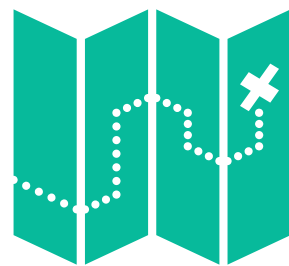


REPHRAIN MAP



Kopo Marvin Ramokapane
Partha Das Chowdhury
Andres Dominguez Hernández
Alicia Cork
Emily Johnstone
Emily Godwin
Ola Michalec

REPHRAIN MAP

AIM

to establish a baseline of current state-of-the-art.

Features

- a living resource (updated regularly)
- inspired by the Mitre ATT&CK framework 11 for technical cyber attacks
- (key distinction) the REPHRAIN Map is socio-technical
- Allow ability to drill down into advances that mitigate against particular online harms.
- a barometer to evaluate the Centre's progress with regards to the baseline
- Communication of research findings and recommendations to bodies outside academia

Users of the MAP

- Academia, industry, law enforcement, policymakers, the general public, and various organisations

APPROACH

Collaborative Approach

Phase 1: Scoping Workshops

- Various scoping sessions with academia, industry, partners, and organisations
- Identified five key components

- ☑ Definition(s)
- ☑ Research Gaps and Challenges
- ☑ Current state of the art
- ☑ Tools and Approaches
- ☑ REPHRAIN Projects

Phase 2: Visual Design

- Drafted visual designs
- Harm centric instead of project centric

Phase 3: Populating the MAP

Data Collection

- Online forms
- Workshops
- One-on-One meetings
- Online searches
- Emails
- 232 Papers from REPHRAIN researchers

Data Curation

- Coding papers

Updating Map

} On Going Process

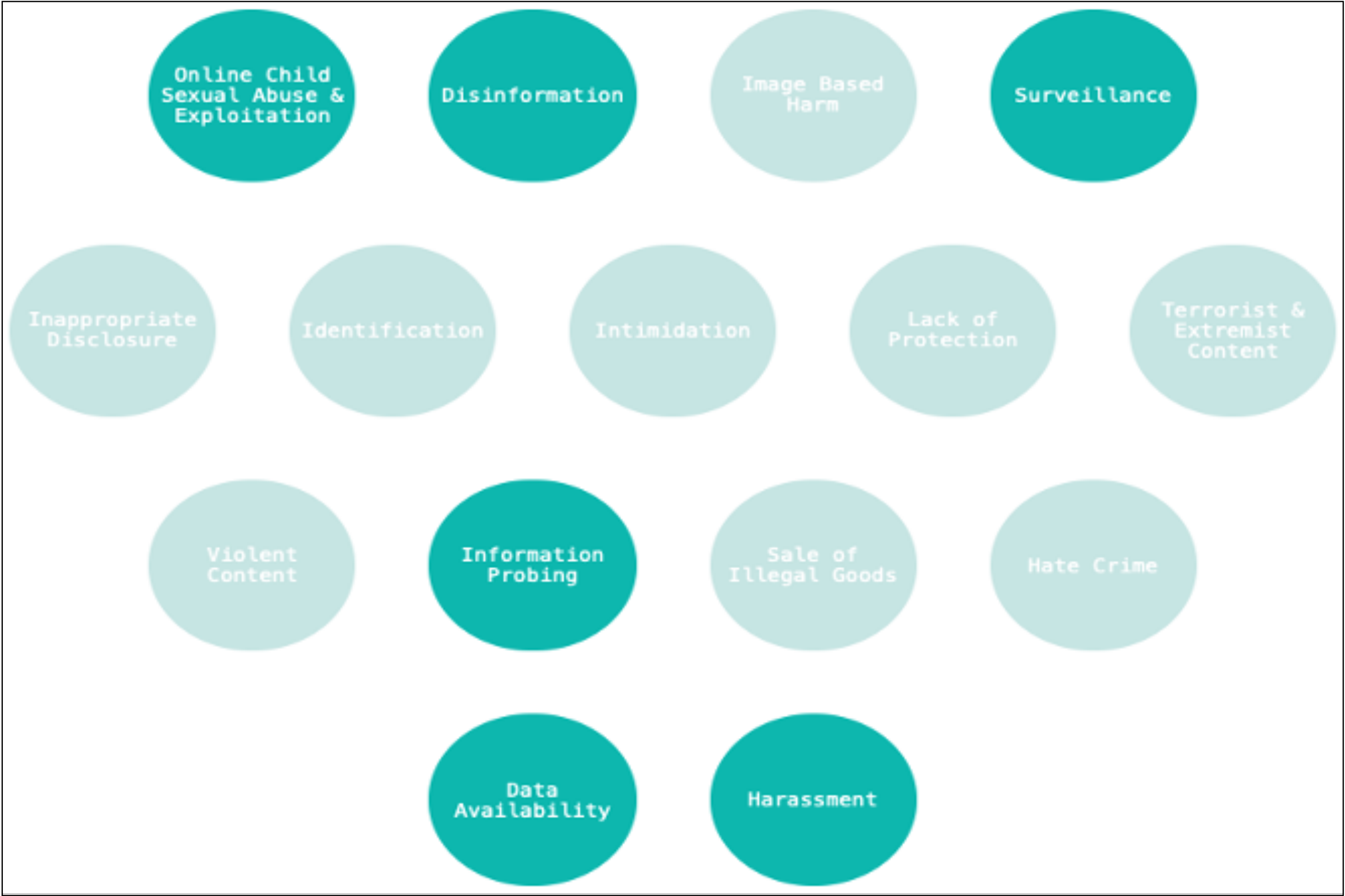
REPHRAIN MAP

Version 0.1

- Harm centric
- Each harm was going to be presented by a circle
- Single entry point
- Navigation stated from individual harms
- Full colour and greyed circles

Online harms


- Disinformation
- Surveillance
- Online Child Sexual Abuse and Exploitation
- Information Probing
- Human trafficking
- Inappropriate/non-consensual disclosure



REPHRAIN MAP

Version 0.1

Go Back to Harms




Definition(s)

Online child communication concurrently

The United Nations defines child knowledgeable parent or caregiver sexual needs trickery, br

Child sexual that involve 2015).

Go Back to Harms



State of the Art

Academic Publications

Liberatore, M., Erdely, R., Kerle, T., Levine, B. N., & Shields, C. (2010). Forensic investigation of peer-to-peer file sharing networks. Digital Investigation, 7, S95–S103. [\(DOI\)](#)

Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. (2016). ICOP: Live forensics to reveal previously unknown criminal media on P2P networks. Digital Investigation, 18, 50–64. [\(DOI\)](#)

Sanchez, L., Grajeda, C., Baggili, I., & Hall, C. (2019). A Practitioner Survey Exploring the Value of Forensic Tools, AI, Filtering, & Safer Presentation for Investigating Child Sexual Abuse Material (CSAM). Digital Investigation, 29, S124–S1 [\(DOI\)](#)

Philppen, A., & Brennan, M. (2019). Child Protection and Safeguarding Technologies: Appropriate or Excessive ‘Solutions’ to Social Problems? Routledge. [\(DOI\)](#)


Policy Documents

[Draft Online Safety Bill](#)

White Papers & Reports

Information is being curated

Go Back to Harms



Challenges

Research Gaps & Challenges

Access to data to train AI models in a secure and privacy preserving way.

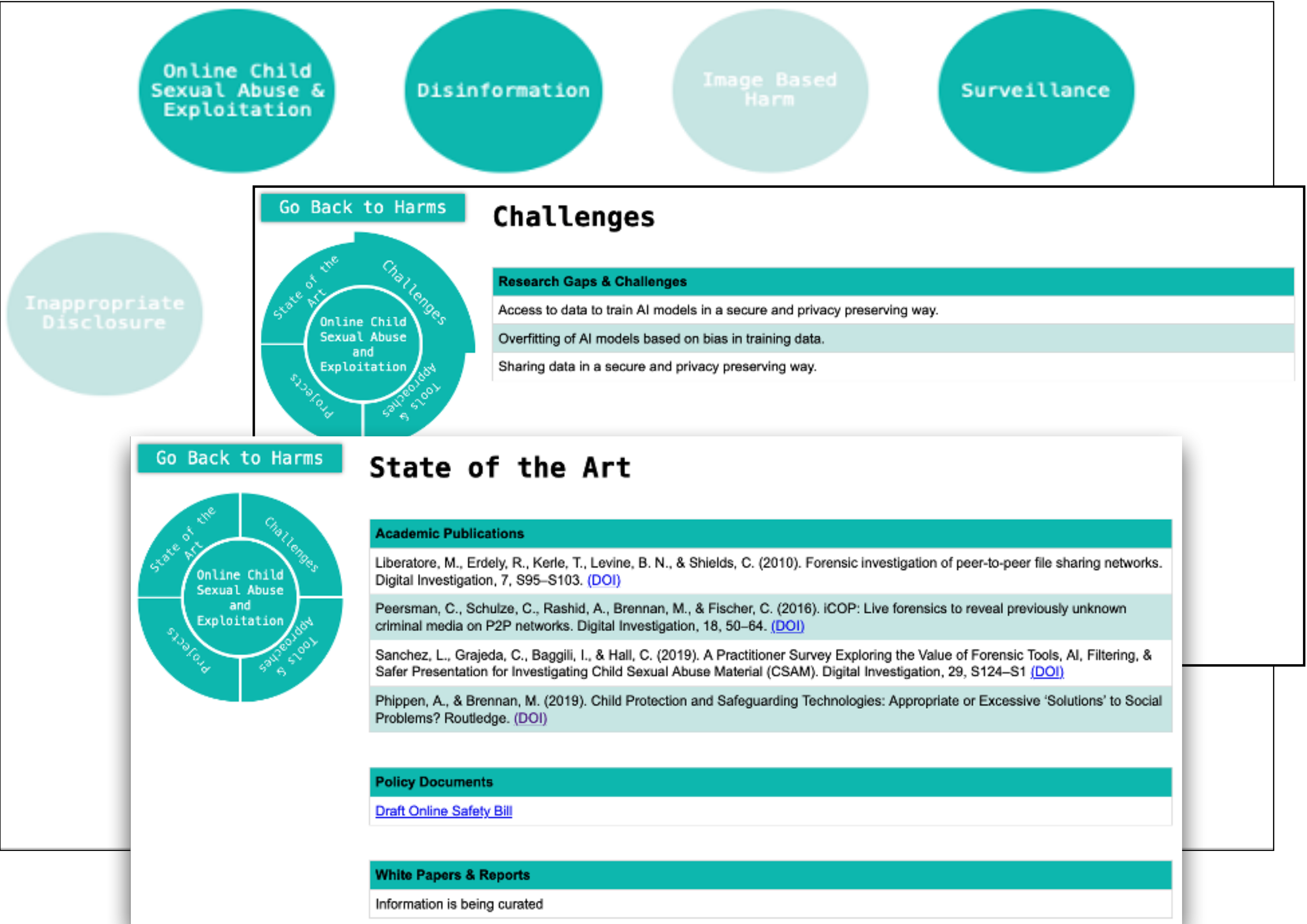
Overfitting of AI models based on bias in training data.

