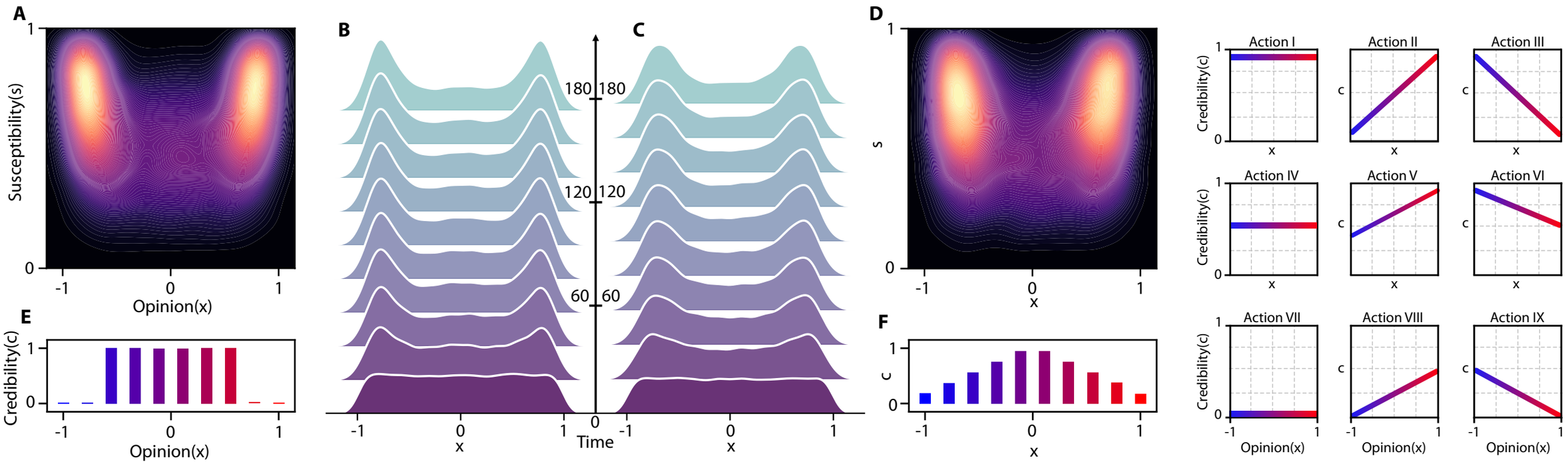
CENTER FOR aUTonomy

$$\max_{\pi^L} \min_{\pi^R} \qquad J := \mathbb{E}\Big[\sum_{k=1}^{\infty} \gamma^k r(x_k)\Big]$$

$$x_{t+1}^i = \underbrace{\frac{a}{A_t^i}\sum_{j=1}^{N}\phi(|x_t^i - x_t^j|)(x_t^j - x_t^i)}_{\text{Social influence}} + \underbrace{\frac{b}{B_t^i}\sum_{m=1}^{M}\psi(|x_t^i - y^m|, c_t^m, a_t^m, s^i)(y^m - x_t^i)}_{\text{Media influence}} + \sigma w_t^i,$$

$$c_{t+1}^m = \lambda c_t^m + (1-\lambda)a_t^m,$$



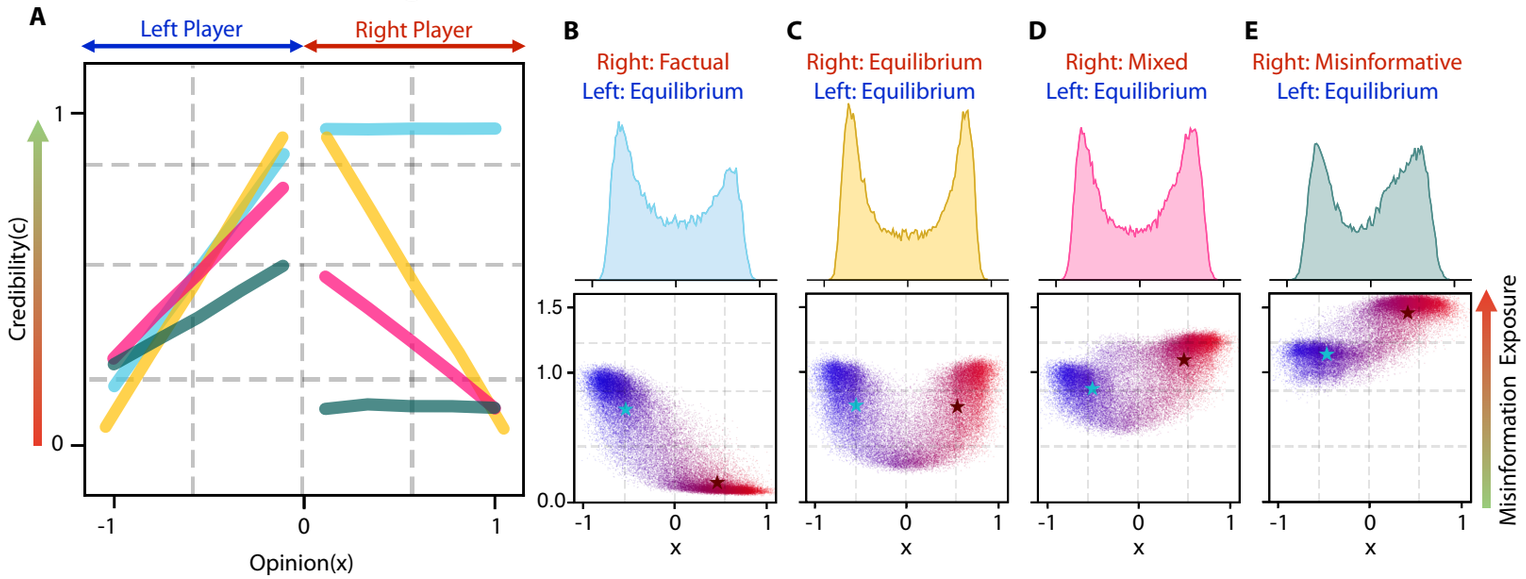


[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Arm Race

CENTER FOR aUTonomy

If a player **increases misinformation dissemination** below equilibrium, the **optimal response** of the other player leads to *greater misinformation sharing*.

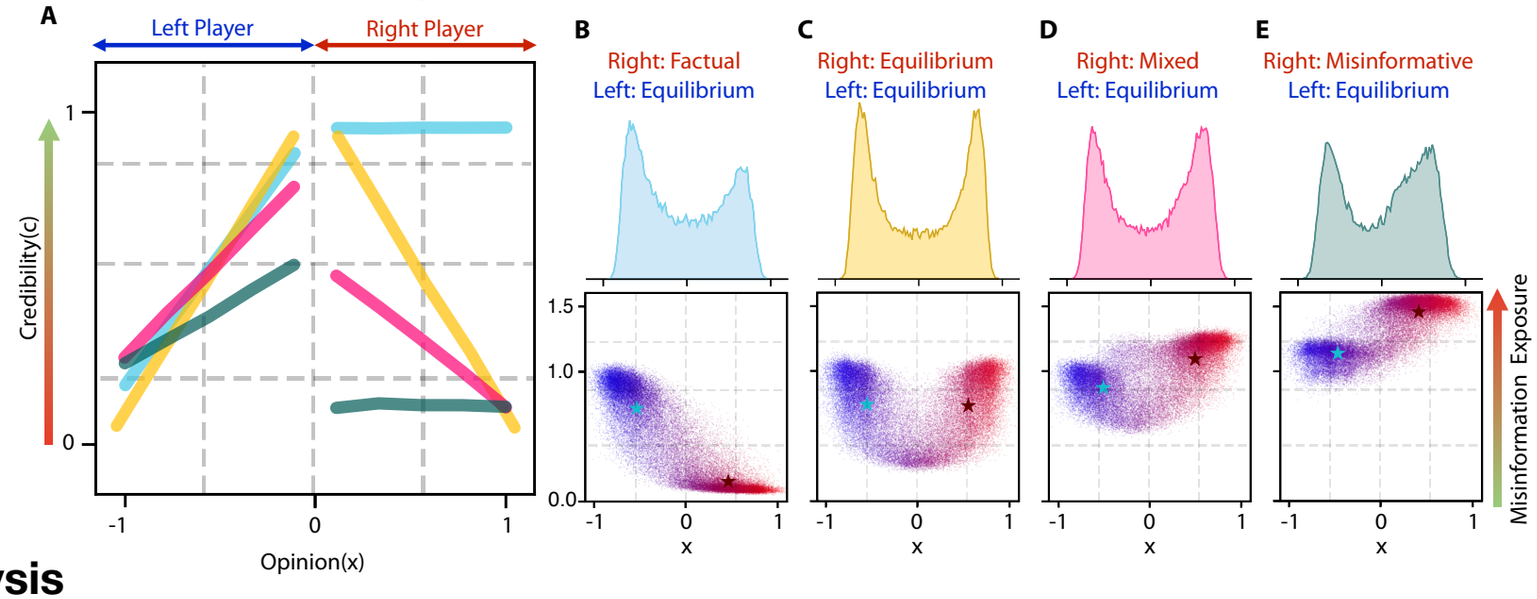


[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

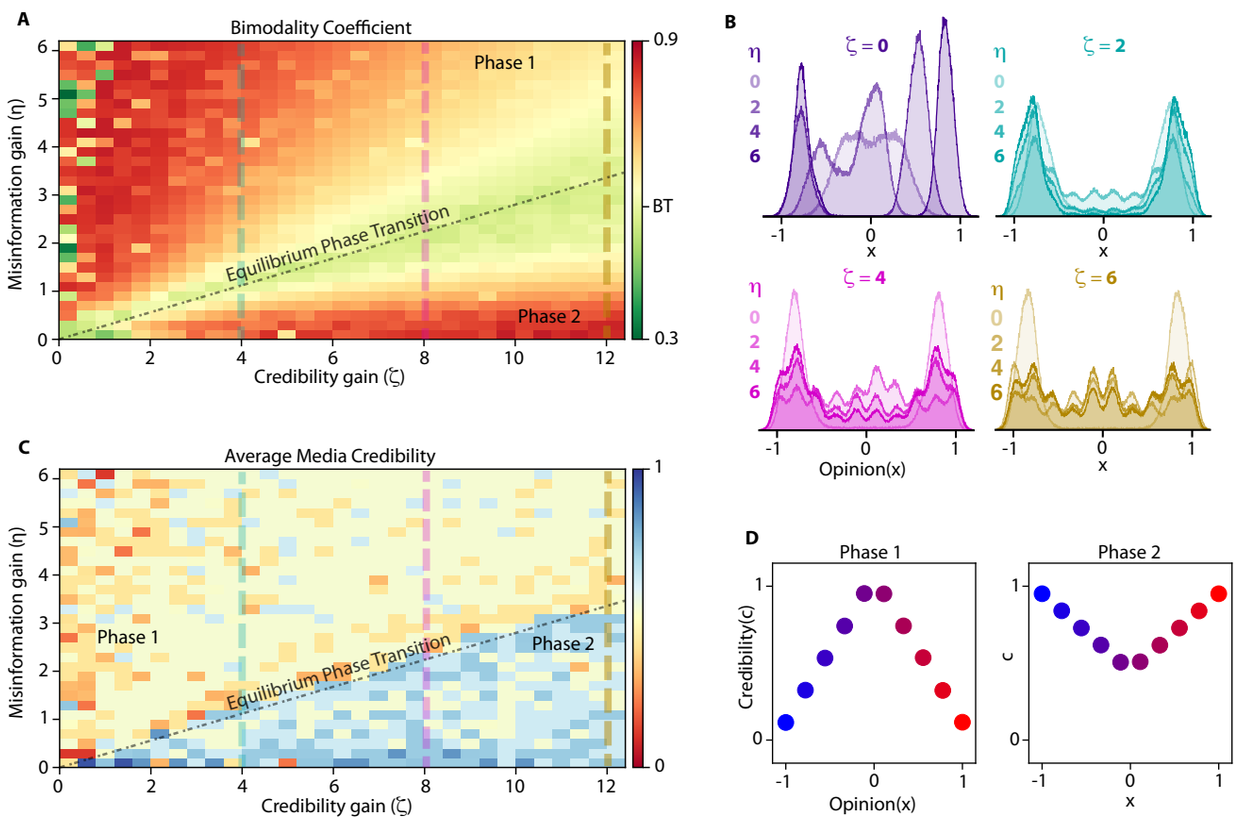
Arm Race

CENTER FOR aUTonomy

If a player **increases misinformation dissemination** below equilibrium, the **optimal response** of the other player leads to *greater misinformation sharing*.



