UNIVERSITY OF BATH

Problematizing and Reconceptualizing Online Harm

Dr Alicia Cork
Prof Adam Joinson
Prof Danae Stanton Fraser
Dr Laura Smith
Dr David Ellis

bath.ac.uk

REPHRAIN

National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online

What harms are included in this work?

Child sexual exploitation and abuse

Terrorist content and activity

Organised immigration crime

Modern slavery

Extreme pornography

Revenge pornography

Harassment and cyberstalking

Hate crime

Sale of illegal goods or services

Content illegally uploaded from prison

Sexting of indecent images by under 18s

Encouraging or assisting suicide

Incitement of violence

Extremist content and activity

Cyberbullying and trolling

Coercive behaviour

Intimidation

Disinformation

Violent content

Promotion of Female Genital Mutilation (FGM)

Advocacy of self-harm

Children accessing pornography

Children accessing in appropriate material (including under 13s using social media and under 18s using data apps; excessive screen time)

Harms in Online Safety Bill

Other harms included in this work

Fraud

Scams

Disinformation

Identity theft

Surveillance

Microtargeting

Profiling

Data theft

Addictions/excessive use

Deepfakes

Malware

Hacking

DDoS attacks

Algorithmic bias

Human trafficking coordination

Excessive online gaming disorders

Inaccessibility

Misinformation

Automated decision making

Excessive online gambling disorders

Threats

Social engineering

Destruction of data