Sharing data in a secure and privacy preserving way.

REPHRAIN MAP

Public Consultation

- ❑ The map was resourceful and potential to be a useful tool
- ❑ Bubbles provided nothing meaningful
- ❑ There was no relationship between the harms
- ❑ Terminology
- ❑ No guidance for users (No use cases)
- ❑ Confusion over full colour and greyed circles

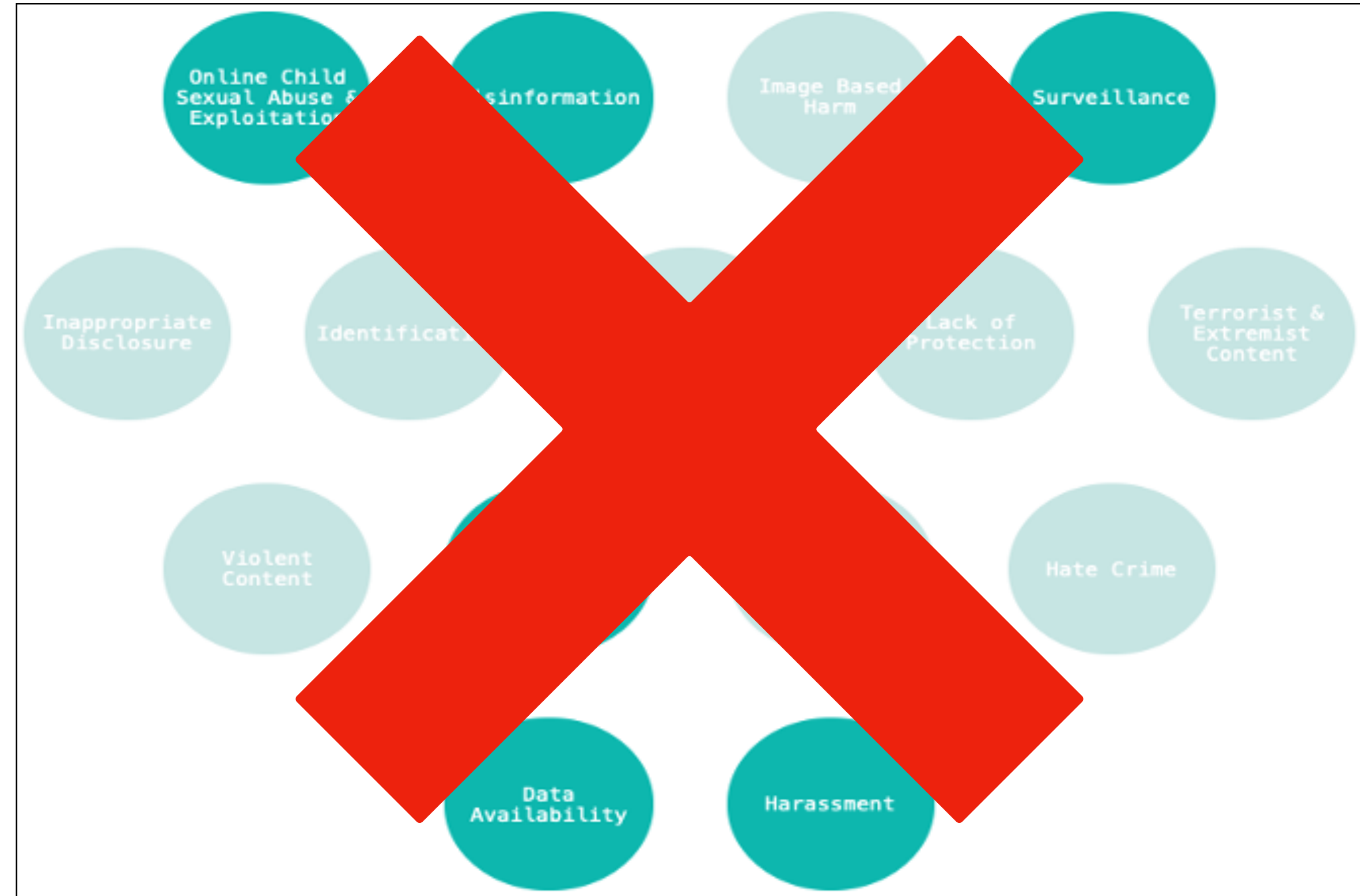


REPHRAIN MAP

New framework of classifying harms

Threat model considered desirable or positive attributes and UN Human Rights list.

- **Privacy**
- **Safety**
- **Reputation**
- **Financial security**
- **Freedom of speech**
- **Fairness**



REPHRAIN MAP

Positive attributes and Harms/Risks

| Privacy | Safety | Reputation | Financial Security | Freedom of Speech | Fairness |
|--|--|--|--|---|--|
| <ul style="list-style-type: none">• Surveillance/ Dataveillance• CSAM• Information Probing• Non-Consensual Disclosure | <ul style="list-style-type: none">• Intimidation/Harassment• Non-Consensual Disclosure• CSAM• Hate Crime• Human Trafficking• Surveillance• Violent Content• Image Based Harm• Sale of Illegal Goods• Institutional Discrimination | <ul style="list-style-type: none">• Image Based Harm• Non-Consensual Disclosure• CSAM• M(D)isinformation• Institutional Discrimination | <ul style="list-style-type: none">• Non-Consensual Disclosure• Surveillance• Human Trafficking• Sale of Illegal Goods• Information Probing• Institutional Discrimination• Bank Fraud | <ul style="list-style-type: none">• Censorship• Self-Censorship/Chilling Effects• Intimidation/Harassment | <ul style="list-style-type: none">• Institutional Discrimination• Intimidation/ Harassment• Image Based Harm• Hate Crime• Surveillance• Information probing |

REPHRAIN MAP

Summary of the major changes

Landing page

- Moved from bubbles to a sangkey diagram
- “Online harms” to "harms, risks and vulnerabilities"
- Two entries: positive attributes and online harms/risks

Categories

- Four Components
 - Description of the harm
 - Research challenges
 - REPHRAIN projects
 - Related resources

New harms and updated terminology,

- Human trafficking / modern day slavery
- Information probing and phishing
- Cyber bullying and harassment