Sensitivity Analysis

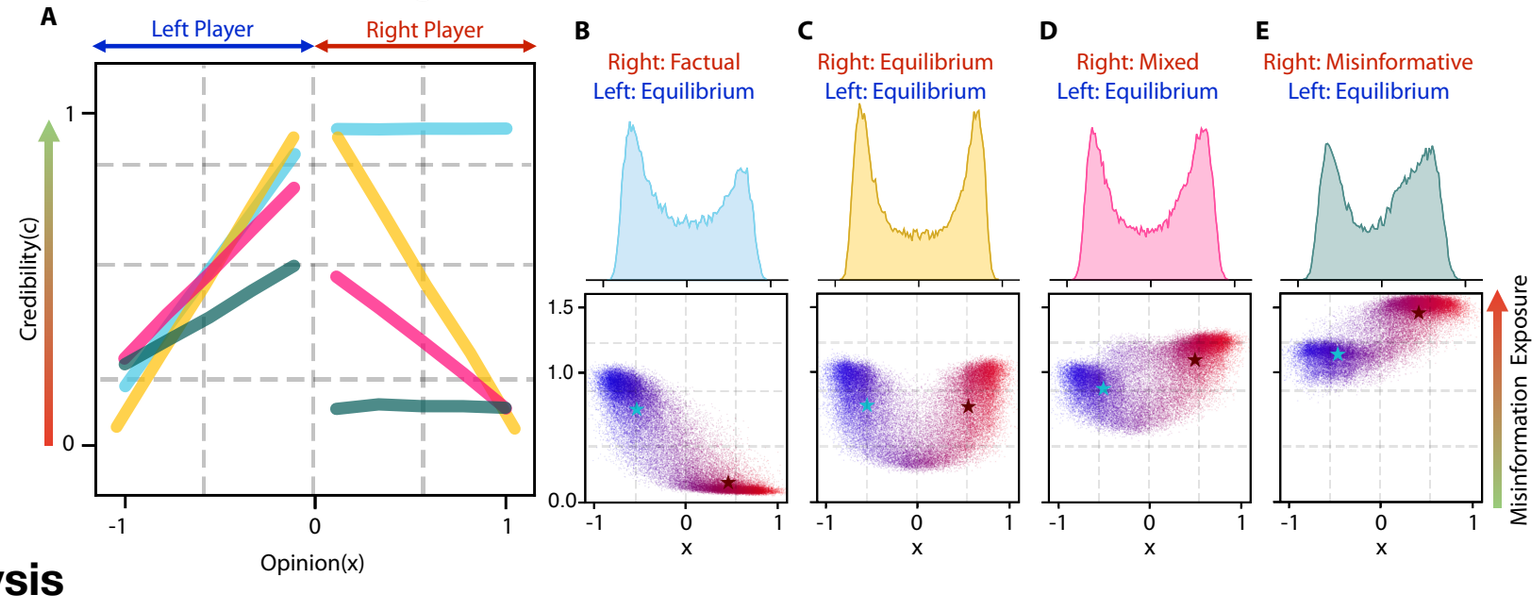


[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

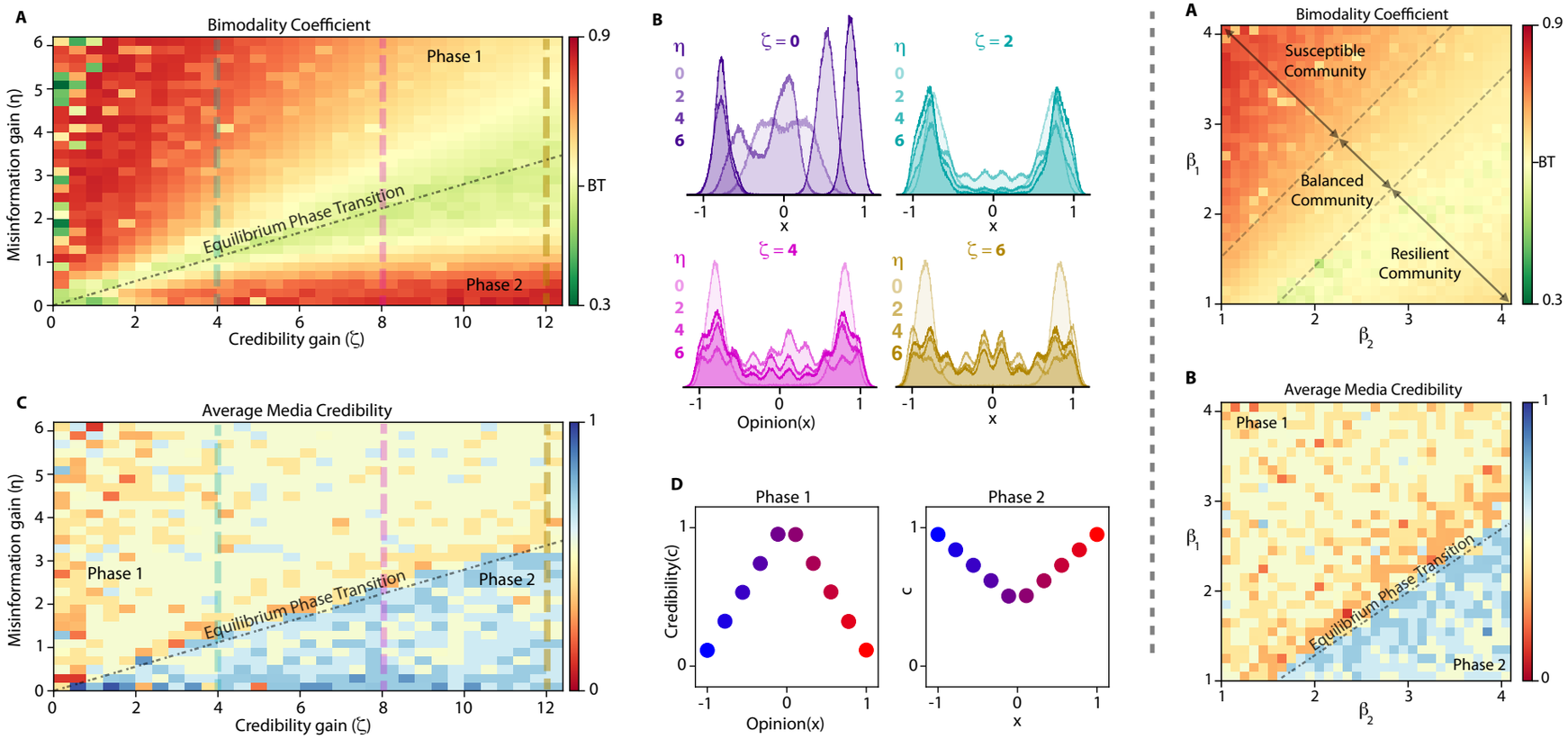
Arm Race

CENTER FOR aUTonomy

If a player **increases misinformation dissemination** below equilibrium, the **optimal response** of the other player leads to *greater misinformation sharing*.



Sensitivity Analysis



[1] Amini, Arash, et al. *How Media Competition Fuels the Spread of Misinformation*. Science Advances [Under review]

Take Away

Misinformation Interventions

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.
- **Community Resilience**: Lowering susceptibility through education and media literacy fosters a less polarized, more informed society.

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.
- **Community Resilience**: Lowering susceptibility through education and media literacy fosters a less polarized, more informed society.
- **Arms Race Dynamics**: Reducing credibility can trigger a misinformation escalation, while credible biased sources encourage factual reporting.

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.
- **Community Resilience**: Lowering susceptibility through education and media literacy fosters a less polarized, more informed society.
- **Arms Race Dynamics**: Reducing credibility can trigger a misinformation escalation, while credible biased sources encourage factual reporting.

Conclusion

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.
- **Community Resilience**: Lowering susceptibility through education and media literacy fosters a less polarized, more informed society.
- **Arms Race Dynamics**: Reducing credibility can trigger a misinformation escalation, while credible biased sources encourage factual reporting.

Conclusion

- Misinformation is a *dynamic* problem.

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.
- **Community Resilience**: Lowering susceptibility through education and media literacy fosters a less polarized, more informed society.
- **Arms Race Dynamics**: Reducing credibility can trigger a misinformation escalation, while credible biased sources encourage factual reporting.

Conclusion

- Misinformation is a *dynamic* problem.
- Is this outcome the *fault* of the media or the information ecosystem?

Misinformation Interventions

- **Debunking Challenges**: While fact-checking reduces misinformation, its impact depends on source credibility and audience alignment.
- **Source Accountability**: Increasing credibility gain by holding sources accountable reduces misinformation and polarization.
- **Community Resilience**: Lowering susceptibility through education and media literacy fosters a less polarized, more informed society.
- **Arms Race Dynamics**: Reducing credibility can trigger a misinformation escalation, while credible biased sources encourage factual reporting.

Conclusion

- Misinformation is a *dynamic* problem.
- Is this outcome the *fault* of the media or the information ecosystem?
- The current information enviromen is *explotaible*

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Points of intervention

[1] van der Linden, S. *Misinformation: susceptibility, spread, and interventions to immunize the public.* Nature Medicine, 2022

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Points of intervention

Option 1: Debunking

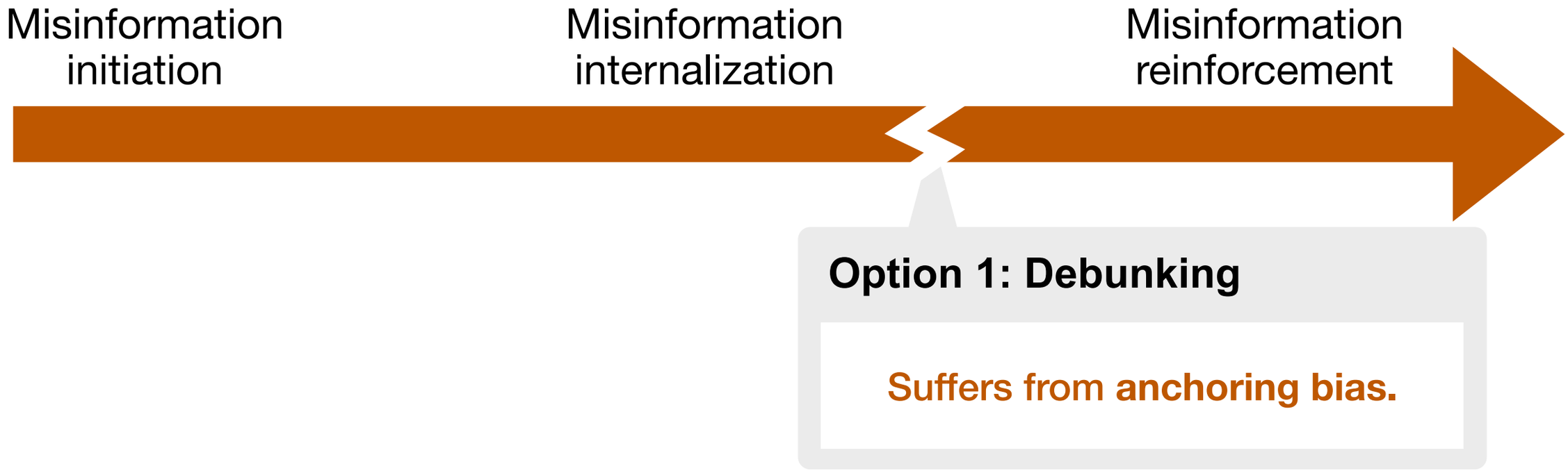
Identify misinformed users.
Correct internalized misinformation by debunking their perceived beliefs.

[1] van der Linden, S. *Misinformation: susceptibility, spread, and interventions to immunize the public.* Nature Medicine, 2022

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Points of intervention



[1] van der Linden, S. *Misinformation: susceptibility, spread, and interventions to immunize the public.* Nature Medicine, 2022

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Points of intervention

Option 2: Pre-bunking

Identify at-risk users.

Educate users to be resilient to future misinformation cascades

Option 1: Debunking

Suffers from anchoring bias.

[1] van der Linden, S. *Misinformation: susceptibility, spread, and interventions to immunize the public.* Nature Medicine, 2022

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Points of intervention

Misinformation initiation → Misinformation internalization → Misinformation reinforcement

Option 2: Pre-bunking

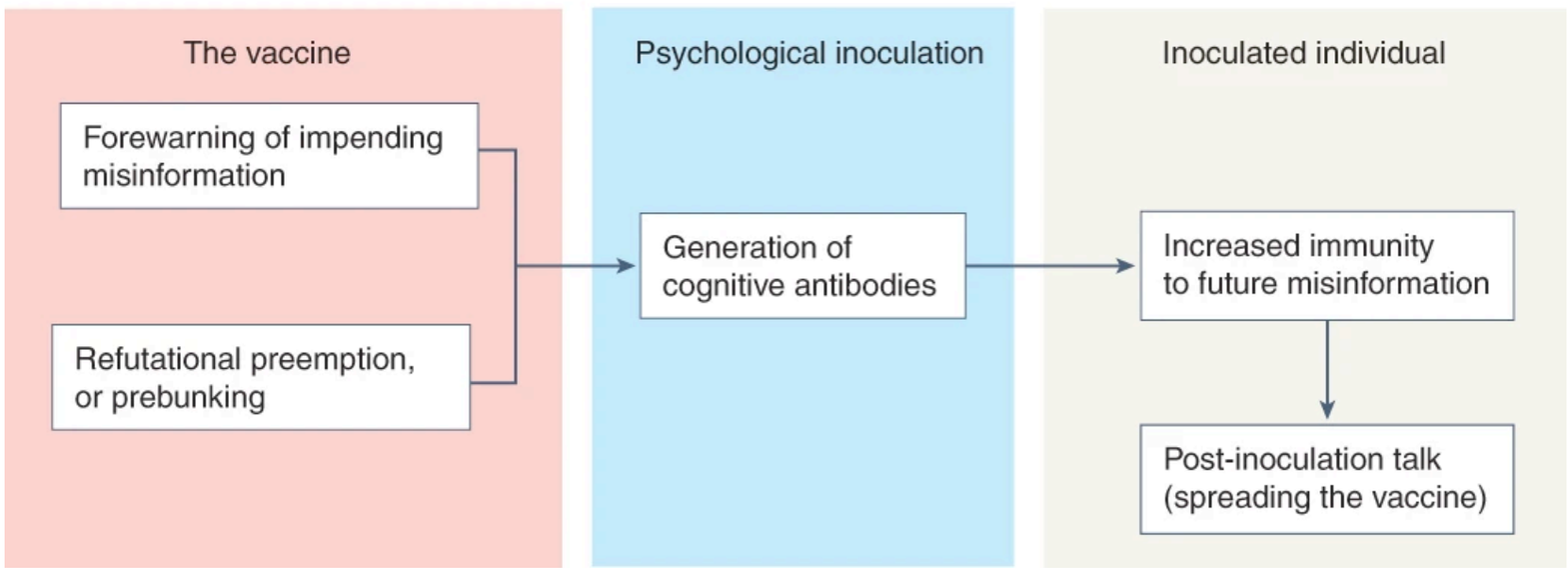
Identify at-risk users.

Educate users to be resilient to future misinformation cascades

Option 1: Debunking

Suffers from anchoring bias.