Technological misuse

Blackmail

Epilepsy trolling

Unauthorized access to IoT devices

Sexual harassment

Pro-eating disorder content

Where we started

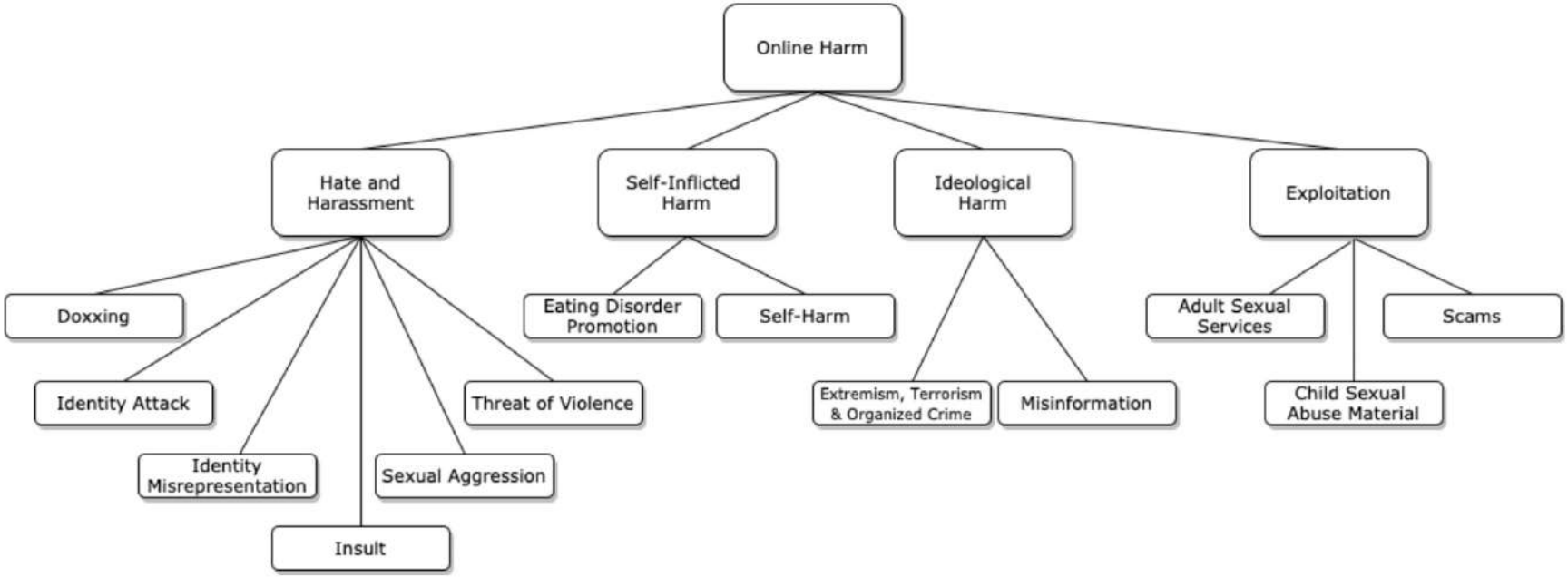
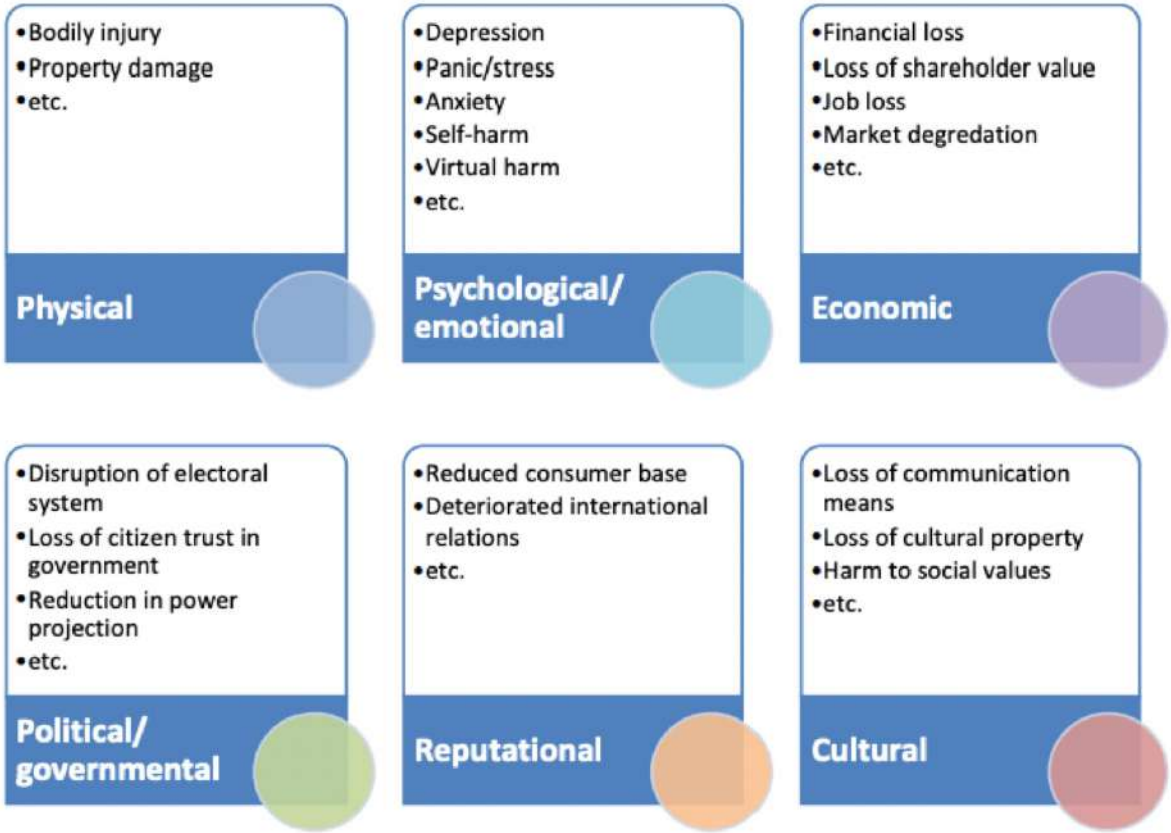


Figure 1: A Typology of Harmful Content

A Unified Taxonomy of Harmful Content (Banko, McKeen & Ray, 2020)

CORE	Content Child as recipient	Contact Child as participant	Conduct Child as actor	Contract Child as consumer
Aggressive	Violent, gory, graphic, racist, hateful and extremist content	Harassment, stalking, hateful behaviour, unwanted surveillance	Bullying, hateful or hostile peer activity e.g. trolling, exclusion, shaming	Identity theft, fraud, phishing, scams, gambling, blackmail, security risks
Sexual	Pornography (legal and illegal), sexualization of culture, body image norms	Sexual harassment, sexual grooming, generation and sharing of child sexual abuse material	Sexual harassment, non-consensual sexual messages, sexual pressures	Sextortion, trafficking for purposes of sexual exploitation, streaming child sexual abuse
Values	Age-inappropriate user-generated or marketing content, mis/disinformation	Ideological persuasion, radicalization and extremist recruitment	Potentially harmful user communities e.g. self-harm, anti-vaccine, peer pressures	Information filtering, profiling bias, polarisation, persuasive design
Cross-cutting	Privacy and data protection abuses, physical and mental health risks, forms of discrimination			

