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

bristol.ac.uk
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
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

**Definition(s)**

Online child sexual abuse and exploitation refers to harmful activities that involve children and young people, typically through digital means. This can include activities such as grooming, explicit communications, and the sharing or production of child sexual abuse images and videos.

**State of the Art**

**Academic Publications**


**Policy Documents**

- Draft Online Safety Bill

**White Papers & Reports**

- Information is being curated

**Challenges**

- Research Gaps & Challenges
  - Access to data to train AI models in a secure and privacy-preserving way.
  - Overfiling of AI models based on bias in training data.
  - Sharing data in a secure and privacy-preserving way.
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

<table>
<thead>
<tr>
<th>Privacy</th>
<th>Safety</th>
<th>Reputation</th>
<th>Financial Security</th>
<th>Freedom of Speech</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Surveillance/ Dataveillance</td>
<td>• Intimidation/Harassment</td>
<td>• Image Based Harm</td>
<td>• Non-Consensual Disclosure</td>
<td>• Censorship</td>
<td>• Institutional Discrimination</td>
</tr>
<tr>
<td>• CSAM</td>
<td>• Non-Consensual Disclosure</td>
<td>• Non-Consensual Disclosure</td>
<td>• Surveillance</td>
<td>• Self-Censorship/Chilling Effects</td>
<td>• Intimidation/Harassment</td>
</tr>
<tr>
<td>• Information Probing</td>
<td>• CSAM</td>
<td>• CSAM</td>
<td>• Human Trafficking</td>
<td>• Intimidation/Harassment</td>
<td>• Image Based Harm</td>
</tr>
<tr>
<td>• Non-Consensual Disclosure</td>
<td>• Hate Crime</td>
<td>• M(D)isinformation</td>
<td>• Sale of Illegal Goods</td>
<td>• Information Probing</td>
<td>• Hate Crime</td>
</tr>
<tr>
<td></td>
<td>• Human Trafficking</td>
<td>• Institutional Discrimination</td>
<td>• Sale of Illegal Goods</td>
<td>• Institutional Discrimination</td>
<td>• Surveillance</td>
</tr>
<tr>
<td></td>
<td>• Surveillance</td>
<td></td>
<td>• Sale of Illegal Goods</td>
<td></td>
<td>• Information probing</td>
</tr>
<tr>
<td></td>
<td>• Violent Content</td>
<td></td>
<td>• Institutional Discrimination</td>
<td></td>
<td>• Image Based Harm</td>
</tr>
<tr>
<td></td>
<td>• Image Based Harm</td>
<td></td>
<td></td>
<td></td>
<td>• Hate Crime</td>
</tr>
<tr>
<td></td>
<td>• Sale of Illegal Goods</td>
<td></td>
<td></td>
<td></td>
<td>• Surveillance</td>
</tr>
<tr>
<td></td>
<td>• Institutional Discrimination</td>
<td></td>
<td></td>
<td></td>
<td>• Information probing</td>
</tr>
</tbody>
</table>
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

Financial Security

Reputation

Fairness

Safety and Wellbeing

Privacy

Freedom of Speech

Bank Fraud
Institutional Discrimination
Human Trafficking
Sale of Illegal Goods
Image Based Harm
Inappropriate Disclosure
Information Probing and Phishing
Surveillance/Data Privacy
Hate Crime
CSAM
Misinformation
Intimidation/Harassment
Violent Content
Censorship
Self-Censorship/Chilling Effects
REPHRAIN MAP v 1.0

Positive Attribute

Financial Security
Reputation
Fairness
Safety and Wellbeing
Privacy
Freedom of Speech

Online harms/risks

- Bank Fraud
- Institutional Discrimination
- Human Trafficking
- Sale of Illegal Goods
- Image Based Harm
- Inappropriate Disclosure
- Information Probing and Phishing
- Surveillance/Data Collection
- Hate Crime
- CSAM
- Mis/disinformation
- Intimidation/Harassment
- Violent Content
- Censorship
- Self-Censorship/Chilling Effects

Search functionality

Search by terms:
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/Data Collection
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/Datacensorship
**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.

**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**
  - Citizen Data Advice Bureau
  - Visit project [12]

- **ClariHi**
  - Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health
  - Visit project [12]

- **MITIGATE**
  - Understanding and Auditing the Impact of Mitigation Strategies on Online Harms
  - Visit project [12]

- **NEWS**
  - Predicting Personality from News Consumption
  - Visit project [12]

- **SURVEY**
  - Global Survey of Policy Approaches to Protecting Citizens Online
  - Visit project [12]

**Related Positive Properties:**

- Reputation
- Safety and Wellbeing
- Freedom of Speech
Description of the Harm/risk

M(D)isinformation

Any information that turns out to be false after it was believed to be true is commonly referred to as misinformation, whereas false 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.

Research Challenges:

These research challenges have evolved from REPRAIN 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 REPRAIN

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 →

NEWS
Predicting Personality from News Consumption
Visit project →

MITIGATE
Understanding and Auditing the Impact of Mitigation Strategies on Online Harms
Visit project →

SURVEY
Global Survey of Policy Approaches to Protecting Citizens Online
Visit project →

Related Resources:

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

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 (2)

Clariti
Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health
Visit project (2)

MITIGATE
Understanding and Auditing the Impact of Mitigation Strategies on Online Harms
Visit project (2)

NEWS
Predicting Personality from News Consumption
Visit project (2)

SURVEY
Global Survey of Policy Approaches to Protecting Citizens Online
Visit project (2)

Related Resources:
- Academic Literature
- Policy Documents
- Other Approaches
- Whitepapers

Related Positive Properties:
- Reputation
- Safety and Wellbeing
- Freedom of Speech
**M(D)isinformation**

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas willful deception is 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.

**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 Positive Properties:**

- Reputation
- Safety and Wellbeing
- Freedom of Speech

**Description of the Harm/risk**

**Related Positive Properties**
M(D)isinformation

Any information that turns out to be false after first considered to be true is commonly referred to as misinformation, whereas willful deception is 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.

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 many 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 [2]

**Clariti**
Social Networks and the Real Danger of Pseudoscience, Fake News and Conspiracy Theories to Public Health
Visit project [2]

**MITIGATE**
Understanding and Auditing the Impact of Mitigation Strategies on Online Harms
Visit project [2]

**NEWS**
Predicting Personality from News Consumption
Visit project [2]

**SURVEY**
Global Survey of Policy Approaches to Protecting Citizens Online
Visit project [2]

Related Positive Properties:
- Reputation
- Safety and Wellbeing
- Freedom of Speech

Related Resources:
- Academic Literature
- Policy Documents
- Other Approaches
- Whitepapers
Filters

Search bar

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

Filters
Search bar
Detailed info about a resource
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

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

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

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
Thank you

rephrain-map@bristol.ac.uk
@REPHRAIN1
https://rephrain-map.co.uk/