Theory of psychological inoculation

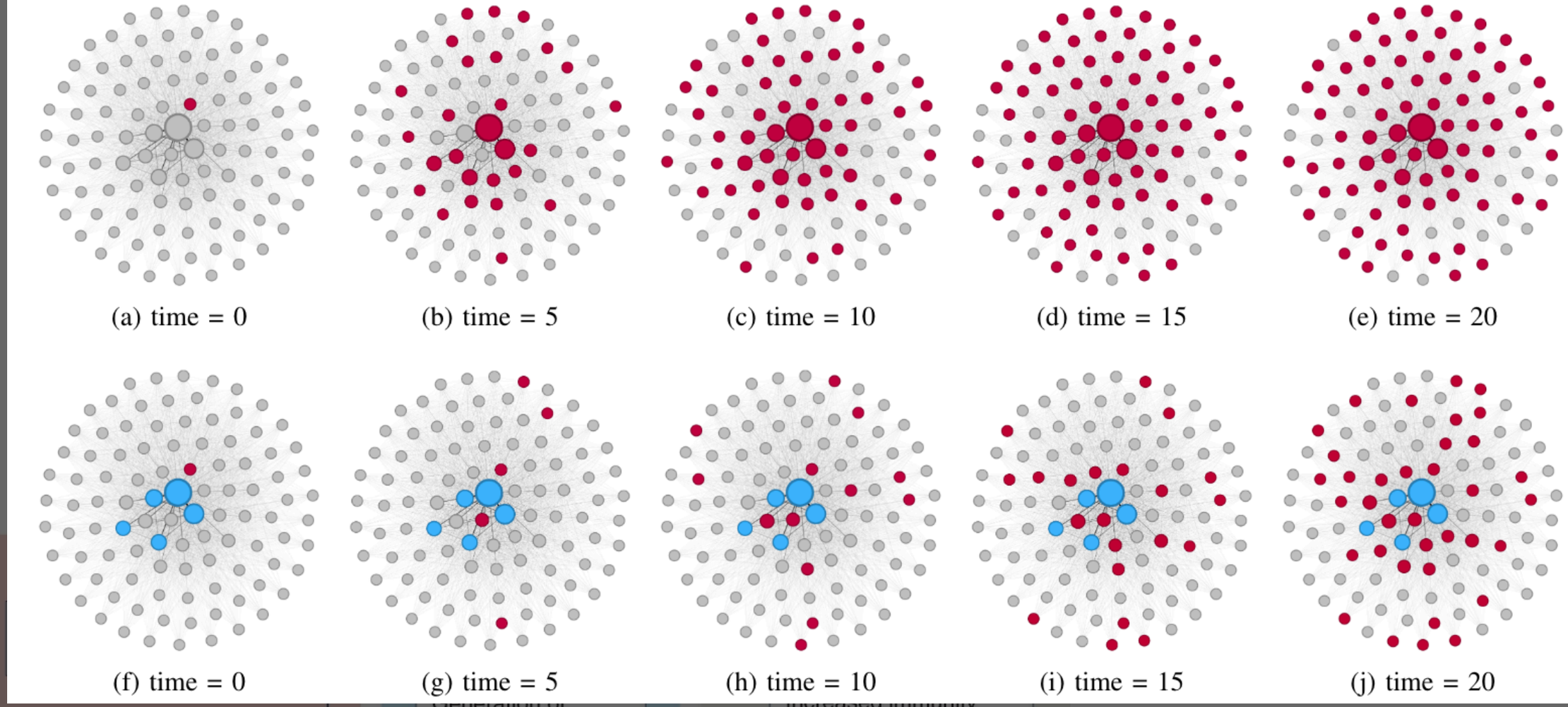


Preemptively inoculating users with correct information immunizes them against misinformation[1].

[1] van der Linden, S. *Misinformation: susceptibility, spread, and interventions to immunize the public.* Nature Medicine, 2022

CENTER FOR aUTonomy

Points of intervention



s with
s them

Refutational preemption, or prebunking

cognitive antibodies

to future misinformation

Post-inoculation talk (spreading the vaccine)

[1] van der Linden, S. *Misinformation: susceptibility, spread, and interventions to immunize the public.* Nature Medicine, 2022

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

When to Inoculate?

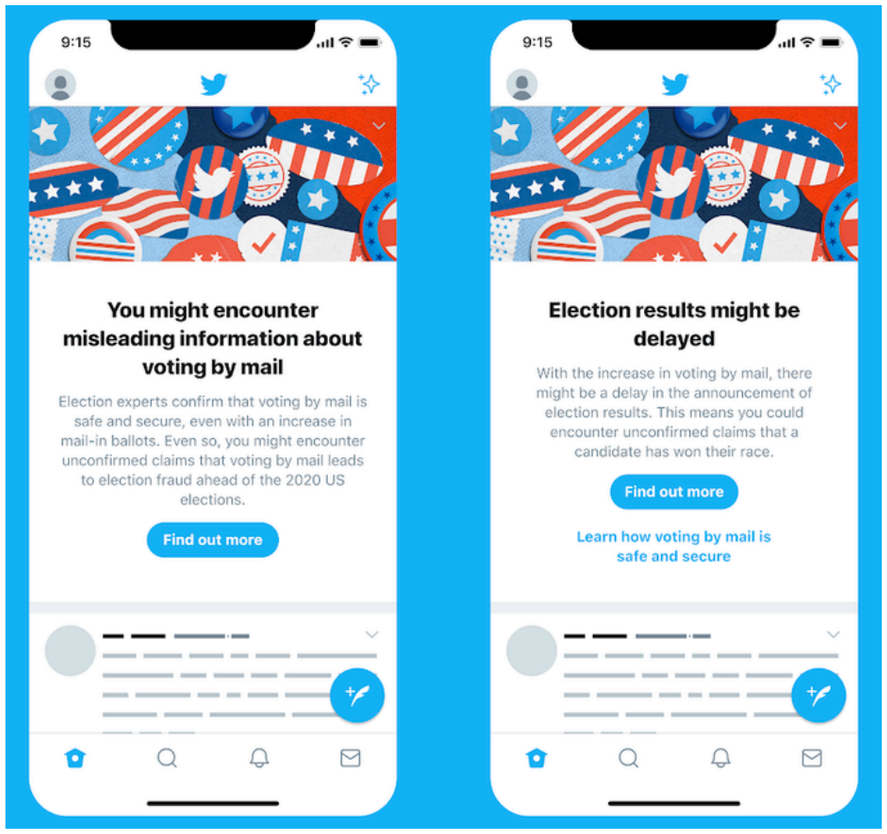
Overusing pre-bunking negatively affects user experience

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

When to Inoculate?



Overusing pre-bunking negatively affects user experience

Content Sequence: $Q = q_1, q_2, q_3, \dots$

Misinformation Misc. Content

$q_i \in \mathbf{M} \cup \mathbf{P} \cup \mathbf{C}$

Pre-bunk

Example:

$$Q = C_1, C_2, C_1, M_1, C_2, M_2, C_4, \dots$$

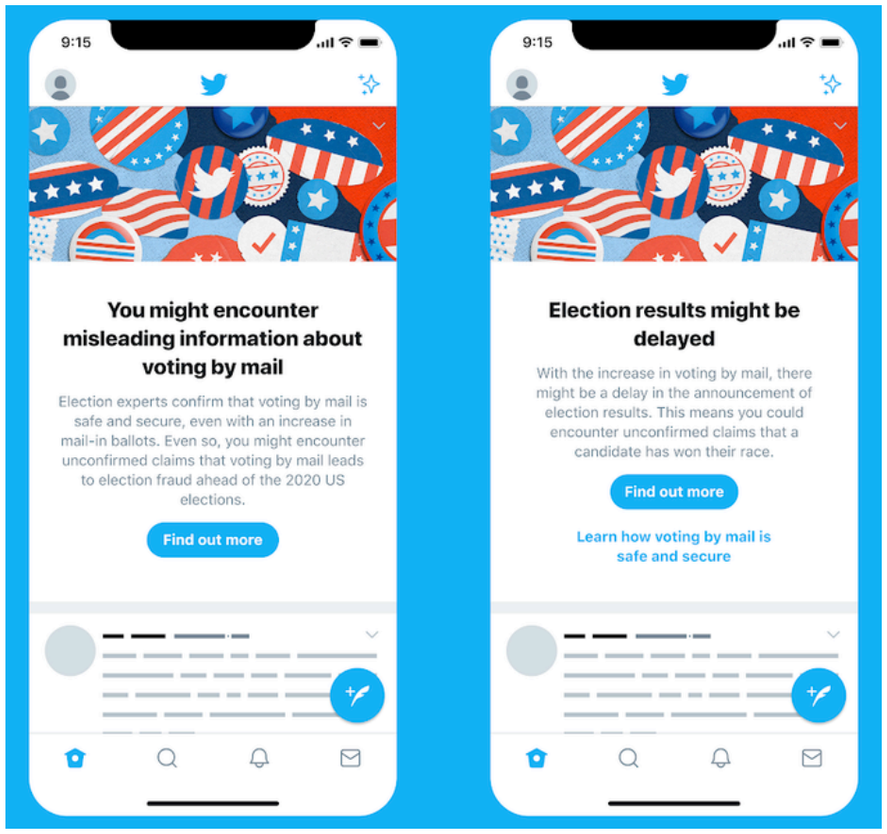
Multiple misinformation sources

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

When to Inoculate?



Overusing pre-bunking negatively affects user experience

Content Sequence: $Q = q_1, q_2, q_3, \dots$

Misinformation Misc. Content

$q_i \in \mathbf{M} \cup \mathbf{P} \cup \mathbf{C}$

Pre-bunk

Example:

$$Q = C_1, C_2, C_1, M_1, C_2, M_2, C_4, \dots$$

Multiple misinformation sources

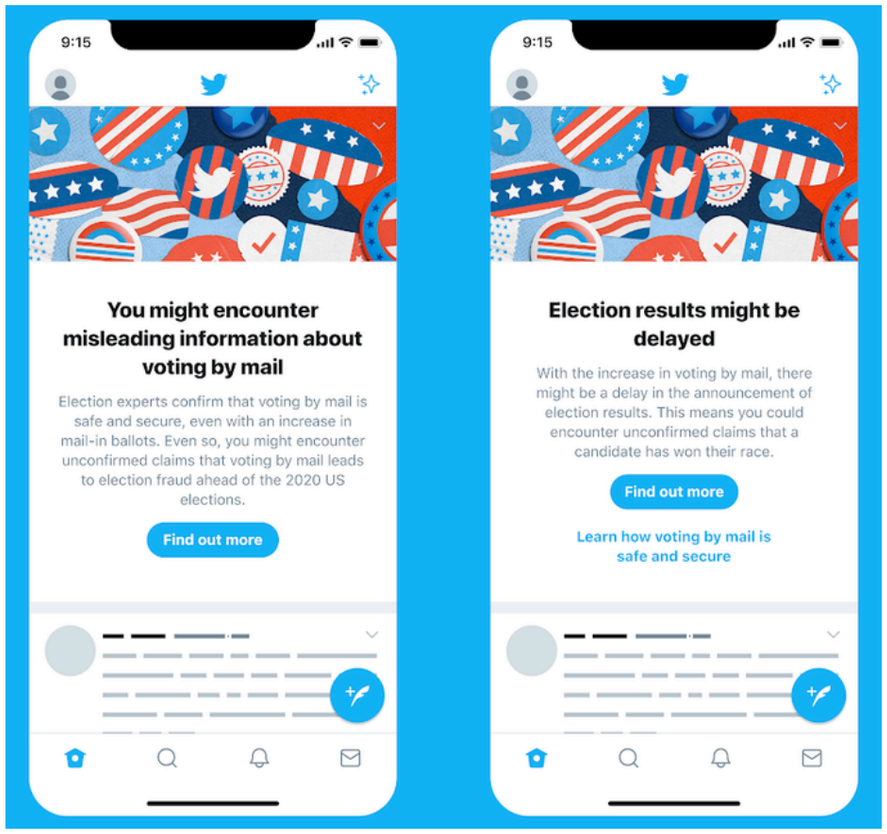
Goal 1: Ensure pre-bunks are delivered before misinformation. Eg: $C_1, P_1, C_2, C_1, M_1, C_2, P_2, M_2, C_4, \dots$

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

When to Inoculate?



Overusing pre-bunking negatively affects user experience

Content Sequence: $Q = q_1, q_2, q_3, \dots$

Misinformation Misc. Content

$q_i \in \mathbf{M} \cup \mathbf{P} \cup \mathbf{C}$

Pre-bunk

Example:

$$Q = C_1, C_2, C_1, M_1, C_2, M_2, C_4, \dots$$

Multiple misinformation sources

Goal 1: Ensure pre-bunks are delivered before misinformation.

Eg: $C_1, P_1, C_2, C_1, M_1, C_2, P_2, M_2, C_4, \dots$

Goal 2: Minimize pre-bunk concentration in feed.

$$\min_{a} \max_{i} c_{i+1} = \beta c_i + \mathbf{1}_{\{q_i \in \mathbf{P}\}}$$

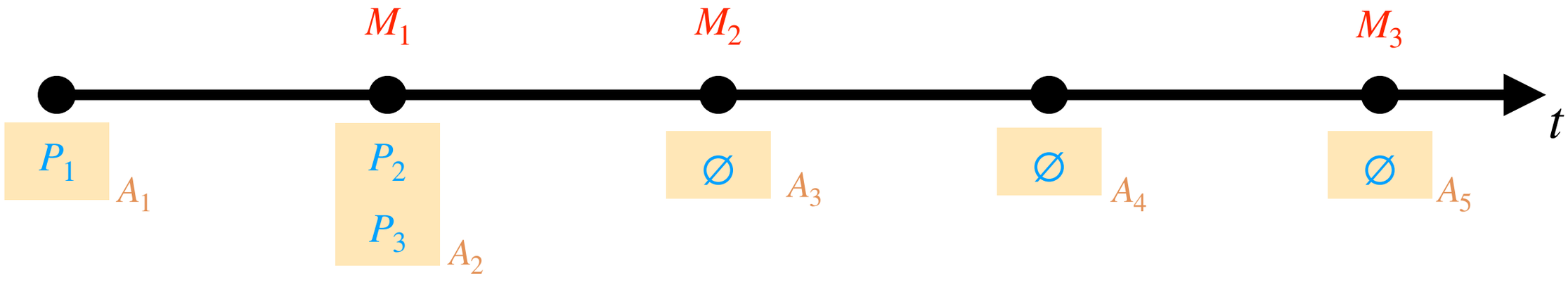
[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Each M_i propagates based on an **SI** epidemic model.

At each time we can decide on which pre-bunks to deliver based on policy $\pi(I_t^1, \ldots I_t^k)$.



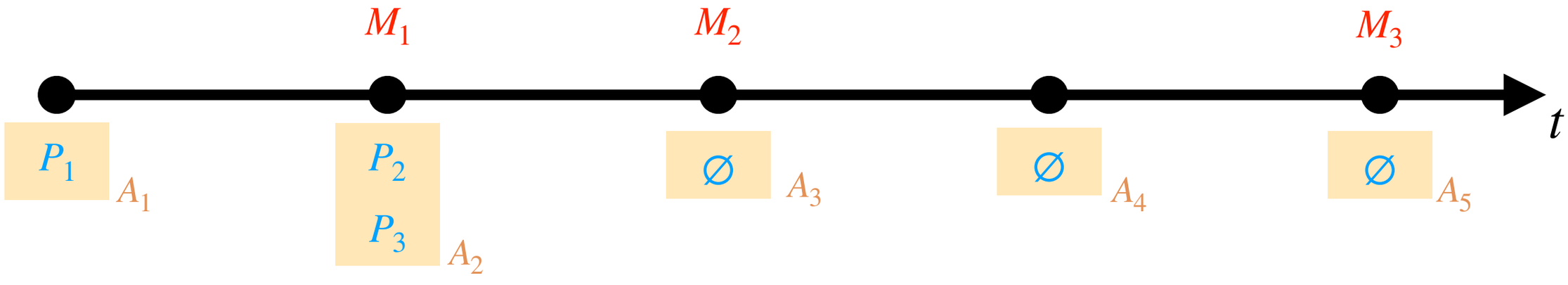
[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Each M_i propagates based on an **SI** epidemic model.