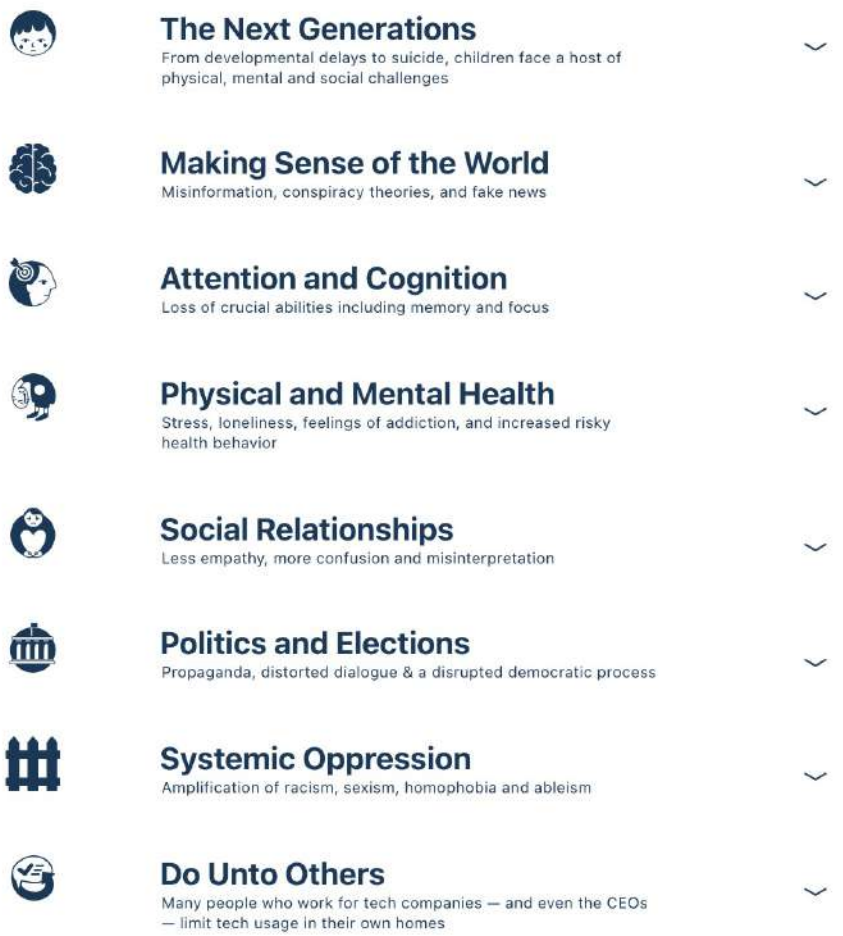
The 4Cs: Classifying Online Risk to Children (Livingstone & Stoilova, 2021)



Cyber Harm: Concepts, Taxonomy and Measurement (Agrafiotis et al., 2016)

Towards a Theory of Harms (Clark & Claffy, 2019):

- Harms to availability of internet access
- Harms to the integrity of the internet experience
- Harms to confidentiality and privacy
- Harms to innovation competition and choice
- Harms to journalism, the marketplace of ideas and the political processes that depend on them.



https://ledger.humanetech.com/

Realisations:

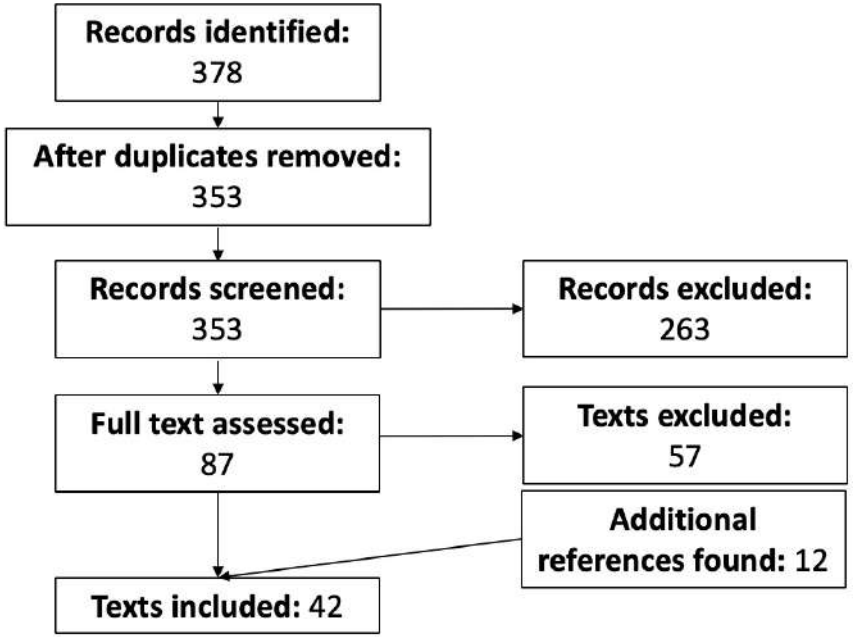
- Linguistic inconsistency in what is meant by ‘Online Harm’
  - Need to discriminate between a risk and the harmful impacts of that risk

- Should we be looking to mitigate exposure to risks or actual harm experienced?
- Many different dimensions of online risk and harm worth considering:
  - Legality
  - Intentionality
  - Likelihood of exposure
  - Adult victims or child victims
  - Severity of harm