Added contributions from REPHRAIN Researchers

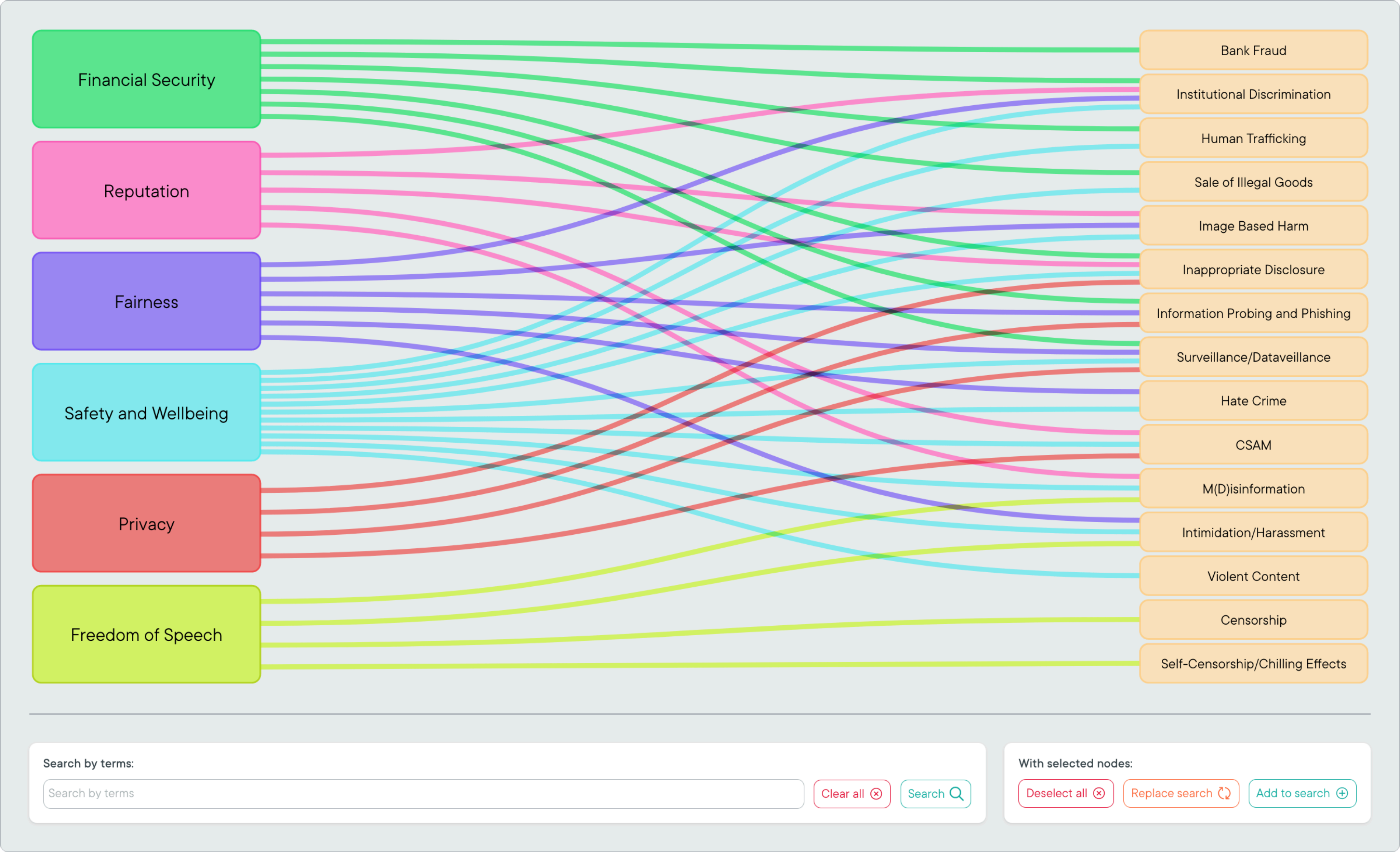
REPHRAIN MAP

Version 1.0

Link: <https://rephrain-map.co.uk>

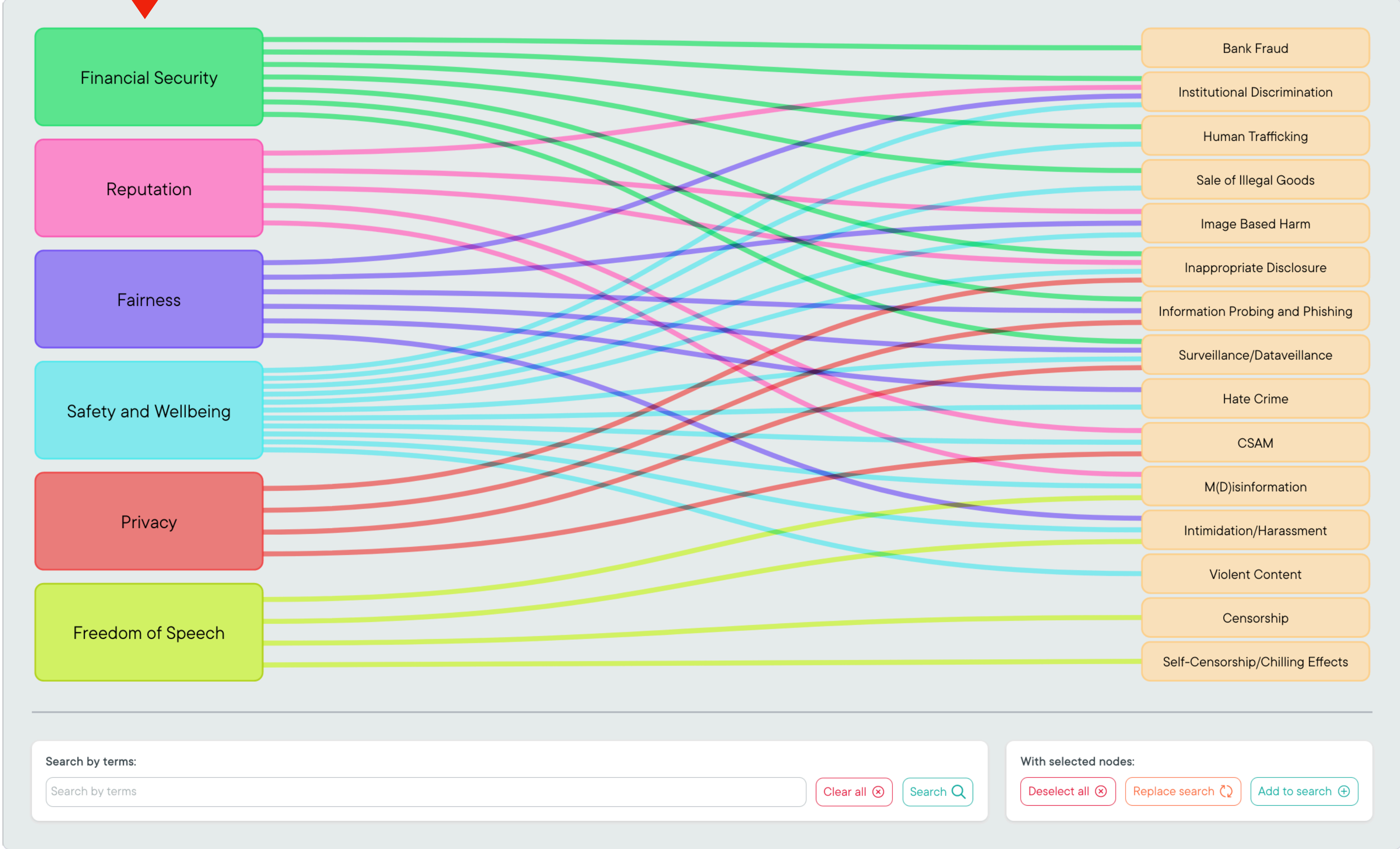
Link: <https://www.rephrain.ac.uk/rephrain-map/>