At each time we can decide on which pre-bunks to deliver based on policy $\pi(I_t^1, \ldots I_t^k)$.



$$\min_{\pi \in \Pi} \max_{t \in [T]} \quad c(\pi, t)$$

$$\text{s.t.} \quad A_1, \ldots, A_T \sim \pi\left(I_t^1 \ldots I_t^k\right),$$

$$\forall k \in \mathbb{N}, \quad \mathbb{P}\left(P_k \in \bigcup_{t=0}^{X_k - 1} A_t\right) = 1.$$

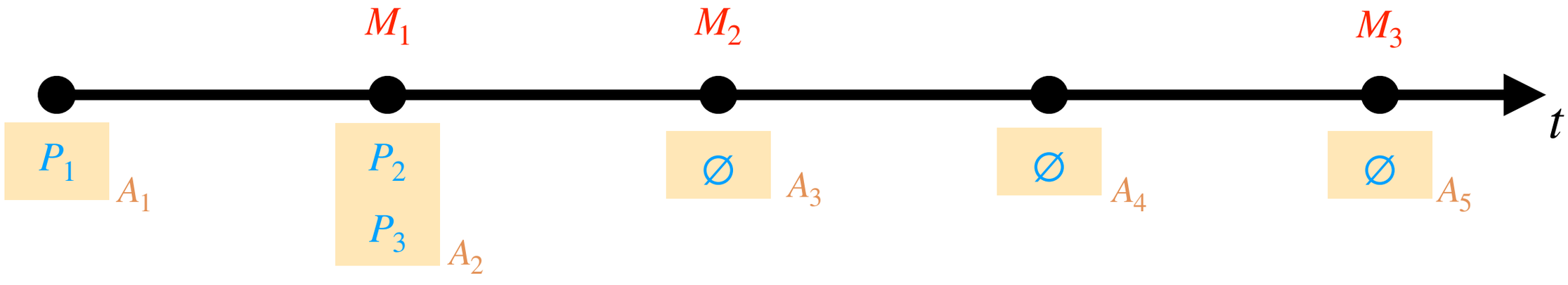
[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Countering Misinformation: Pre-bunking

CENTER FOR aUTonomy

Each M_i propagates based on an **SI** epidemic model.

At each time we can decide on which pre-bunks to deliver based on policy $\pi(I_t^1, \dots I_t^k)$.



$$\min_{\pi\in\Pi} \max_{t\in[T]} \quad c(\pi, t) \longrightarrow c(\pi, t) = \sum_{\tau=1}^{t} \beta^{t-\tau} \mathbb{E}_\pi \left[\left| A_\tau \right| \right]$$

$$\text{s.t.} \quad A_1, \dots, A_T \sim \pi\left(I_t^1 \dots I_t^k\right), \longrightarrow \pi \text{ determines the actions } A_t.$$

$$\forall k \in \mathbb{N}, \quad \mathbb{P}\left(P_k \in \bigcup_{t=0}^{X_k - 1} A_t \right) = 1. \longrightarrow \text{Deliver pre-bunks before misinformation}$$

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Two Baseline Policies

We want to ensure $\forall k \in \mathbb{N}, \quad \mathbb{P}\left(P_k \in \bigcup_{t=0}^{X_k - 1} A_t \right) = 1.$

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Two Baseline Policies

CENTER FOR aUTonomy

We want to ensure $\forall k \in \mathbb{N}, \quad \mathbb{P}\left(P_k \in \bigcup_{t=0}^{X_k-1} A_t\right) = 1.$

Policy 2: Deliver pre-bunk P_k whenever there are misinformed (I_t^k) nodes in the **neighborhood.**

$$A_t = \left\{ P_k \in P \mid I_t^k \cap N_{\text{in}}(c) \neq \varnothing, P_k \notin \bigcup_{\tau=1}^{t-1} A_\tau \right\}$$

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Two Baseline Policies

CENTER FOR aUTonomy

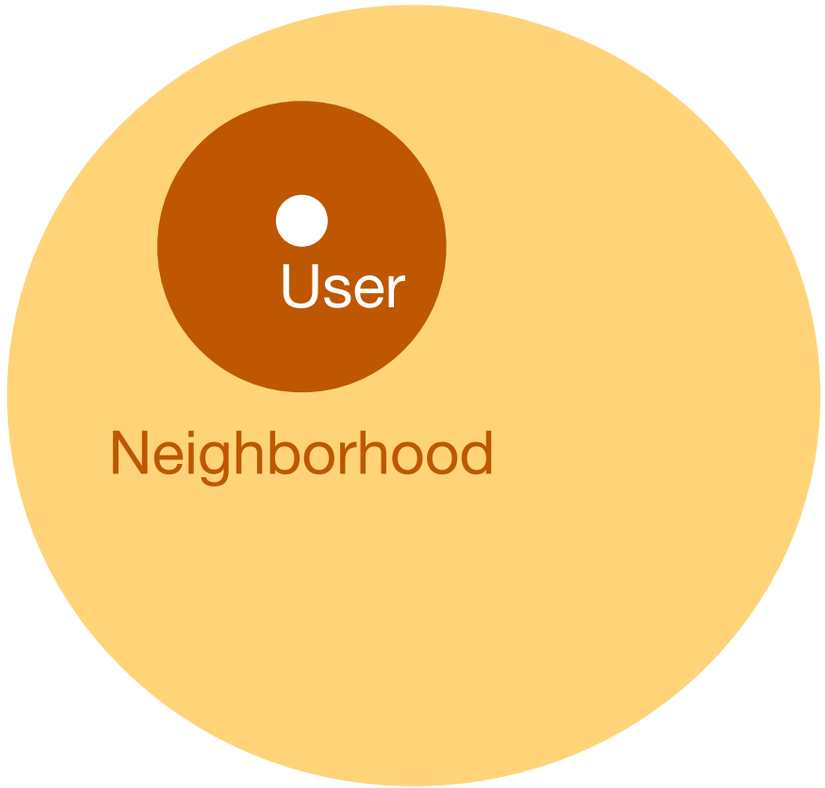
We want to ensure $\forall k \in \mathbb{N}, \quad \mathbb{P}\left(P_k \in \bigcup_{t=0}^{X_k-1} A_t\right) = 1.$

Policy 2: Deliver pre-bunk P_k whenever there are misinformed (I_t^k) nodes in the **neighborhood.**

$$A_t = \left\{ P_k \in P \mid I_t^k \cap N_{\text{in}}(c) \neq \varnothing, P_k \notin \bigcup_{\tau=1}^{t-1} A_\tau \right\}$$

Policy 1: Deliver pre-bunk P_k whenever $I_t^k \neq \varnothing$ in the **observed network.**

$$A_t = \left\{ P_k \in P \mid I_t^k \neq \varnothing, P_k \notin \bigcup_{\tau=1}^{t-1} A_\tau \right\}$$



[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Two Baseline Policies

CENTER FOR aUTonomy

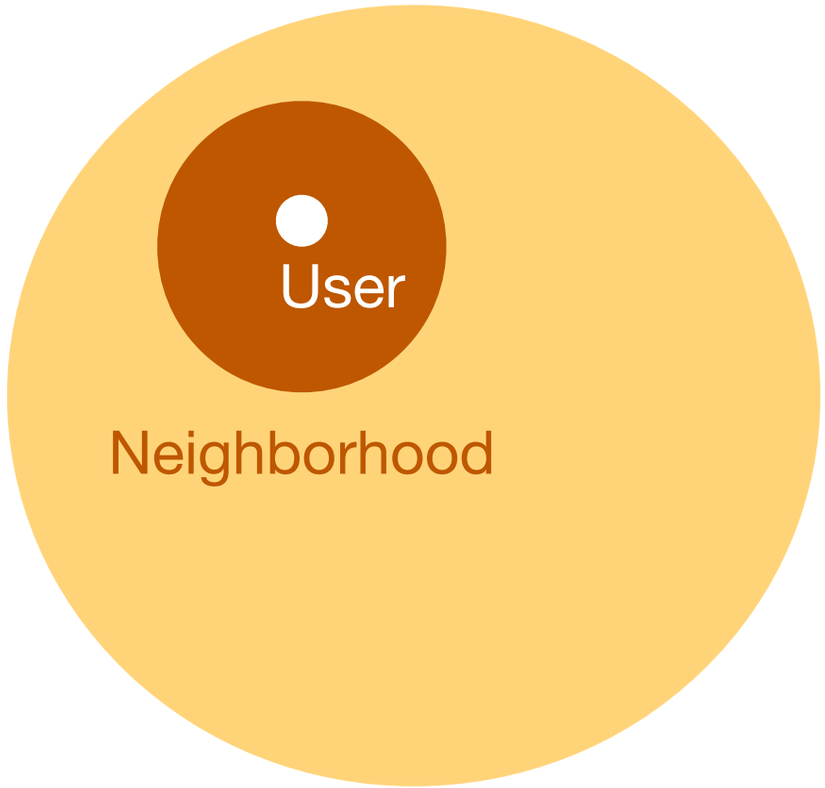
We want to ensure $\forall k \in \mathbb{N}, \quad \mathbb{P}\left(P_k \in \bigcup_{t=0}^{X_k-1} A_t\right) = 1.$

Policy 2: Deliver pre-bunk P_k whenever there are misinformed (I_t^k) nodes in the **neighborhood.**

$$A_t = \left\{ P_k \in P \mid I_t^k \cap N_{\text{in}}(c) \neq \emptyset, P_k \notin \bigcup_{\tau=1}^{t-1} A_\tau \right\}$$

Policy 1: Deliver pre-bunk P_k whenever $I_t^k \neq \emptyset$ in the **observed network.**

$$A_t = \left\{ P_k \in P \mid I_t^k \neq \emptyset, P_k \notin \bigcup_{\tau=1}^{t-1} A_\tau \right\}$$



Both policies yield feasible but suboptimal solutions.

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Equal Pre-bunk Spacing

CENTER FOR aUTonomy

Proxy Problem

Optimally deliver $\lfloor \epsilon T \rfloor$ pre-bunks until time T.

$$\min_{t_1 \dots t_{\lfloor \epsilon T \rfloor}} \max_{n \in \{0 \dots \lfloor \epsilon T \rfloor\}} \sum_{i=1}^{n} \beta^{t_n - t_j},$$

$$\text{s.t.} \quad 0 \leq t_{i-1} \leq t_i \leq T, \quad \forall i \in \{1 \dots \lfloor \epsilon T \rfloor\}.$$

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Equal Pre-bunk Spacing

CENTER FOR aUTonomy

Proxy Problem

Optimally deliver $\lfloor \epsilon T \rfloor$ pre-bunks until time T.

$$\min_{t_1 \dots t_{\lfloor \in T \rfloor}} \max_{n \in \{0 \dots \lfloor \epsilon T \rfloor\}} \sum_{i=1}^{n} \beta^{t_n - t_j},$$
$$\text{s.t.} \quad 0 \le t_{i-1} \le t_i \le T, \quad \forall i \in \{1 \dots \lfloor \epsilon T \rfloor\}.$$

Theorem: Let $t_0 \dots t_{\lfloor \in T \rfloor} \in \mathbb{R}$ be an optimal solution to above problem , then there exists h such that $t_i - t_{i-1}$ is constant across all $i = h+1, \dots, \lfloor \epsilon T \rfloor$. Specifically there exists a constant α such that[1],

$$t_h - t_{h-1} = \log_\beta(\alpha - 1) - \log_\beta(h),$$
$$t_i - t_{i-1} = \log_\beta(\alpha - 1) - \log_\beta(\alpha).$$

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Equal Pre-bunk Spacing

CENTER FOR aUTonomy

Proxy Problem

Optimally deliver $\lfloor \epsilon T \rfloor$ pre-bunks until time T.

$$\min_{t_1 \dots t_{\lfloor \in T \rfloor}} \max_{n \in \{0 \dots \lfloor \epsilon T \rfloor\}} \sum_{i=1}^{n} \beta^{t_n - t_j},$$
$$\text{s.t.} \quad 0 \le t_{i-1} \le t_i \le T, \quad \forall i \in \{1 \dots \lfloor \epsilon T \rfloor\}.$$

Theorem: Let $t_0 \dots t_{\lfloor \in T \rfloor} \in \mathbb{R}$ be an optimal solution to above problem , then there exists h such that $t_i - t_{i-1}$ is constant across all $i = h+1, \dots, \lfloor \epsilon T \rfloor$. Specifically there exists a constant α such that[1],

$$t_h - t_{h-1} = \log_\beta(\alpha - 1) - \log_\beta(h),$$
$$t_i - t_{i-1} = \log_\beta(\alpha - 1) - \log_\beta(\alpha).$$

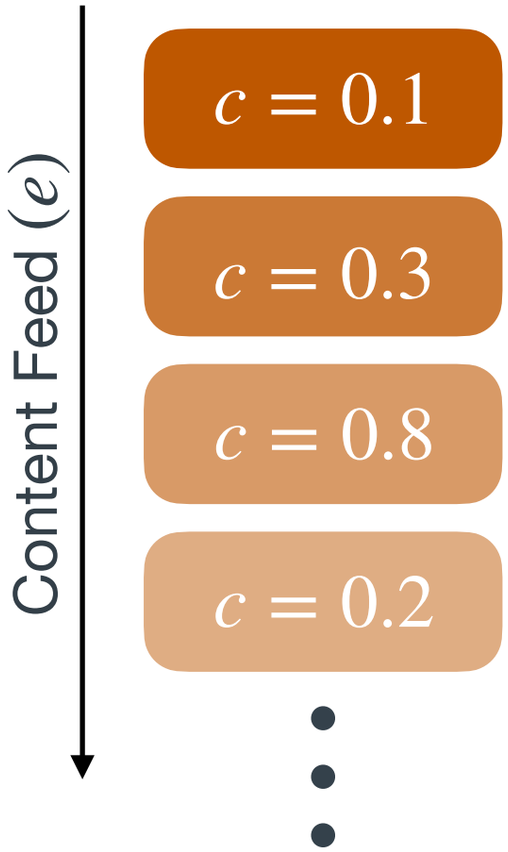
Optimal solution of the proxy problem evenly distributes pre-bunks.