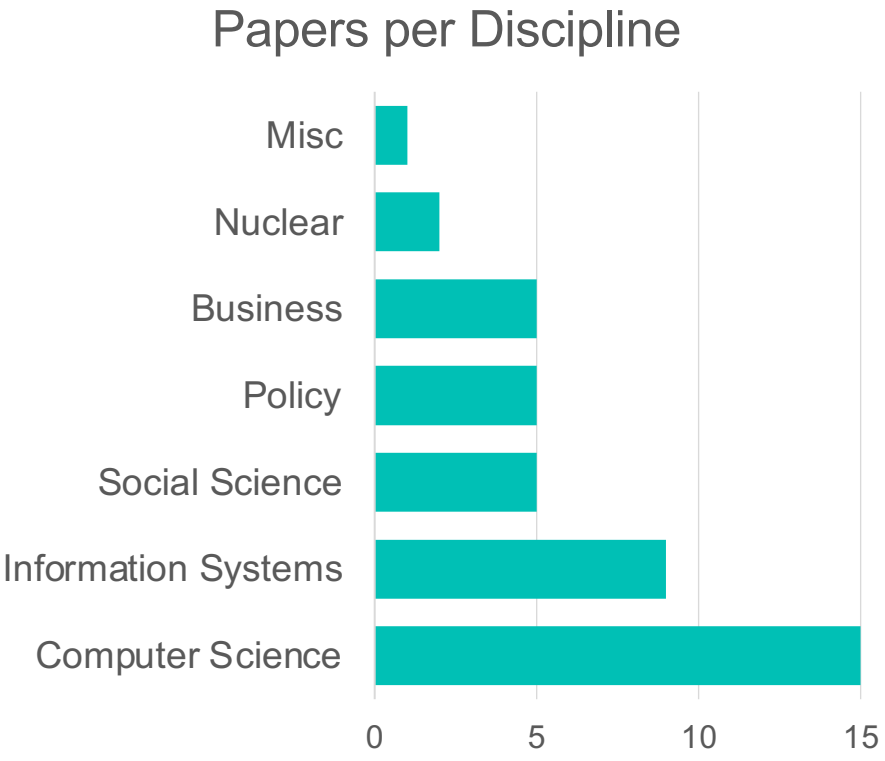
Systematic Review

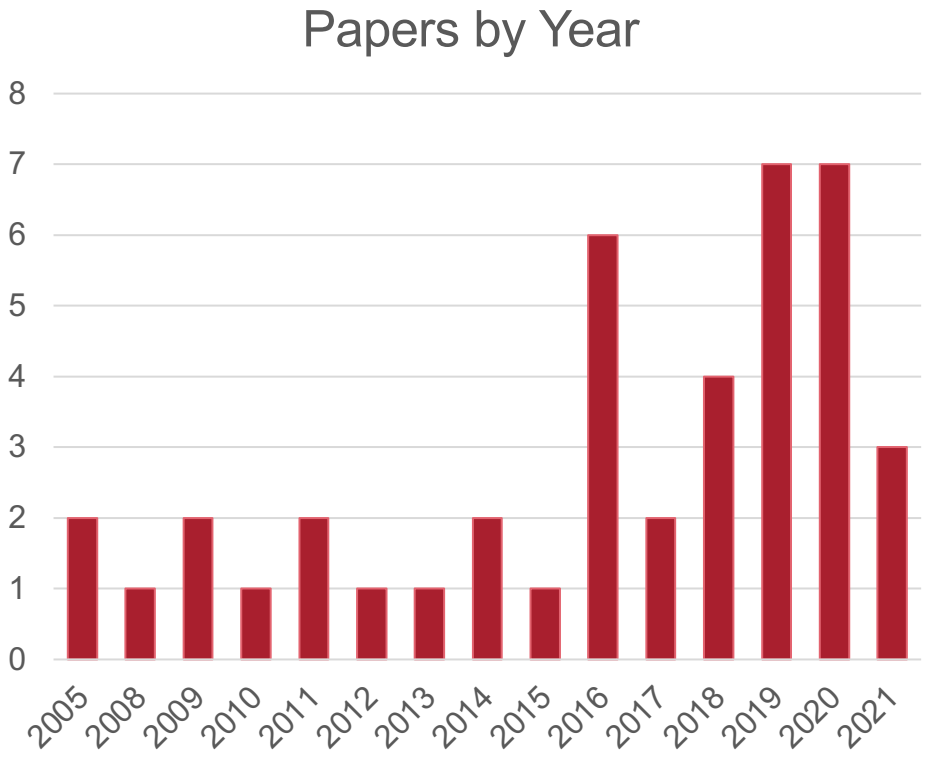
- Systematic Review of all papers that contain a **Taxonomy** of **Online Harm** or **Risk**

- Aims:
  - Understand the **dimensions of importance** in categorizing harms and risks online
  - Understand the **methodology** of how researchers have developed categorization frameworks
  - Understand differences between **disciplines** e.g., cybersecurity versus public policy

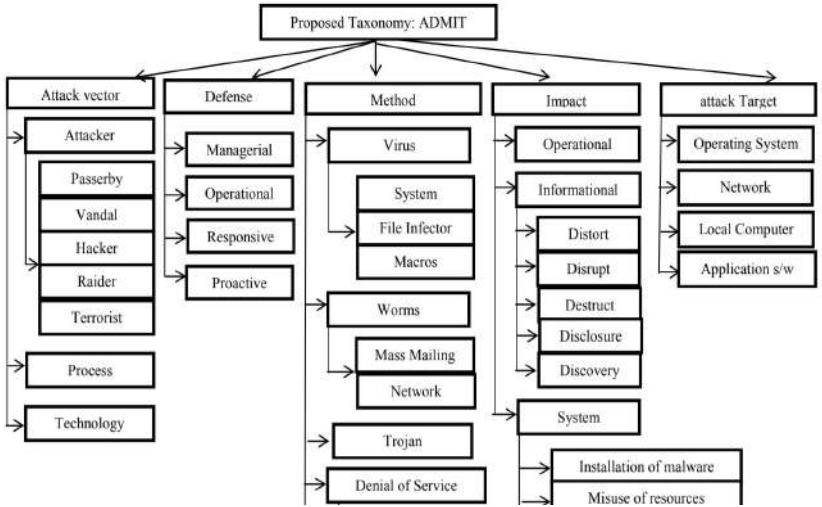


Descriptive Results





Cybersecurity vs cybersafety



Greater focus on the perpetrator and **vulnerabilities of the technology.**

igure 1: Taxonomy of risk of harm online and related offline activities (O'Connell and Bryce, 2005)

Greater focus on the **vulnerability of the victim** and how they interact with technology.

Original Aim

- Aimed to discriminate between risks and the harmful impacts of the risks

- However:
  - Difficulty assessing when 'actual harm' has been caused; some harms are anticipatory
  - The link between actions and harm caused are very rarely direct
  - Many 'nuances in the context' that may determine the possibility of harm
  - Does there need to be a victim for something to be harmful? The role of awareness in understanding harm
  - Subjective definitions of harm
    - Feinberg's harm principle states a harm is a 'setback to one's interests' (1984)
    - Harm is a normative judgment that reflects underlying social judgments about good and bad (Lin, 2006)

What about risk?