REPHRAIN MAP v 1.0

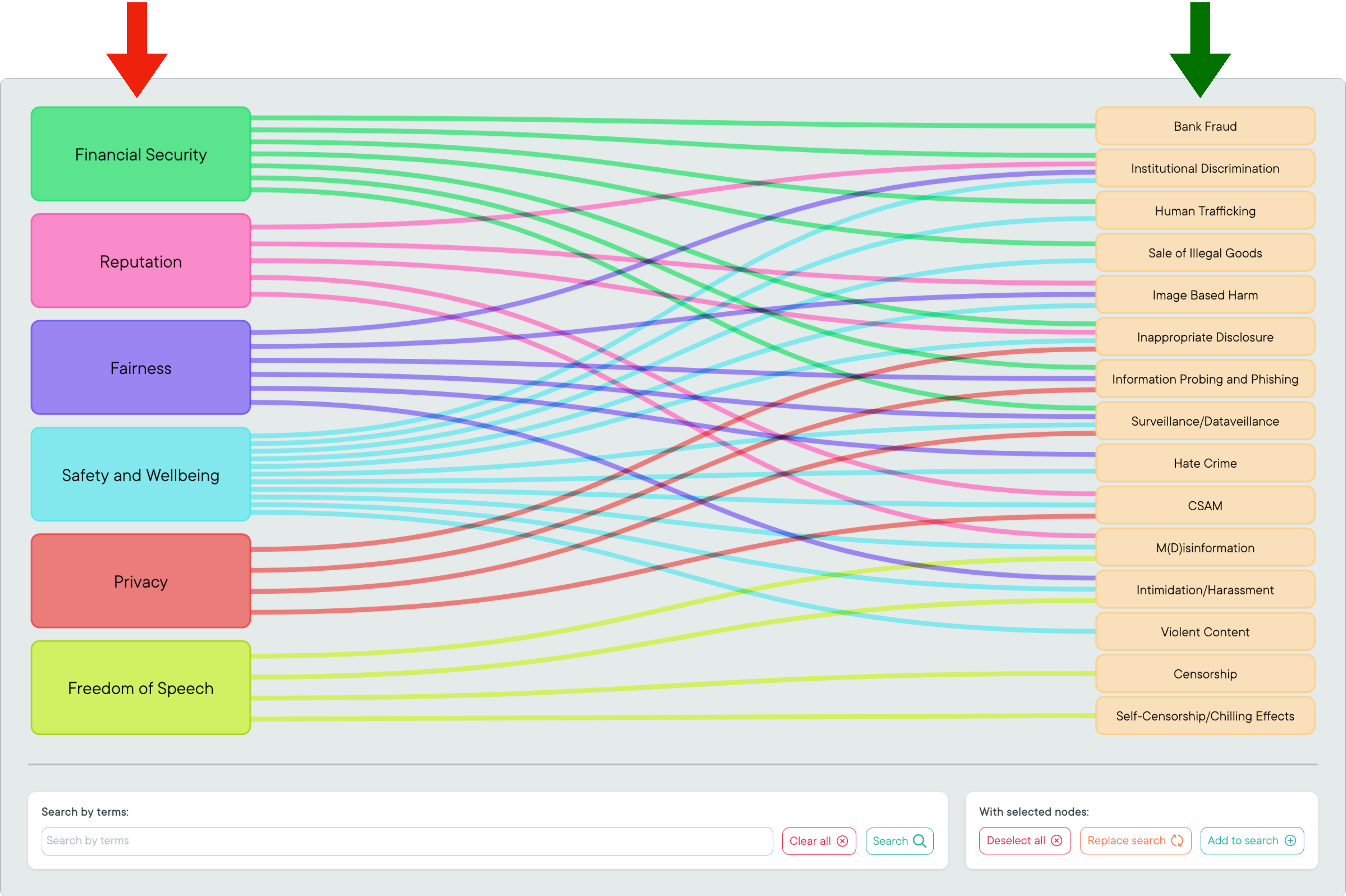


Positive Attribute

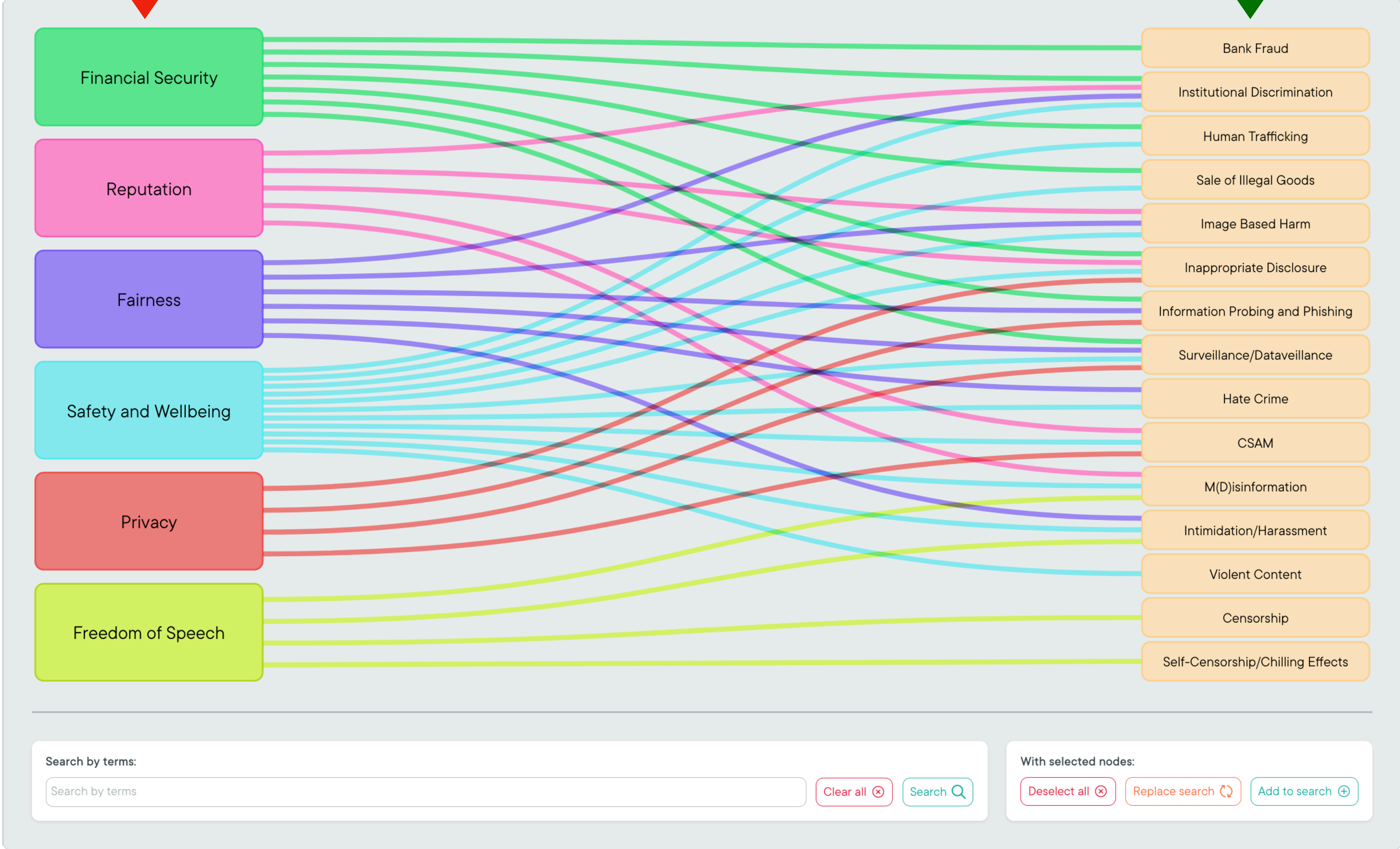
REPHRAIN MAP v 1.0



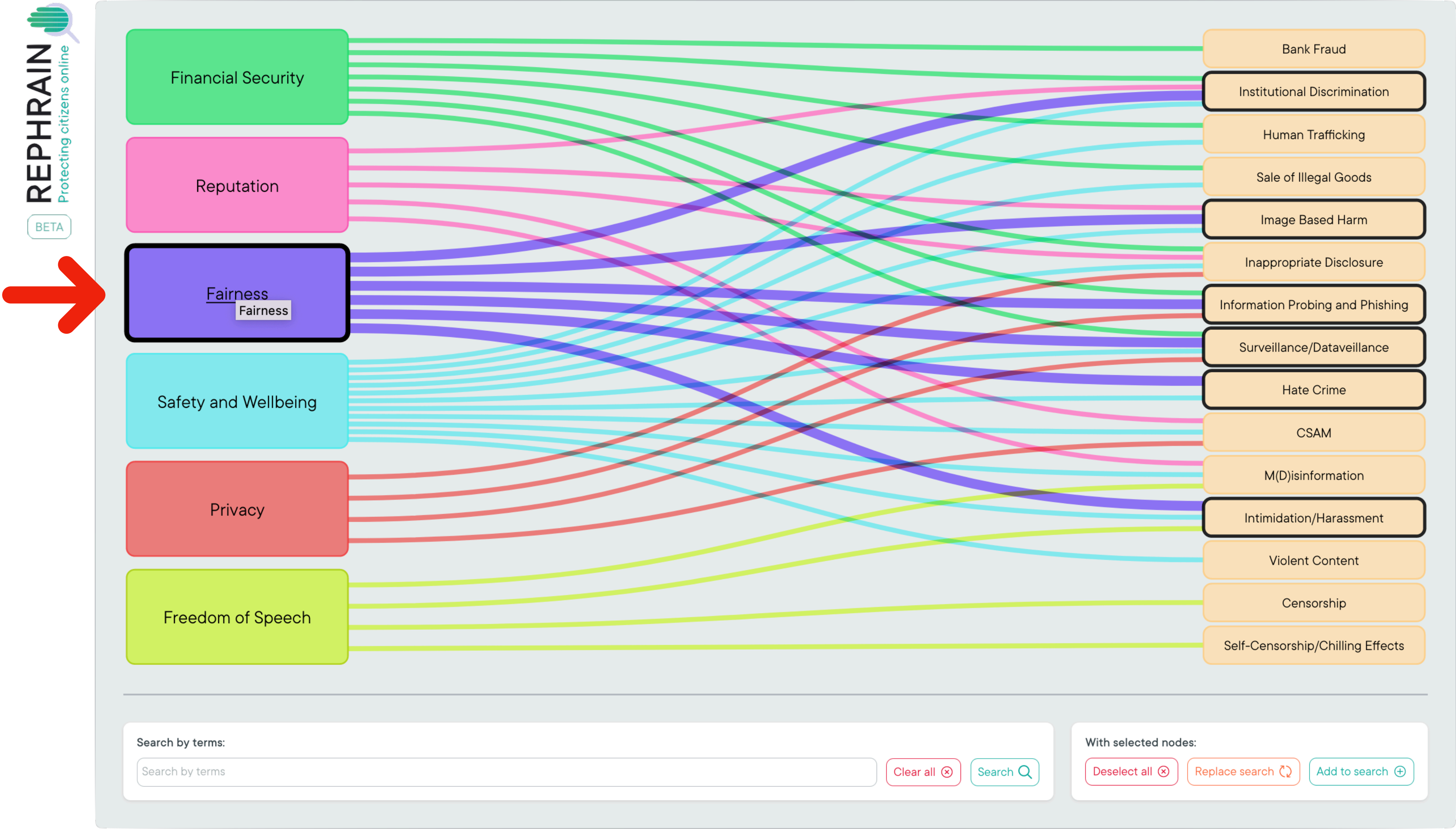
REPHRAIN MAP v 1.0



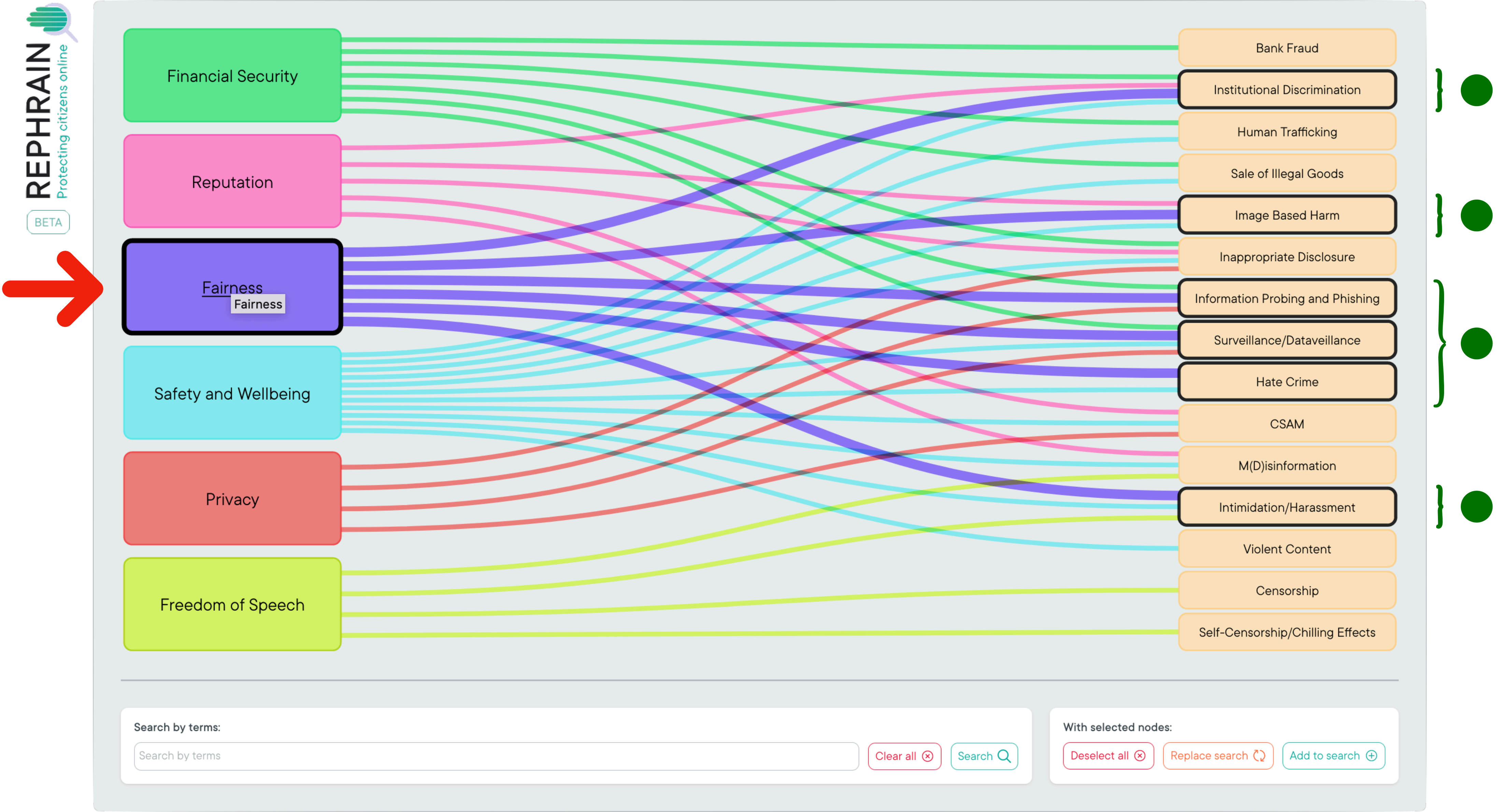
REPHRAIN MAP v 1.0



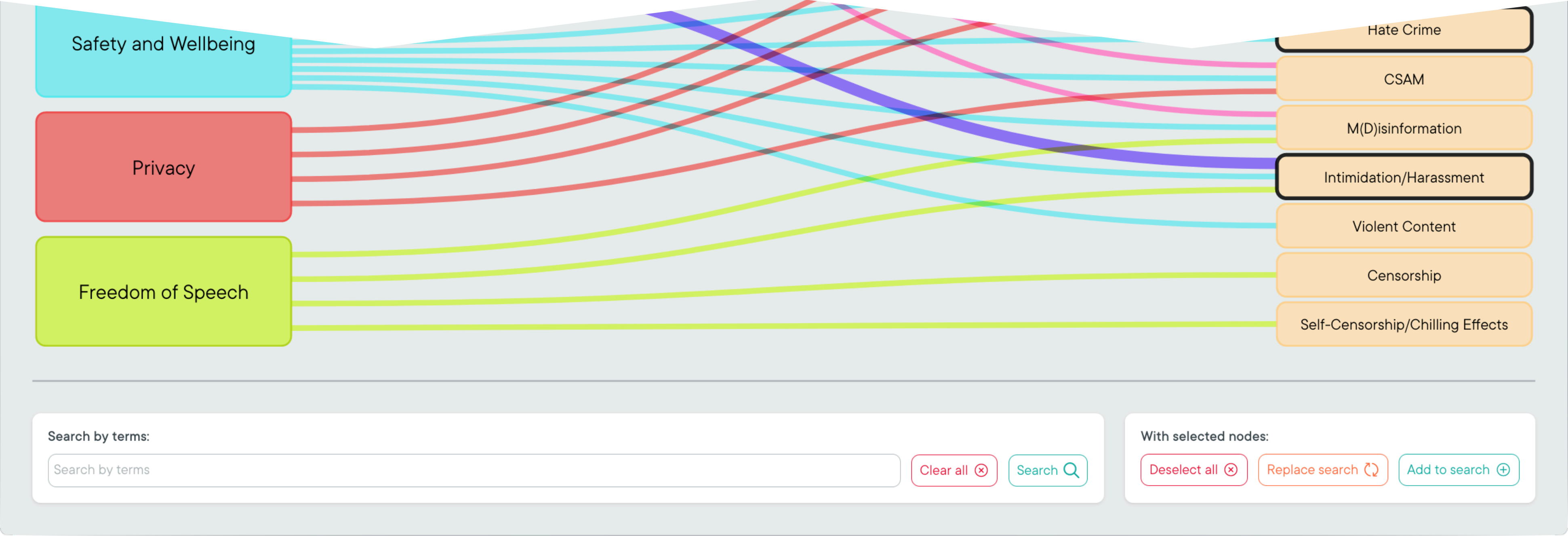
REPHRAIN MAP v 1.0



REPHRAIN MAP v 1.0



REPHRAIN MAP v 1.0



Positive Property:

Fairness

Participating online should be inclusive, respectful, and free from biases. This category includes harms that are a consequence of bias by systems.