[1] Y. E. Bayiz, U. Topcu, *Prebunking Design as a Defense Mechanism Against Misinformation Propagation on Social Networks,* Submitted to CDC 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$

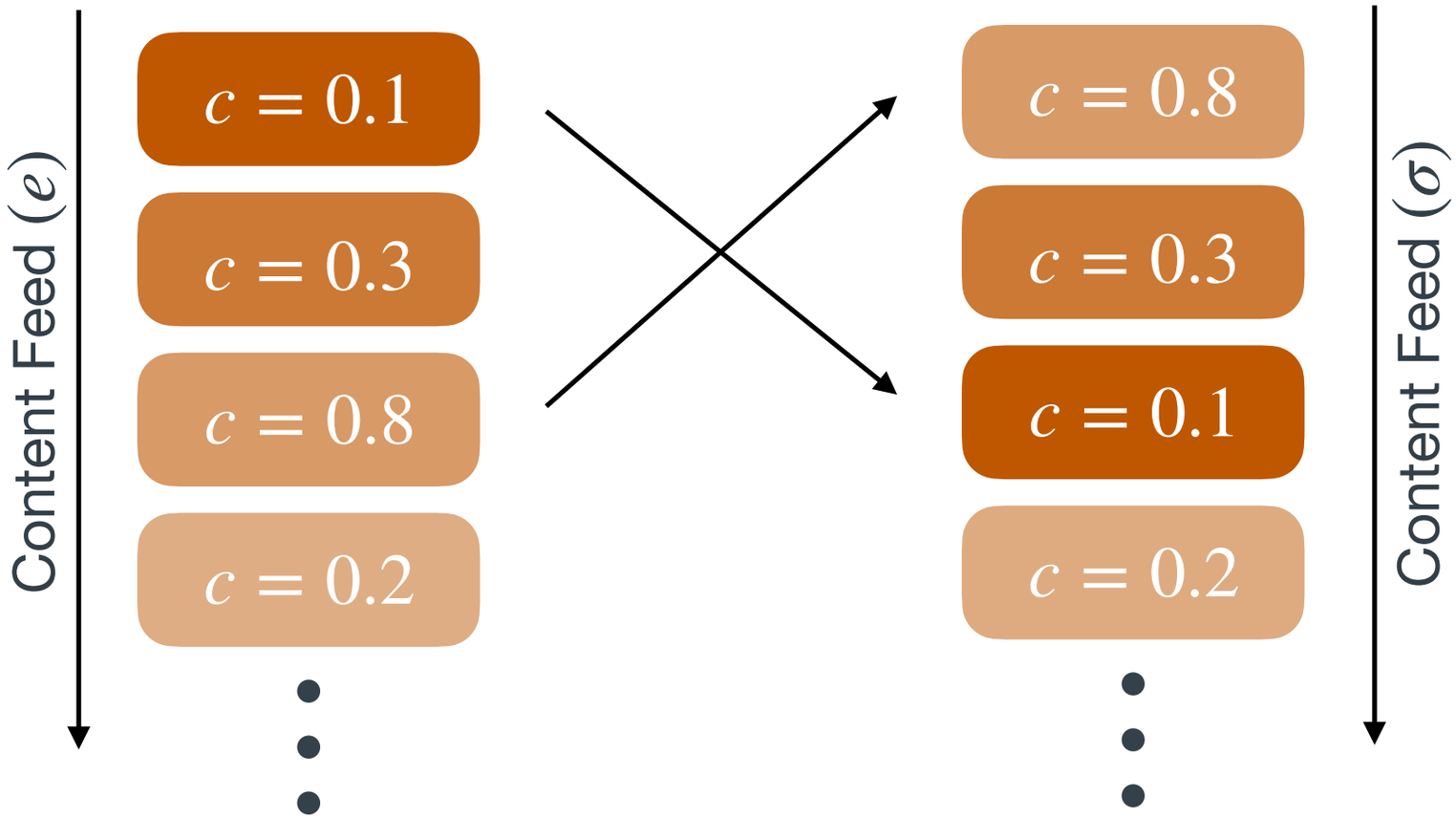


[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$

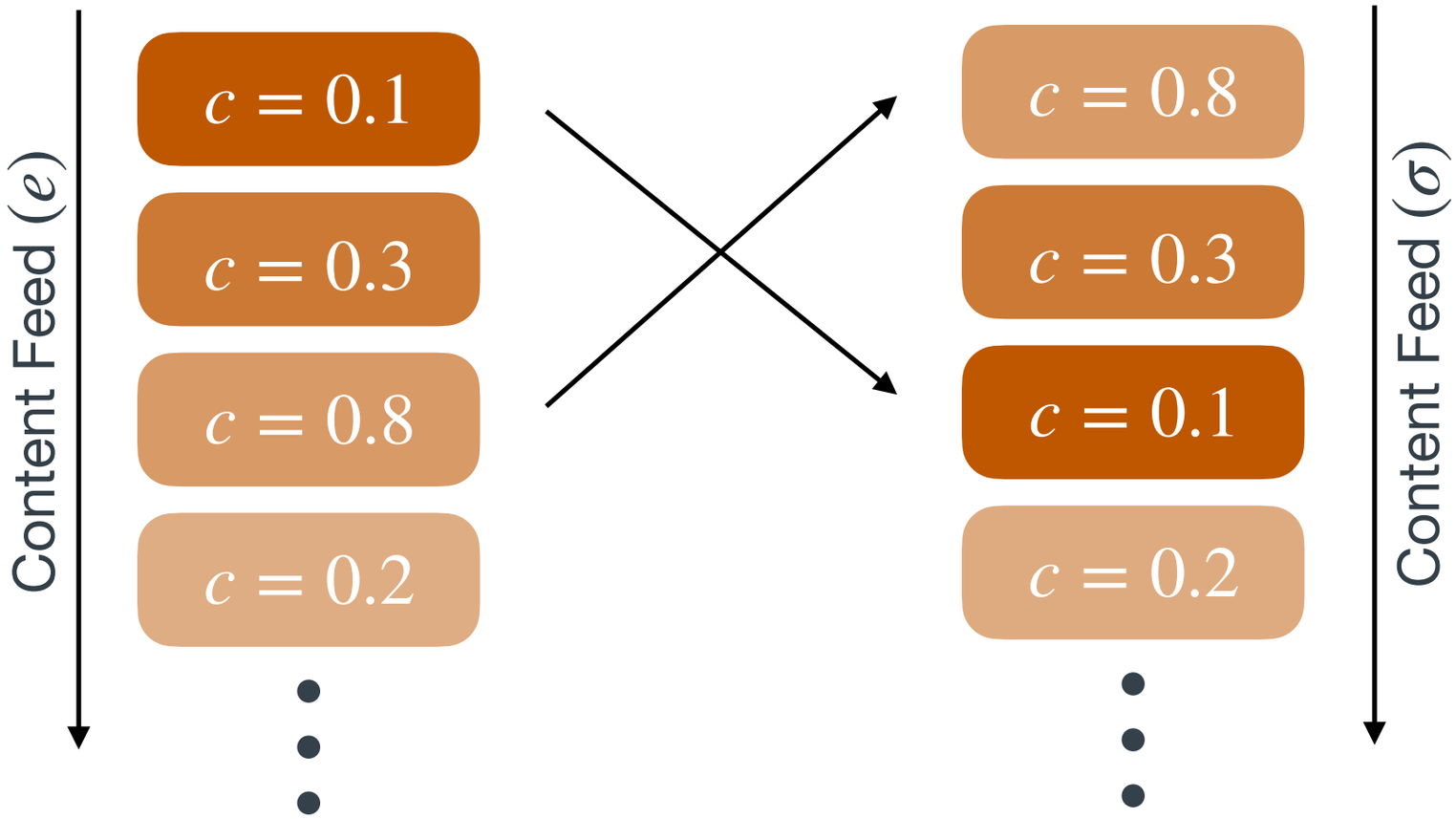


[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$



Assume that user sees p^{th} post with probability

$$q_p = \lambda^p \longrightarrow \textbf{persistence}$$

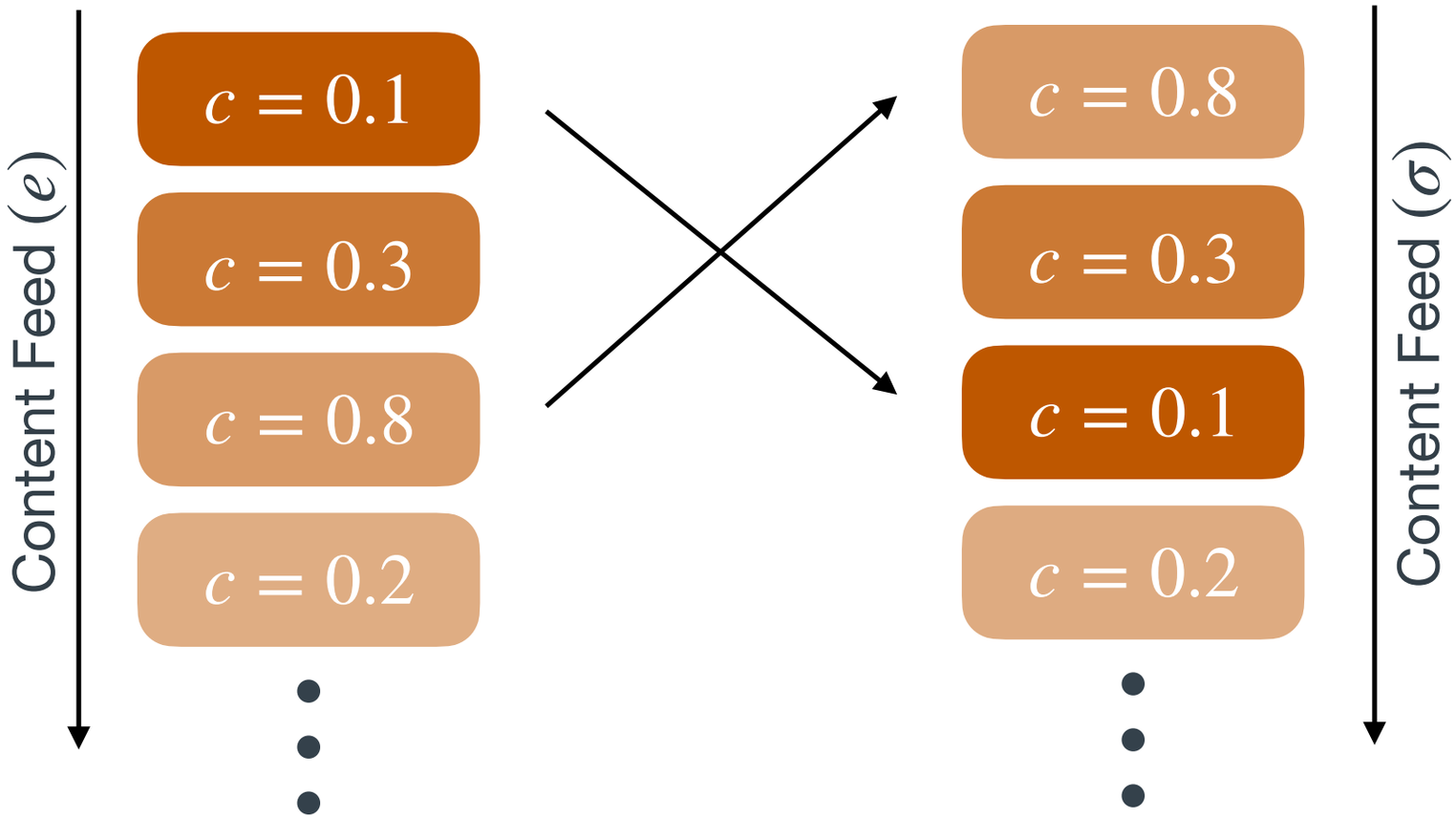
[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$

Identity permutation optimizes platform objectives



Assume that user sees p^{th} post with probability

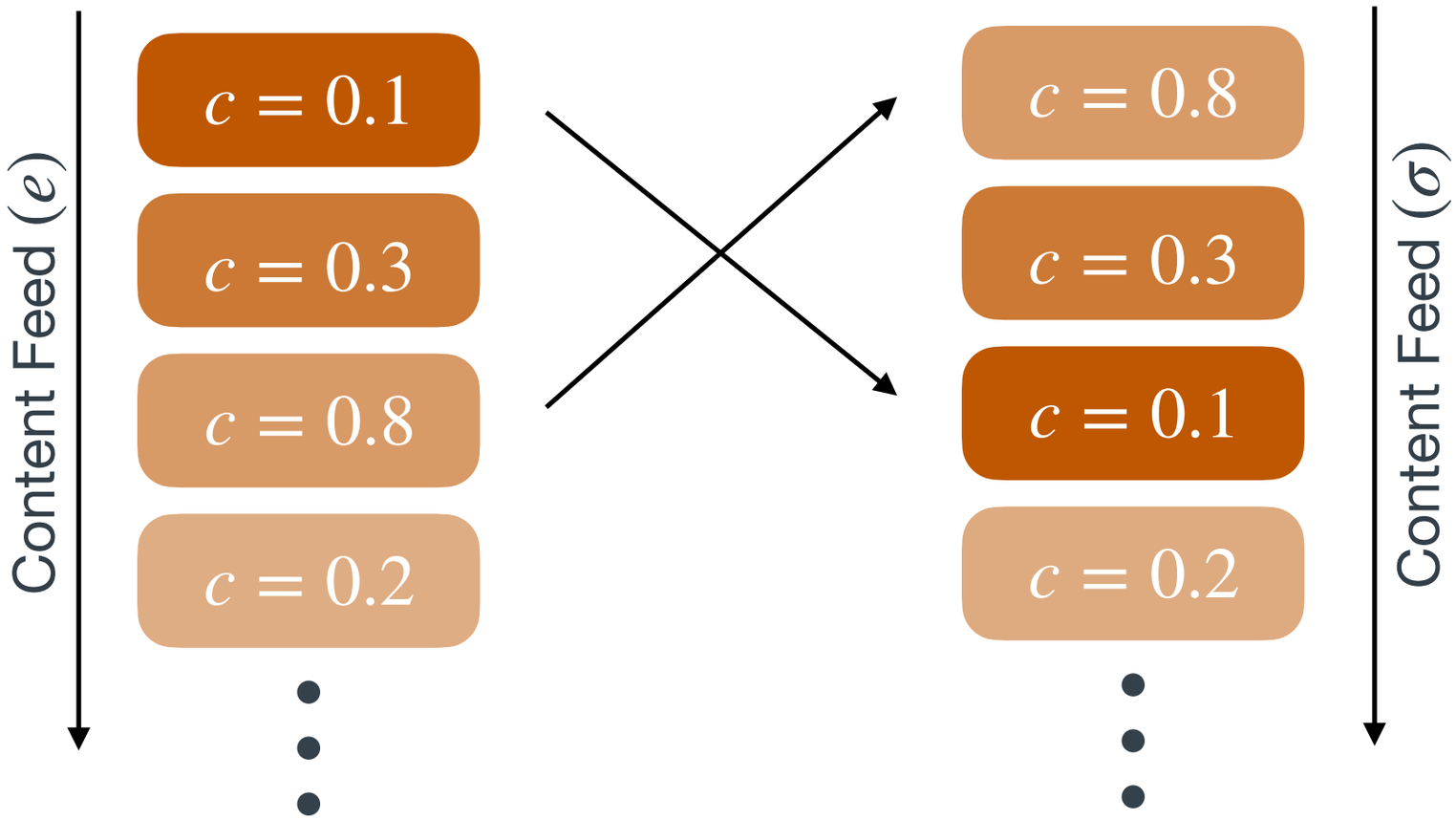
$$q_p = \lambda^p \longrightarrow \textbf{persistence}$$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$