- Taxonomies demonstrate there is no ‘best practice’ way to categorise risk
- Research on risk is reactive and retrospective
- Anything could constitute a risk; surveillance could be risky, adoption of virtual reality technology could be risky, cryptocurrencies could be risky, using social media could be risky…

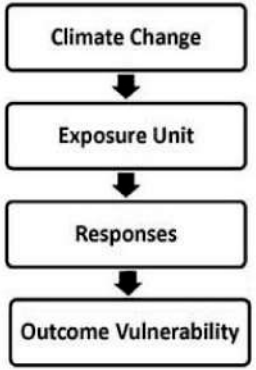
Vulnerability?

Vulnerability

- Not all 'risks' lead to a harm – and likelihood of harm being experienced varies across people/time/contexts/technologies
- A combination of factors *might* increase the likelihood that a harm is experienced as a consequence of an event…
  - Low trust in institutions makes a country vulnerable to misinformation
  - Shared passwords make a person vulnerable to data loss
- Mitigation works by addressing vulnerabilities at multiple levels/points of interaction

Vulnerability

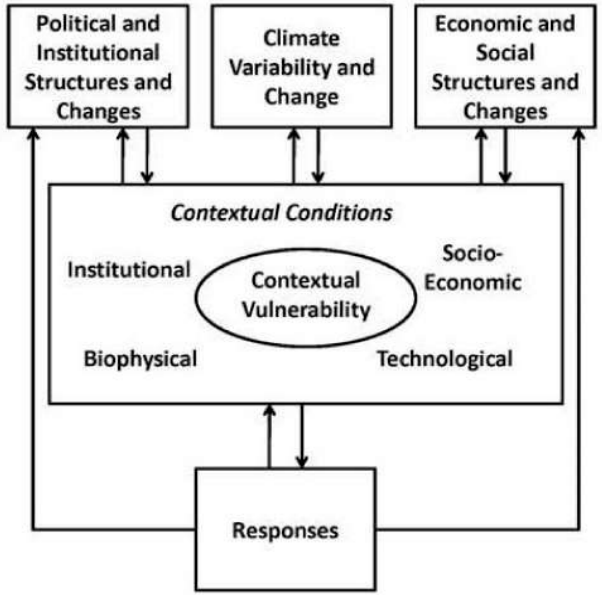
Two types of vulnerability: outcome and contextual (O'Brien et al, 2007).



'Outcome vulnerability is considered a linear result of the projected impacts of climate change on a particular exposure unit'

Current focus of much online harms research:
'risk-hazard' approach – attempting to understand the impact of a hazard based on exposure to it and sensitivity of the exposed entity (Turner et al., 2003).