Related Harms:

Hate Crime

Image Based Harm

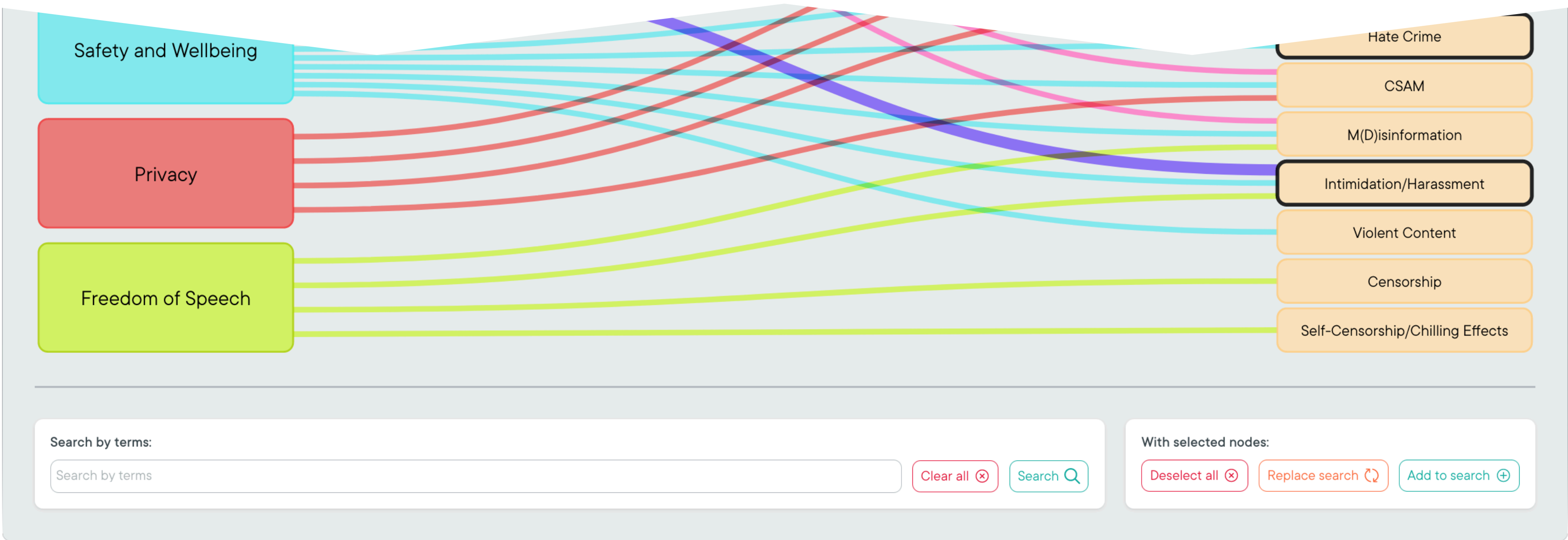
Information Probing and Phishing

Institutional Discrimination

Intimidation/Harassment

Surveillance/Dataveillance

REPHRAIN MAP v 1.0



Positive
Property

Positive Property:

Fairness

Participating online should be inclusive, respectful, and free from biases. This category includes harms that are a consequence of bias by systems.