Identity permutation optimizes platform objectives

$$e = \arg\max_{\sigma} V(\sigma)$$

$$V(\sigma) = \sum_{p=1}^{K} q_p v(\sigma(p)),$$

Assume that user sees p^{th} post with probability

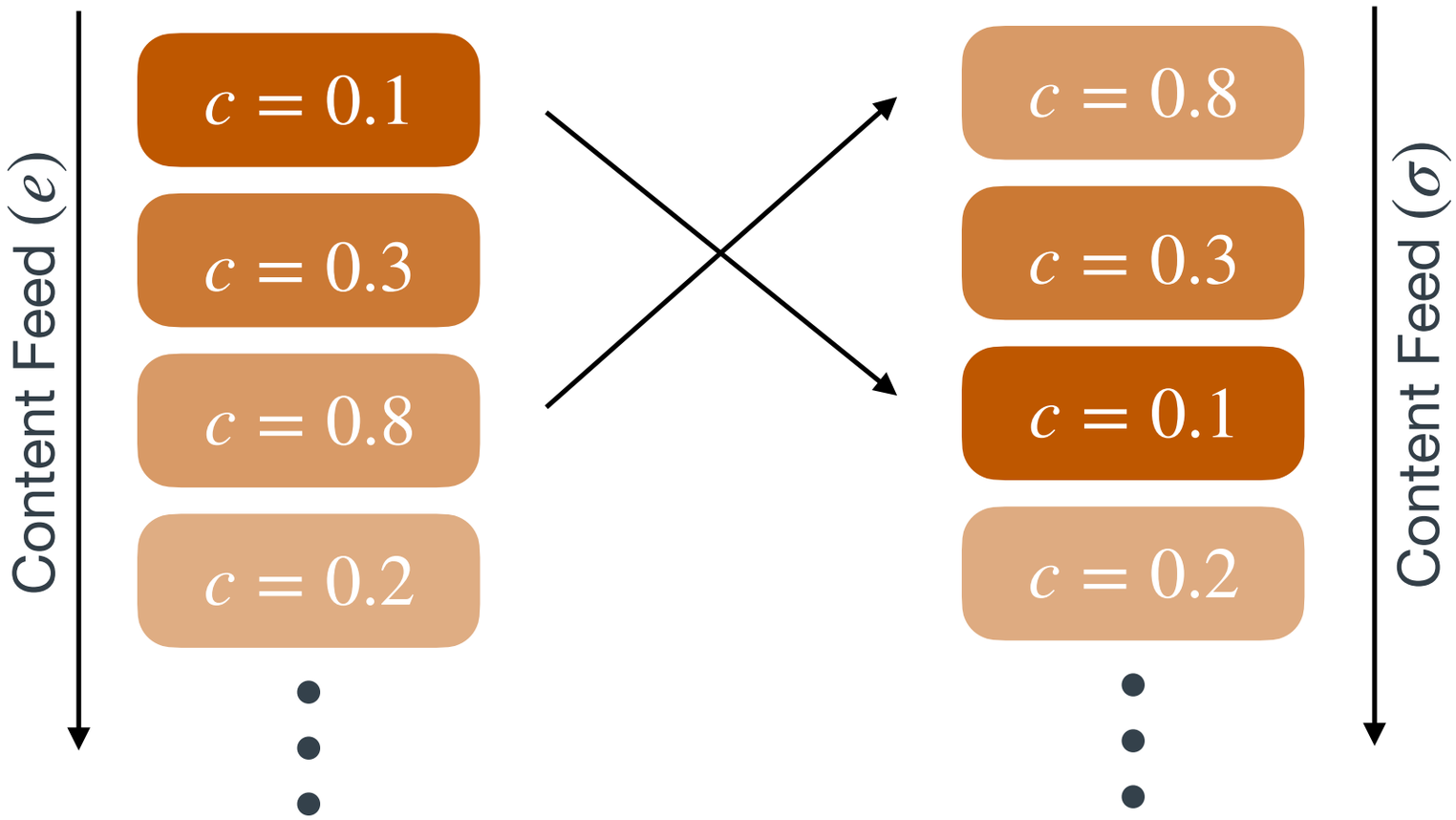
$$q_p = \lambda^p \longrightarrow \textbf{persistence}$$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$



Identity permutation optimizes platform objectives

$$e = \arg\max_{\sigma} V(\sigma)$$

$$V(\sigma) = \sum_{p=1}^{K} q_p v(\sigma(p)),$$

$$\max_{\sigma \in S_K} \quad \alpha V(\sigma) + \sum_{p=1}^{K} q_p c_{\sigma(p)}$$

Assume that user sees p^{th} post with probability

$$q_p = \lambda^p \longrightarrow \textbf{persistence}$$

Problem: $V(\sigma)$ is not known

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Misinformation is vaguely defined.

- Fake News
- Propaganda
- Scams
- Rumors

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Misinformation is vaguely defined.

- Fake News
- Propaganda
- Scams
- Rumors

Instead of labeling **misinformation** we use a probabilistic **credibility** measure

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Misinformation is vaguely defined.

- Fake News
- Propaganda
- Scams
- Rumors

Instead of labeling **misinformation** we use a probabilistic **credibility** measure

Democratic View

Suppose a randomly selected user labels post p as $Z_p \sim \text{Bernoulli}(c)$

Define credibility as $E[Z_p] = c$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Misinformation is vaguely defined.

- Fake News
- Propaganda
- Scams
- Rumors

Instead of labeling **misinformation** we use a probabilistic **credibility** measure

Democratic View

Suppose a randomly selected user labels post p as $Z_p \sim \text{Bernoulli}(c)$

Define credibility as $E[Z_p] = c$

Evidence-based View

Parse top k news articles

Determine credibility based on the consensus

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Misinformation is vaguely defined.

- Fake News
- Propaganda
- Scams
- Rumors

Instead of labeling **misinformation** we use a probabilistic **credibility** measure

Democratic View

Suppose a randomly selected user labels post p as $Z_p \sim \text{Bernoulli}(c)$

Define credibility as $E[Z_p] = c$

Evidence-based View

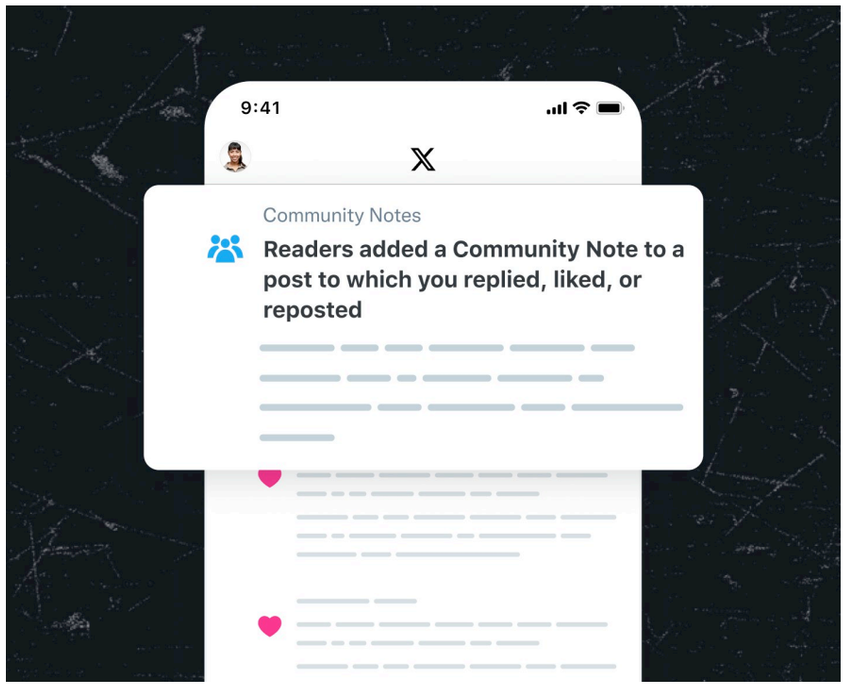
Parse top k news articles

Determine credibility based on the consensus

We use a hybrid approach that utilizes both views.

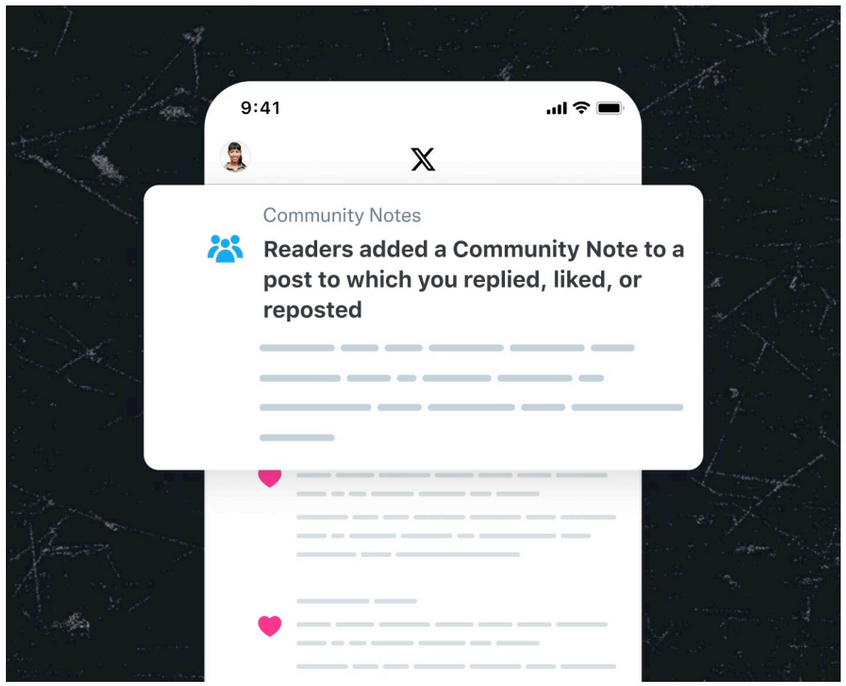
[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Community Notes



[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Community Notes

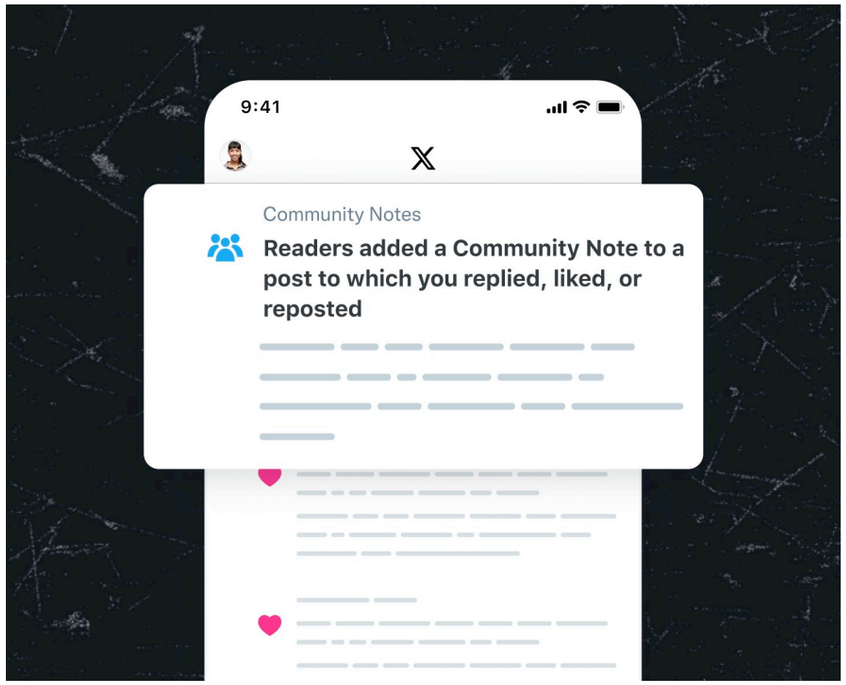


$$\hat{c}_p^{\text{human}} = \frac{\sum_{i=1}^{n} \left(z_{p,i} \cdot h_{p,i}^{+} + (1 - z_{p,i}) \cdot h_{p,i}^{-} \right)}{\sum_{i=1}^{n} \left(h_{p,i}^{+} + h_{p,i}^{-} \right)},$$

- Each post p comes with community notes
- Each community note i comes with a verdict $z_{p,i}$
- Eat community note also comes with ratings:
  - $h_{p,i}^{+}$: Number of ‘helpful’ ratings
  - $h_{p,i}^{-}$: Number of ‘not helpful’ ratings

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Community Notes



Push towards note verdict if helpful

Push away from note verdict if not helpful

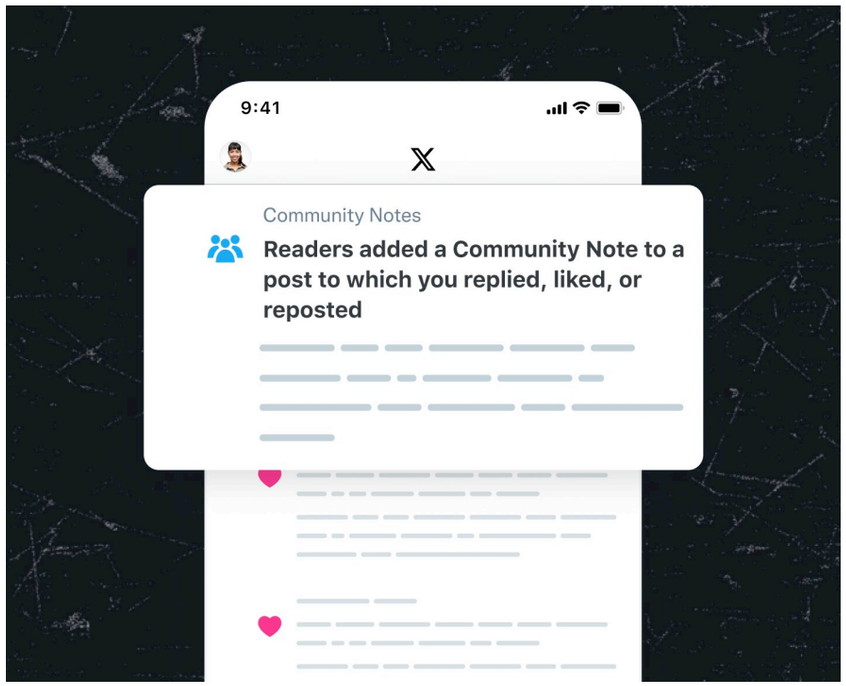
$$\hat{c}_p^{\text{human}} = \frac{\sum_{i=1}^{n} \left(z_{p,i} \cdot h_{p,i}^{+} \right) + \left((1 - z_{p,i}) \cdot h_{p,i}^{-} \right)}{\sum_{i=1}^{n} \left(h_{p,i}^{+} + h_{p,i}^{-} \right)},$$

Total number of ratings across all posts

- Each post p comes with community notes
- Each community note i comes with a verdict $z_{p,i}$
- Eat community note also comes with ratings:
  - $h_{p,i}^{+}$: Number of ‘helpful’ ratings
  - $h_{p,i}^{-}$: Number of ‘not helpful’ ratings

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Community Notes



- Each post p comes with community notes
- Each community note i comes with a verdict $z_{p,i}$
- Eat community note also comes with ratings:
  - $h^{+}_{p,i}$: Number of ‘helpful’ ratings
  - $h^{-}_{p,i}$: Number of ‘not helpful’ ratings

Push towards note verdict if helpful

Push away from note verdict if not helpful

$$\hat{c}_p^{\text{human}} = \frac{\sum_{i=1}^{n} \left(z_{p,i} \cdot h^{+}_{p,i} \right) + \left((1 - z_{p,i}) \cdot h^{-}_{p,i} \right)}{\sum_{i=1}^{n} \left(h^{+}_{p,i} + h^{-}_{p,i} \right)},$$

Total number of ratings across all posts

Assuming ratings and notes come from uniformly sampled users

$\hat{c}_p^{\text{human}}$ is an unbiased estimator of $c = E[Z_p]$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Enhancing Human Scores

CENTER FOR aUTonomy

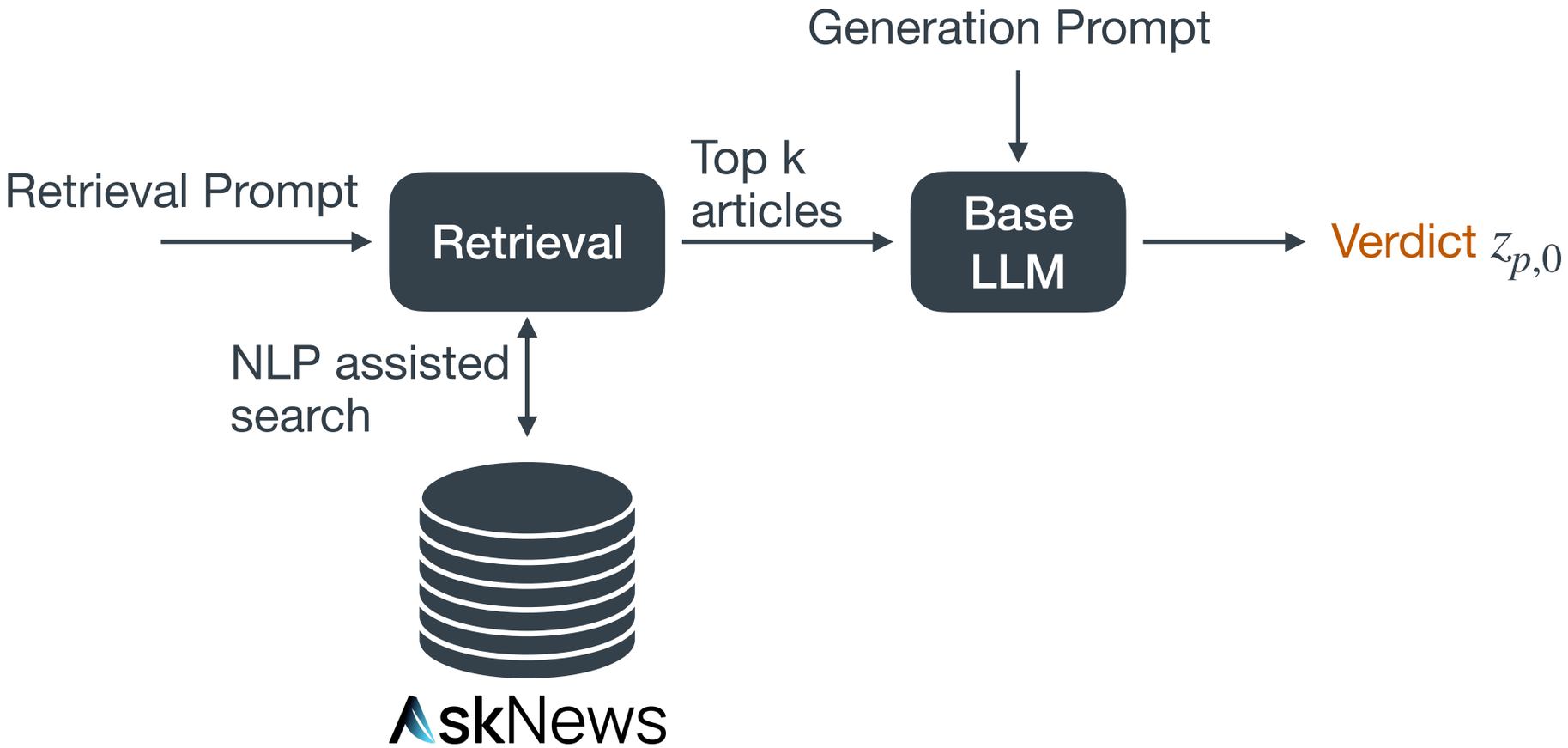
Generate artificial community notes using retrieval augmented generation (RAG)

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Enhancing Human Scores

CENTER FOR aUTonomy

Generate artificial community notes using retrieval augmented generation (RAG)

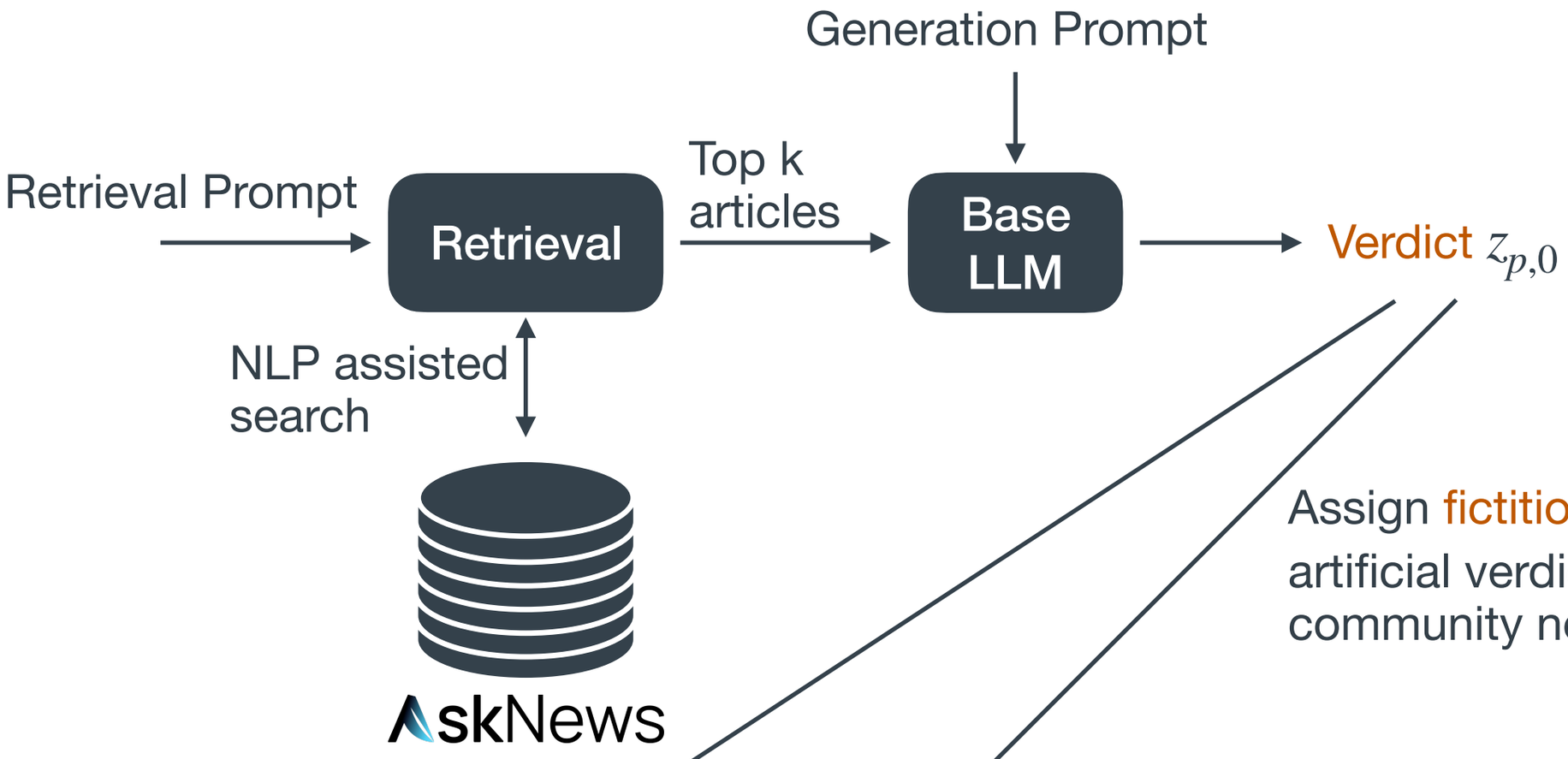


[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Enhancing Human Scores

CENTER FOR aUTonomy

Generate artificial community notes using retrieval augmented generation (RAG)



$$\hat{c}_p = \frac{z_{p,0} \cdot k^+ + (1 - z_{p,0}) \cdot k^- + \sum_{i=1}^{n} \left(z_{p,i} \cdot h_{p,i}^+ + (1 - z_{p,i}) \cdot h_{p,i}^- \right)}{\sum_{i=1}^{n} \left(h_{p,i}^+ + h_{p,i}^- \right) + k^+ + k^-}$$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Enhancing Human Scores

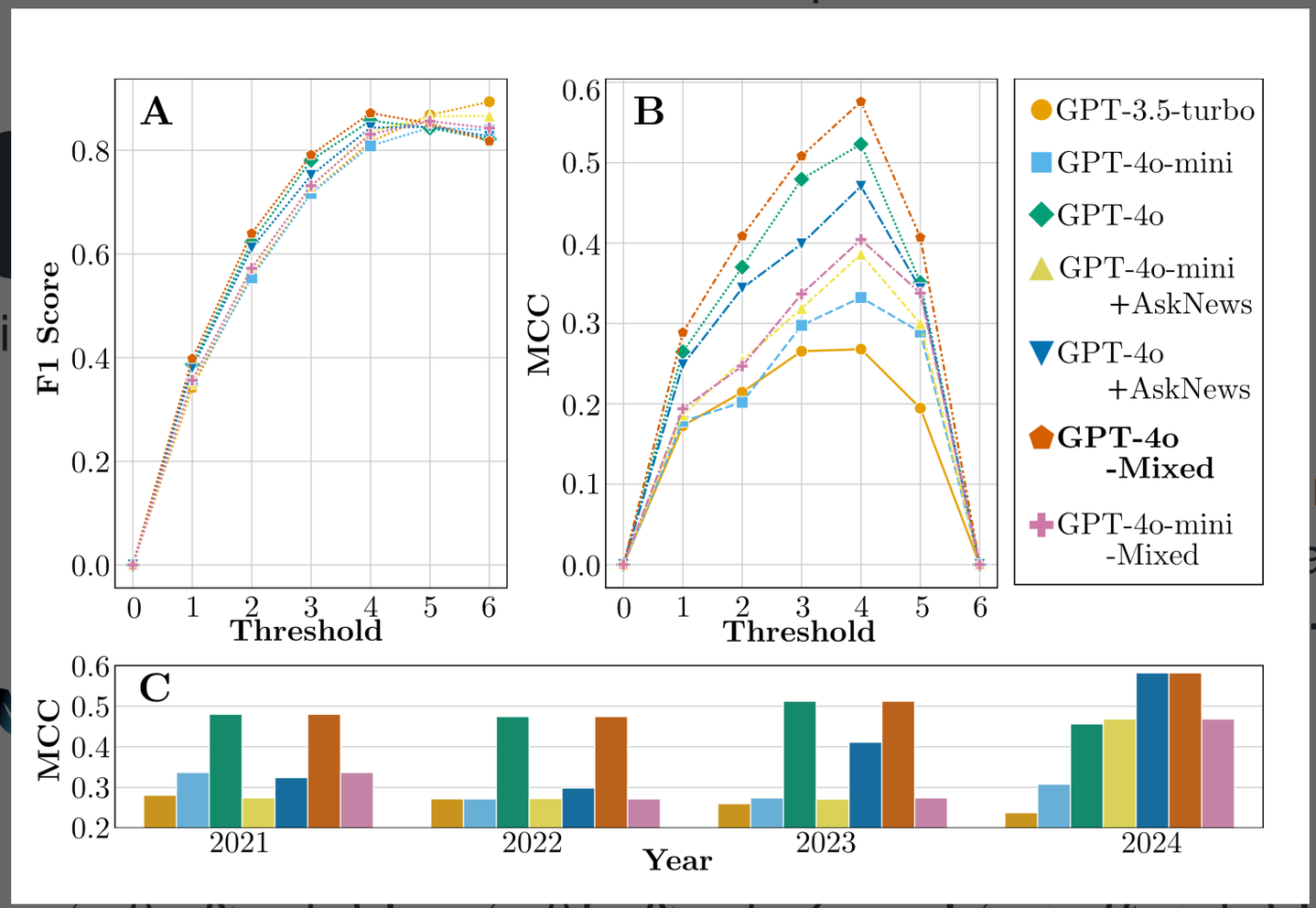
CENTER FOR aUTonomy

Generate artificial community notes using retrieval augmented generation (RAG)