Related Harms:

Hate Crime

Image Based Harm

Information Probing and Phishing

Institutional Discrimination

Intimidation/Harassment

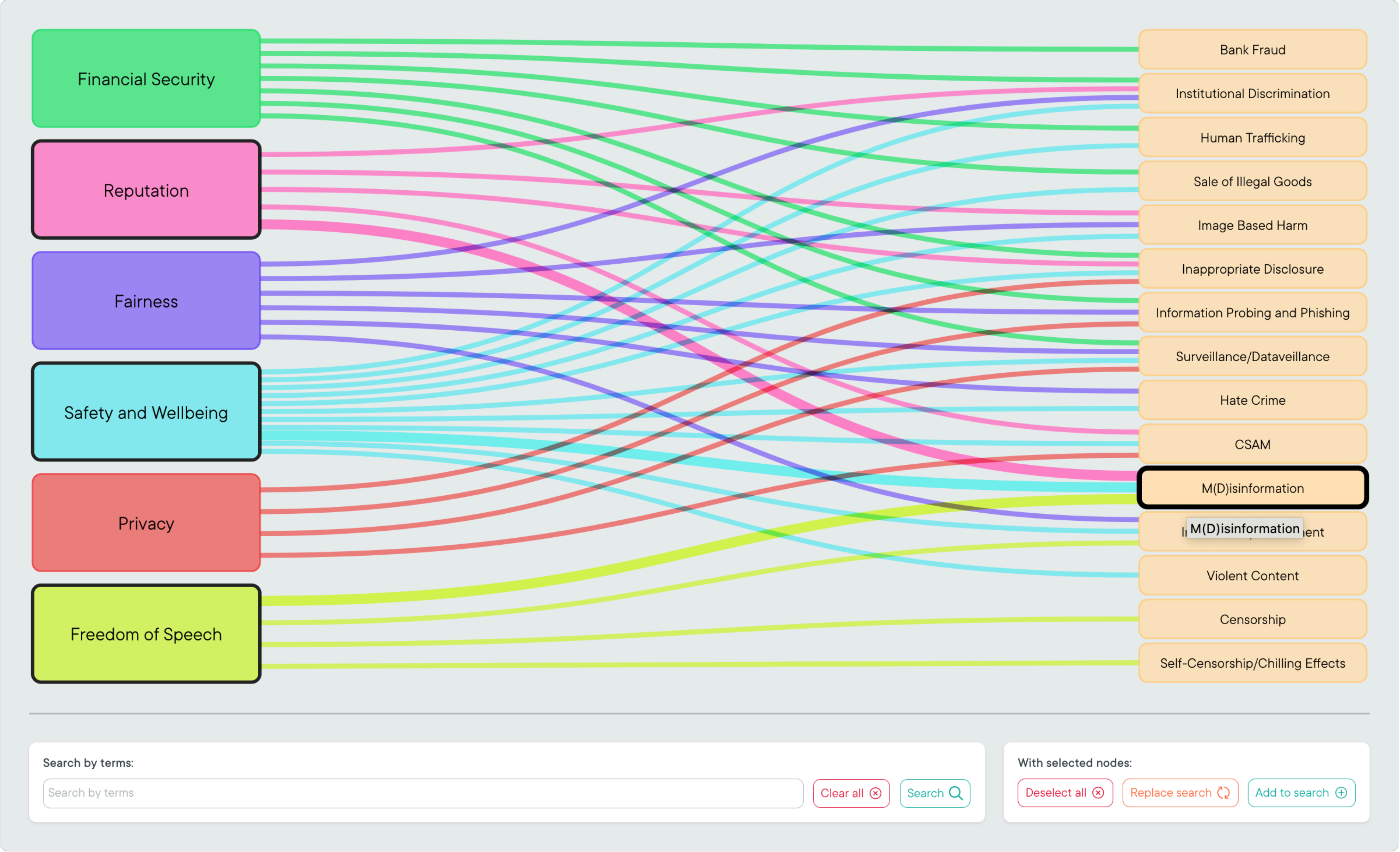
Surveillance/Dataveillance

Online
harms/risks

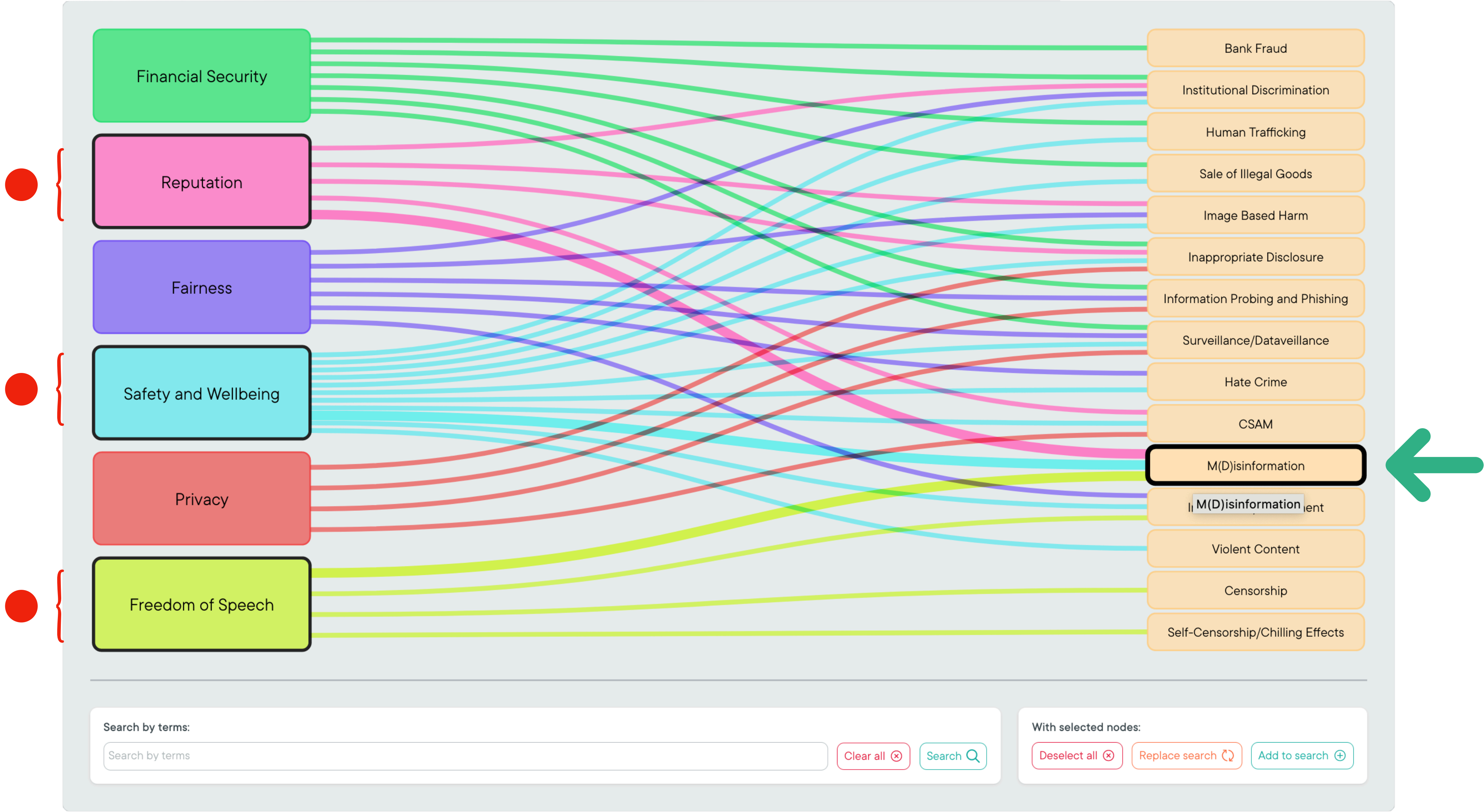
POSITIVE PROPERTY: FAIRNESS



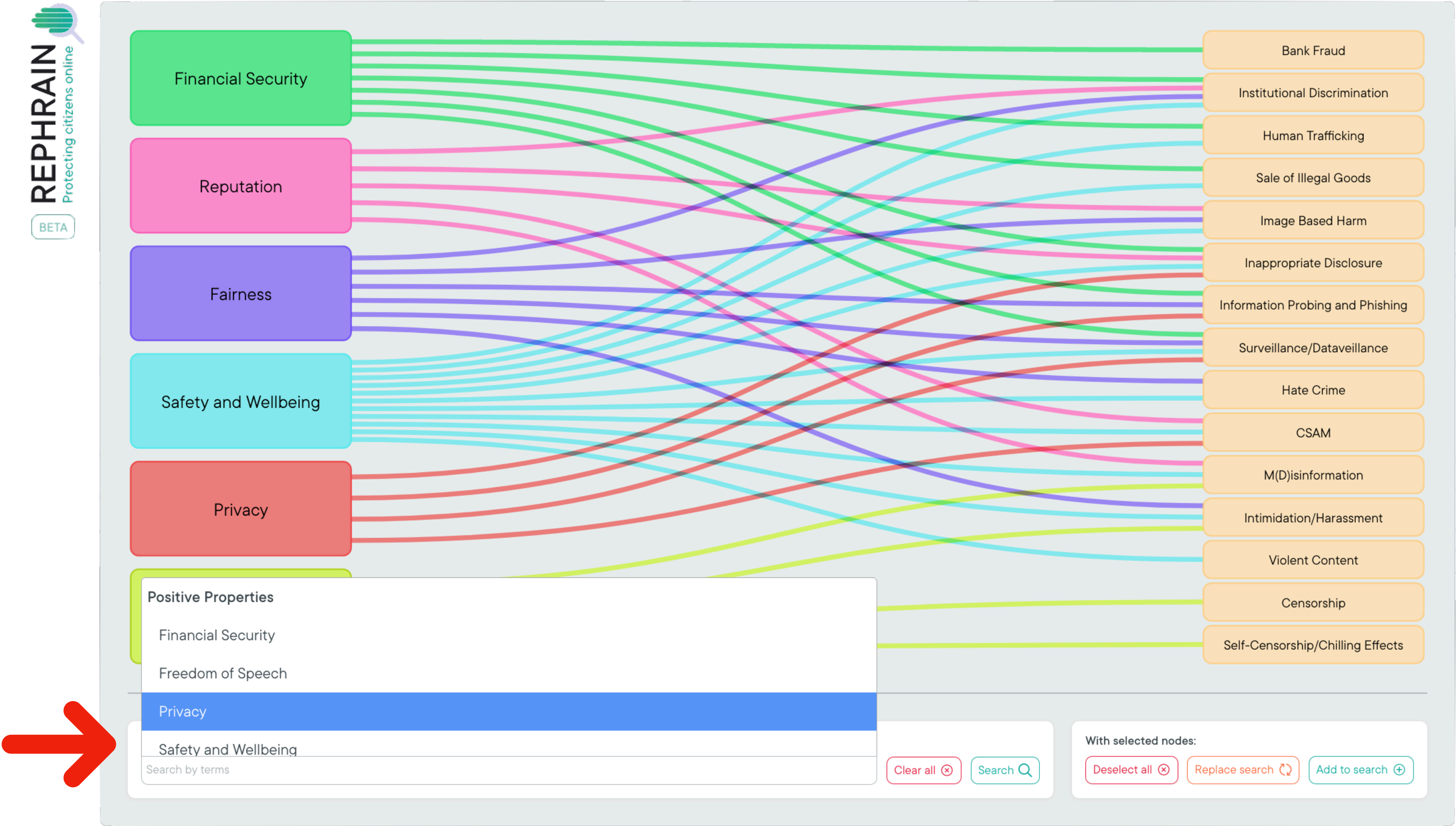
REPHRAIN MAP v 1.0



REPHRAIN MAP v 1.0



REPHRAIN MAP v 1.0



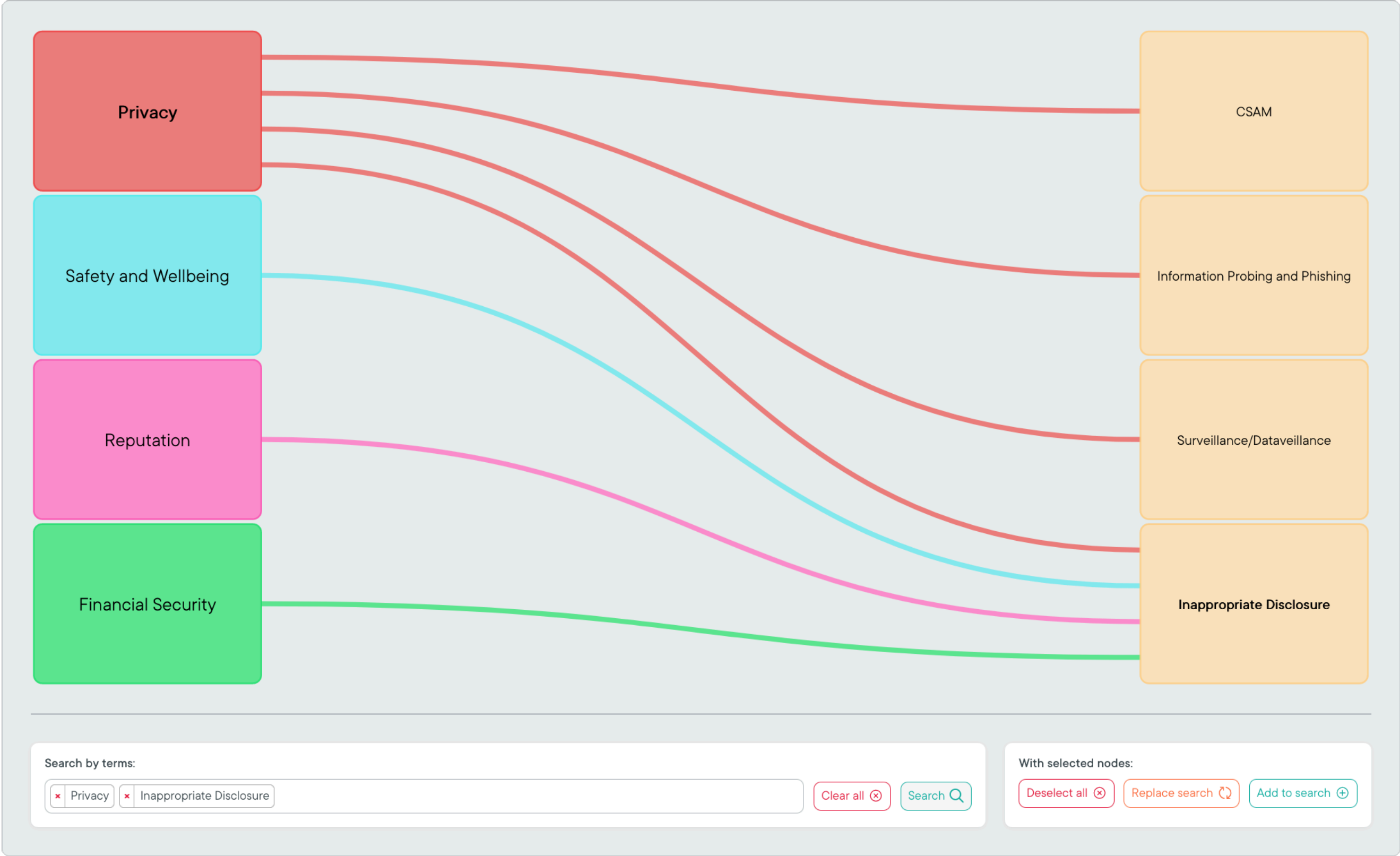
Search: Positive Attribute



Search: **Positive Attribute**



Search: **Positive Attribute + Online harm**



Harm:

M(D)isinformation

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas wilful deceptions are labelled disinformation. Disinformation thus refers to the subset of misinformation that is spread intentionally, although the psychological effect—and harm—on the recipient is likely to be the same regardless of intent.

Related Positive Properties:

Reputation

Safety and Wellbeing

Freedom of Speech

Research Challenges:

These research challenges have evolved from REPHRAIN researchers working in the area.

1. What constitutes disinformation? We currently rely on professional journalists' verdicts, but sometimes they disagree
2. Current research on automatic misinformation detection is almost exclusively in English, despite facts arising in all languages around the world
3. Current research on automatic misinformation detection only uses a couple of modalities (text and images), despite there being many other features available in real-world situations, such as the social network of the person stating the claim or the replies to the claim. High-quality datasets are quite scarce.
4. High-quality datasets are quite scarce
5. Locating the relevant social contexts that are sharing and discussing a relevant fact-checked claim
6. Ensuring that as many languages are represented as possible

REPHRAIN Projects:

Relevant ongoing and past projects funded under REPHRAIN.

| | | |
|--|---|--|
| Bureau Citizens Data Advice Bureau Visit project  | Clariti Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health Visit project  | MITIGATE Understanding and Auditing the Impact of Mitigation Strategies on Online Harms Visit project  |
| NEWS Predicting Personality from News Consumption Visit project  | SURVEY Global Survey of Policy Approaches to Protecting Citizens Online Visit project  | |

Related Resources:

- All
- Academic Literature
- Policy Documents
- Other Approaches
- Whitepapers

Description of
the Harm/risk

Harm:

M(D)isinformation

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas wilful deceptions are labelled disinformation. Disinformation thus refers to the subset of misinformation that is spread intentionally, although the psychological effect—and harm—on the recipient is likely to be the same regardless of intent.

Related Positive Properties:

Reputation

Safety and Wellbeing

Freedom of Speech

Related Positive
Properties

Research Challenges:

These research challenges have evolved from REPHRAIN researchers working in the area.

1. What constitutes disinformation? We currently rely on professional journalists' verdicts, but sometimes they disagree
2. Current research on automatic misinformation detection is almost exclusively in English, despite facts arising in all languages around the world
3. Current research on automatic misinformation detection only uses a couple of modalities (text and images), despite there being many other features available in real-world situations, such as the social network of the person stating the claim or the replies to the claim. High-quality datasets are quite scarce.
4. High-quality datasets are quite scarce
5. Locating the relevant social contexts that are sharing and discussing a relevant fact-checked claim
6. Ensuring that as many languages are represented as possible

REPHRAIN Projects:

Relevant ongoing and past projects funded under REPHRAIN.

Bureau

Citizens Data Advice Bureau

[Visit project](#)

Clariti

Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health

[Visit project](#)

MITIGATE

Understanding and Auditing the Impact of Mitigation Strategies on Online Harms

[Visit project](#)

NEWS

Predicting Personality from News Consumption

[Visit project](#)

SURVEY

Global Survey of Policy Approaches to Protecting Citizens Online

[Visit project](#)

Related Resources:

All

Academic Literature

Policy Documents

Other Approaches

Whitepapers

Description of
the Harm/risk

Research
Challenges

Related Positive
Properties

Harm:

M(D)isinformation

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas wilful deceptions are labelled disinformation. Disinformation thus refers to the subset of misinformation that is spread intentionally, although the psychological effect—and harm—on the recipient is likely to be the same regardless of intent.

Related Positive Properties:

Reputation

Safety and Wellbeing

Freedom of Speech

Research Challenges:

These research challenges have evolved from REPHRAIN researchers working in the area.

- 1. What constitutes disinformation? We currently rely on professional journalists' verdicts, but sometimes they disagree
- 2. Current research on automatic misinformation detection is almost exclusively in English, despite facts arising in all languages around the world
- 3. Current research on automatic misinformation detection only uses a couple of modalities (text and images), despite there being many other features available in real-world situations, such as the social network of the person stating the claim or the replies to the claim. High-quality datasets are quite scarce.
- 4. High-quality datasets are quite scarce
- 5. Locating the relevant social contexts that are sharing and discussing a relevant fact-checked claim
- 6. Ensuring that as many languages are represented as possible

REPHRAIN Projects:

Relevant ongoing and past projects funded under REPHRAIN.

Bureau

Citizens Data Advice Bureau

Visit project

Clariti

Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health

Visit project

MITIGATE

Understanding and Auditing the Impact of Mitigation Strategies on Online Harms

Visit project

NEWS

Predicting Personality from News Consumption

Visit project

SURVEY

Global Survey of Policy Approaches to Protecting Citizens Online

Visit project

Related Resources:

All

Academic Literature

Policy Documents

Other Approaches

Whitepapers

Description of
the Harm/risk

Harm:

M(D)isinformation

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas wilful deceptions are labelled disinformation. Disinformation thus refers to the subset of misinformation that is spread intentionally, although the psychological effect—and harm—on the recipient is likely to be the same regardless of intent.