Retrieval Prompt

NLP assi
search

Generation Prompt



ratings k_p^+, k_p^- to
and treat it like a

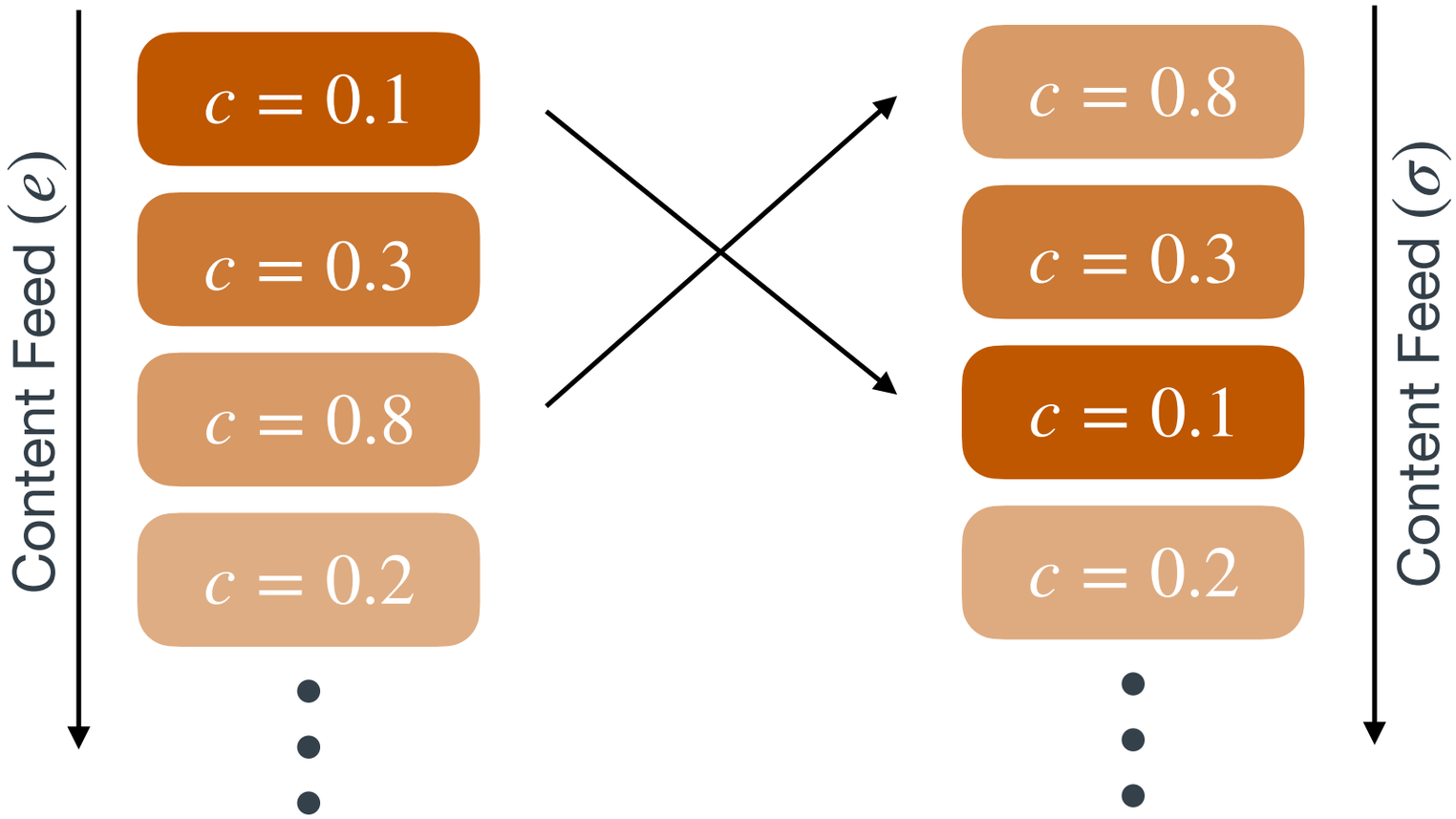
$$\hat{c}_p = \frac{z_{p,0} \cdot k^+ + (1 - z_{p,0}) \cdot k^- + \sum_{i=1} \left(z_{p,i} \cdot h_{p,i}^+ + (1 - z_{p,i}) \cdot h_{p,i}^- \right)}{\sum_{i=1}^{n} \left(h_{p,i}^+ + h_{p,i}^- \right) + k^+ + k^-}$$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$



$$\max_{\sigma \in S_K} \quad \alpha V(\sigma) + \sum_{p=1}^{K} q_p c_{\sigma(p)}$$

Problem: $V(\sigma)$ is not known

($V(\sigma)$ may not even exist)

Assume identity permutation optimizes platform objectives

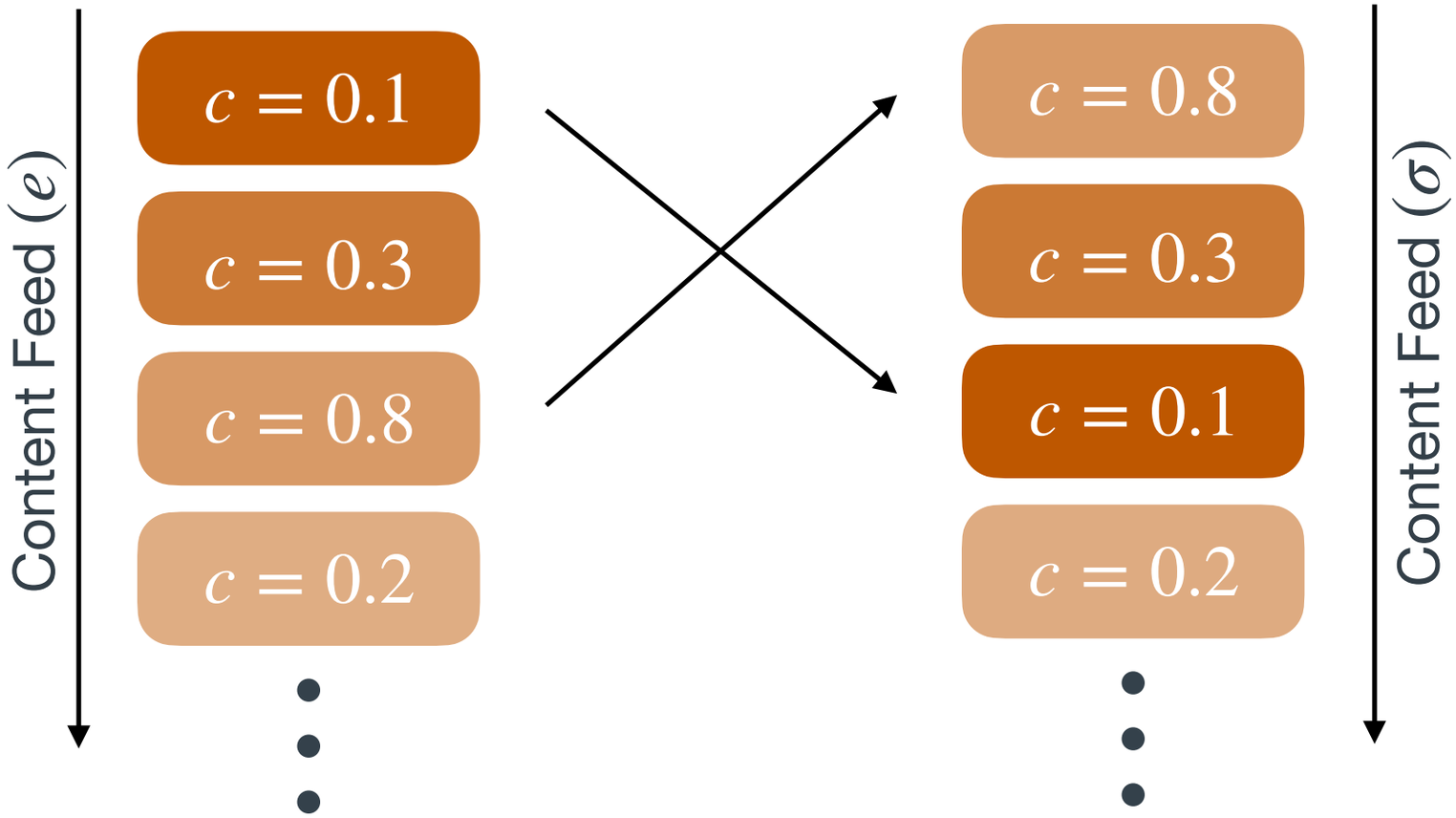
$$e = \arg\max_{\sigma} V(\sigma) \qquad V(\sigma) = \sum_{p=1}^{K} q_p v(\sigma(p)),$$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Re-ranking for Credibility

CENTER FOR aUTonomy

Assumption: Content arrives already sorted in a **content feed** based on original value assignment $V(\sigma)$



Assume identity permutation optimizes platform objectives

$$e = \arg\max_{\sigma} V(\sigma) \qquad V(\sigma) = \sum_{p=1}^{K} q_p v(\sigma(p)),$$

$$\max_{\sigma \in S_K} \quad \alpha V(\sigma) + \sum_{p=1}^{K} q_p c_{\sigma(p)}$$

Problem: $V(\sigma)$ is not known

($V(\sigma)$ may not even exist)

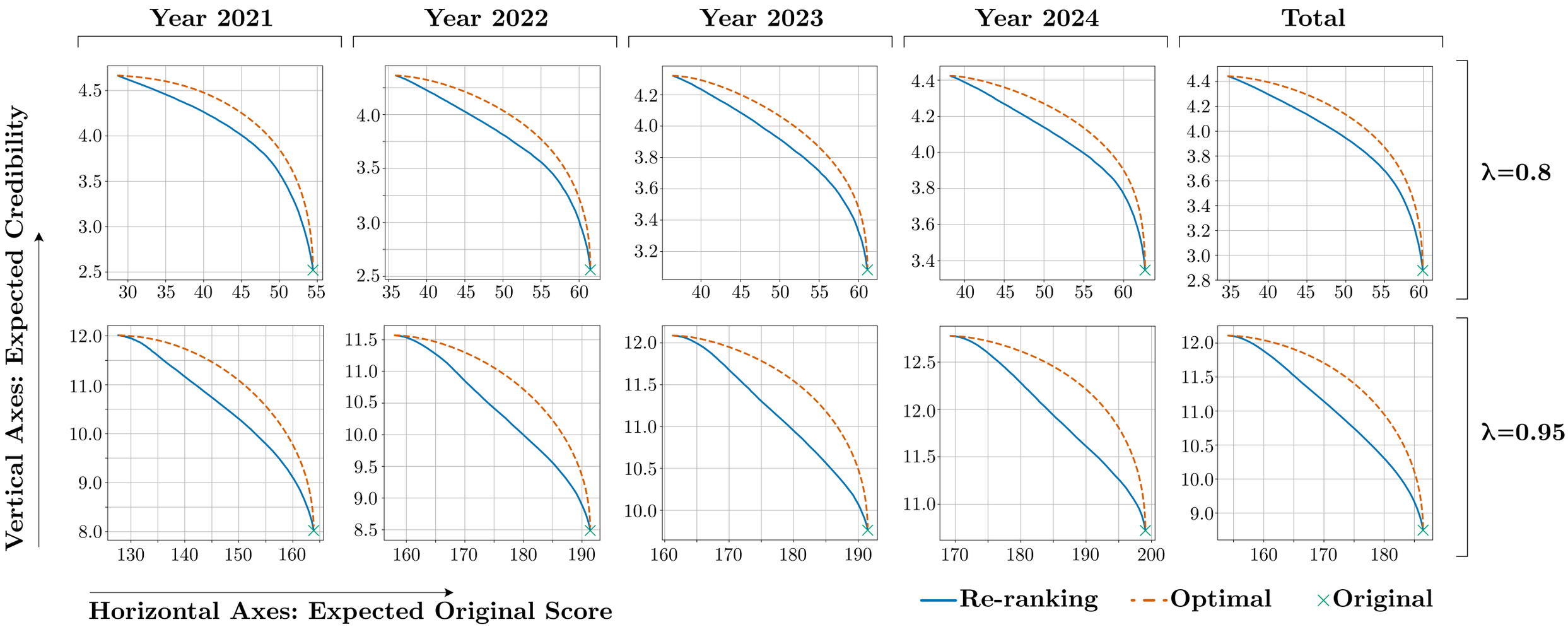
Solution: Solve a surrogate problem

$$\min_{\sigma} D(\sigma) - \sum_{p=1}^{K} q_p c_{\sigma(p)}$$

$$D(\sigma) = \sum_{i=1}^{K} |\sigma(i) - i|$$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Optimality Gap



What is the cost of remaining oblivious to original cost?

Theoretically "Optimal" solutions: $\max_{\sigma \in S_K} \quad \alpha V(\sigma) + \sum_{p=1}^{K} q_p c_{\sigma(p)}$

[1] Y. E. Bayiz, U. Topcu, *Re-ranking with Semi-automated Misinformation Detection for Increasing Credibility in Social Media Feeds,* Submitted to SIGIR 2025

Concluding Remarks

CENTER FOR aUTonomy

- **Challenges**:
  - Information ecosystem is not completely understood
  - Information ecosystem is fragile against adversarial attacks
  - Adversaries have initiatives
- **Future Work**:
  - How to design a robust social network?
  - Long term evolution of the network
  - How to integrate multiple information modalities?

Conclusion

Control and decision making enable new solutions for this challenge that goes well beyond fact-checking.