Related Positive Properties:

Reputation

Safety and Wellbeing

Freedom of Speech

Related Positive
Properties

Research
Challenges

Research Challenges:

These research challenges have evolved from REPHRAIN researchers working in the area.

1. What constitutes disinformation? We currently rely on professional journalists' verdicts, but sometimes they disagree
2. Current research on automatic misinformation detection is almost exclusively in English, despite facts arising in all languages around the world
3. Current research on automatic misinformation detection only uses a couple of modalities (text and images), despite there being many other features available in real-world situations, such as the social network of the person stating the claim or the replies to the claim. High-quality datasets are quite scarce.
4. High-quality datasets are quite scarce
5. Locating the relevant social contexts that are sharing and discussing a relevant fact-checked claim
6. Ensuring that as many languages are represented as possible

REPHRAIN
Projects

REPHRAIN Projects:

Relevant ongoing and past projects funded under REPHRAIN.

Bureau

Citizens Data Advice Bureau

[Visit project](#)

Clariti

Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health

[Visit project](#)

MITIGATE

Understanding and Auditing the Impact of Mitigation Strategies on Online Harms

[Visit project](#)

NEWS

Predicting Personality from News Consumption

[Visit project](#)

SURVEY

Global Survey of Policy Approaches to Protecting Citizens Online

[Visit project](#)

Related Resources:

All

Academic Literature

Policy Documents

Other Approaches

Whitepapers

Description of
the Harm/risk

Harm:

M(D)isinformation

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas wilful deceptions are labelled disinformation. Disinformation thus refers to the subset of misinformation that is spread intentionally, although the psychological effect—and harm—on the recipient is likely to be the same regardless of intent.

Related Positive Properties:

Reputation

Safety and Wellbeing

Freedom of Speech

Related Positive
Properties

Research
Challenges

Research Challenges:

These research challenges have evolved from REPHRAIN researchers working in the area.

1. What constitutes disinformation? We currently rely on professional journalists' verdicts, but sometimes they disagree
2. Current research on automatic misinformation detection is almost exclusively in English, despite facts arising in all languages around the world
3. Current research on automatic misinformation detection only uses a couple of modalities (text and images), despite there being many other features available in real-world situations, such as the social network of the person stating the claim or the replies to the claim. High-quality datasets are quite scarce.
4. High-quality datasets are quite scarce
5. Locating the relevant social contexts that are sharing and discussing a relevant fact-checked claim
6. Ensuring that as many languages are represented as possible

REPHRAIN
Projects

REPHRAIN Projects:

Relevant ongoing and past projects funded under REPHRAIN.

Bureau

Citizens Data Advice Bureau

[Visit project](#)

Clariti

Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health

[Visit project](#)

MITIGATE

Understanding and Auditing the Impact of Mitigation Strategies on Online Harms

[Visit project](#)

NEWS

Predicting Personality from News Consumption

[Visit project](#)

SURVEY

Global Survey of Policy Approaches to Protecting Citizens Online

[Visit project](#)

Related
Resources

Related Resources:

All

Academic Literature

Policy Documents

Other Approaches

Whitepapers

Filters

Search bar

HARM: HATE CRIME

Harm:

Hate Crime

Related Positive Properties:

Safety and Wellbeing

Fairness

Related Resources:

All

Academic Literature

Policy Documents

Other Approaches

Whitepapers

Search by terms:

Search by terms

Clear all

Search

| Resource Type | Resource title | Resource authors | Resource external link | Details |
|-------------------------------------|--|------------------|------------------------|---------|
| Academic Literature | Discovering and interpreting conceptual biases in online communities | Ferrer, X. ... | | |

Filters

Search bar

Harm:

Hate Crime

Related Positive Properties:

Safety and Wellbeing

Fairness

Related Resources:

All

Academic Literature

Policy Documents

Other Approaches

Whitepapers

Search by terms:

Search by terms

Clear all

Search

| Resource Type | Resource title | Resource authors | Resource external link | Details |
|---------------------|--|------------------|------------------------|---------|
| Academic Literature | Discovering and interpreting conceptual biases in online communities | Ferrer, X. ... | | |

Detailed info about
a resource

Filters

Search bar

Harm:

Hate Crime

Related Positive Properties:

Safety and Wellbeing

Fairness

Related Resources:

AllAcademic LiteraturePolicy DocumentsOther ApproachesWhitepapers

Search by terms:

Search by terms

Clear all

Search

| Resource Type | Resource title | Resource authors | Resource external link | Details |
|---------------------|--|------------------|------------------------|---------|
| Academic Literature | Discovering and interpreting conceptual biases in online communities | Ferrer, X. ... | | |

Detailed info about
a resource

Resources (e.g., papers) are
coded according to:

- Harm being addressed
- Adopted methodology
- Platform or technology
- Targeted group/Victims
- Perpetrators

All Resource details

Authors:

Ferrer, X., van Nuenen, T., Such, J. M., Criado, N.

Harms:

Hate Crime

Publication:

IEEE Transactions on Knowledge and Data Engineering

Resource Type:

Academic Literature

Methodologies:

Natural language processing

Word embeddings

Perpetrators:

(none)

Platforms/Technologies:

Forums

Fringe platforms

Victims:

VR users

Bystanders

Title:

Discovering and interpreting conceptual biases in online communities

External link:

(none)

LESSONS LEARNED

REPHRAIN MAP as a METHOD

- Shows how different bodies of knowledge link
- Brings research from different disciplines together
- Translate concepts from specific discipline-specific jargon
- Visualises complex areas of research

| Victims | Perpetrator | Platform/Technology | Methodology |
|-------------------------|----------------------------|--------------------------|-------------------------|
| Bystanders | Campaign groups | AI systems | Anomaly detection |
| Children | Darknet communities | Internet of Things | Case studies |
| Consumers | Extremist groups | Cloud systems | Usability Studies |
| IPV victims | Government agencies | Contact tracing apps | Detection system |
| Online dating users | IPV perpetrators | Content-sharing services | Digital forensics |
| Political organisations | Law enforcement | Critical infrastructure | Digital traces |
| Sex workers | Nation State | Darknet markets | Ethnography |
| Social media users | Online fitness communities | Emails | Experimental |
| Teenagers | Organized crime groups | E-recruitment platforms | Focus groups |
| Refugees | Romance Scammers | Virtual Reality | Interviews |
| Bystanders | Sex Offenders | Social Media Platforms | Surveys |
| Women | Social Media Users | Smartphones | Social Network Analysis |

REPHRAIN MAP as a METHOD

- Shows how different bodies of knowledge link
- Brings research from different disciplines together
- Translate concepts from specific discipline-specific jargon
- Visualises complex areas of research

REPHRAIN MAP as a MEDIUM

- Shows what areas of research have been covered
- The research gaps that need to be bridged
- Existing tools or approaches to tackle online harm/risks
- What REPHRAIN is doing and areas that still need attention

| Victims | Perpetrator | Platform/Technology | Methodology |
|-------------------------|----------------------------|--------------------------|-------------------------|
| Bystanders | Campaign groups | AI systems | Anomaly detection |
| Children | Darknet communities | Internet of Things | Case studies |
| Consumers | Extremist groups | Cloud systems | Usability Studies |
| IPV victims | Government agencies | Contact tracing apps | Detection system |
| Online dating users | IPV perpetrators | Content-sharing services | Digital forensics |
| Political organisations | Law enforcement | Critical infrastructure | Digital traces |
| Sex workers | Nation State | Darknet markets | Ethnography |
| Social media users | Online fitness communities | Emails | Experimental |
| Teenagers | Organized crime groups | E-recruitment platforms | Focus groups |
| Refugees | Romance Scammers | Virtual Reality | Interviews |
| Bystanders | Sex Offenders | Social Media Platforms | Surveys |
| Women | Social Media Users | Smartphones | Social Network Analysis |

REPHRAIN MAP as a METHOD

- Shows how different bodies of knowledge link
- Brings research from different disciplines together
- Translate concepts from specific discipline-specific jargon
- Visualises complex areas of research

REPHRAIN MAP as a MEDIUM

- Shows what areas of research have been covered
- The research gaps that need to bridged
- Existing tools or approaches to tackle online harm/risks
- What REPHRAIN is doing and areas that still need attention

REPHRAIN MAP as a PROVOCATION

- Debates around the concept of online harm
- Appropriateness of terms “MAP of online harms” vs “MAP of technology-mediated harms”
- Outdated terms

| Victims | Perpetrator | Platform/Technology | Methodology |
|-------------------------|----------------------------|--------------------------|-------------------------|
| Bystanders | Campaign groups | AI systems | Anomaly detection |
| Children | Darknet communities | Internet of Things | Case studies |
| Consumers | Extremist groups | Cloud systems | Usability Studies |
| IPV victims | Government agencies | Contact tracing apps | Detection system |
| Online dating users | IPV perpetrators | Content-sharing services | Digital forensics |
| Political organisations | Law enforcement | Critical infrastructure | Digital traces |
| Sex workers | Nation State | Darknet markets | Ethnography |
| Social media users | Online fitness communities | Emails | Experimental |
| Teenagers | Organized crime groups | E-recruitment platforms | Focus groups |
| Refugees | Romance Scammers | Virtual Reality | Interviews |
| Bystanders | Sex Offenders | Social Media Platforms | Surveys |
| Women | Social Media Users | Smartphones | Social Network Analysis |

Thank you

 rephrain-map@bristol.ac.uk

 @REPHRAIN1

 <https://rephrain-map.co.uk/>