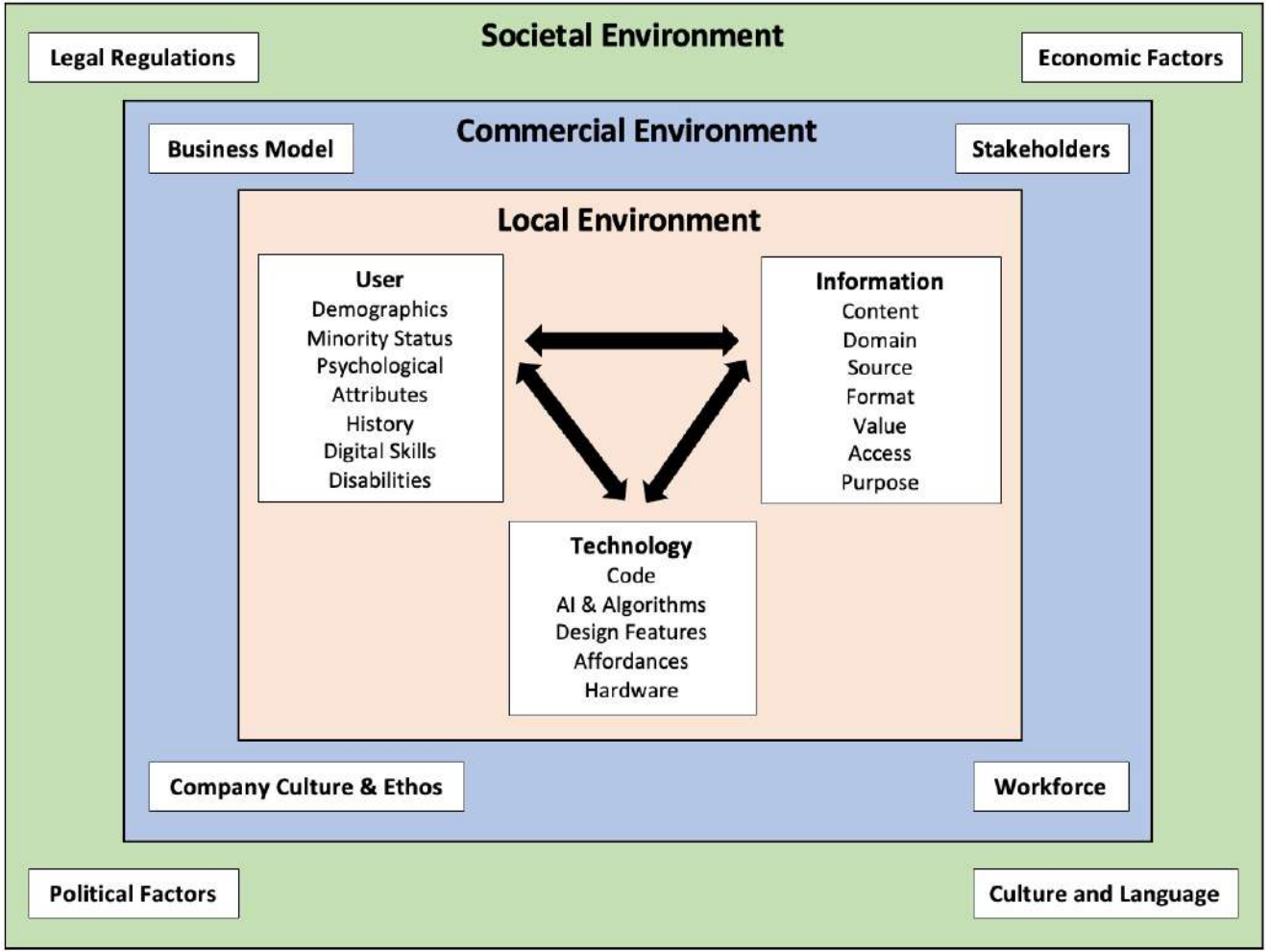
bath.ac.uk



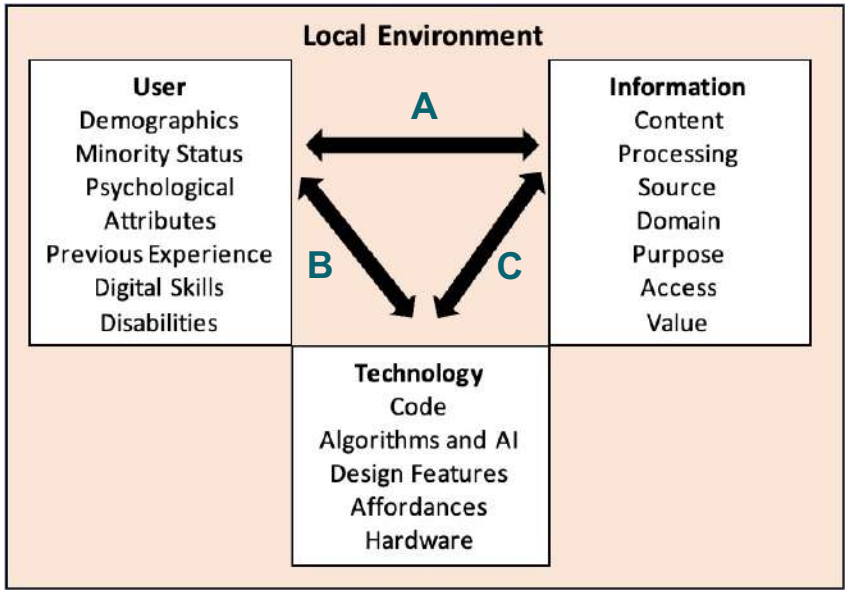
'Contextual vulnerability is based on a processual and multidimensional view of climate–society interactions. […] From this perspective, reducing vulnerability involves altering the context in which climate change occurs'

What we should be thinking about: modelling the online context and how this system of interacting components leads to contextual vulnerability.

Contextual Vulnerability Online



The Local Environment



Local Environment Interactions:

A: How do features of the user interact with features of the information in a way that creates vulnerability?

- User supplied information risks: *what data is being collected or produced by a user?*
- User access to information risks: *what content is being accessed or viewed by a user?*

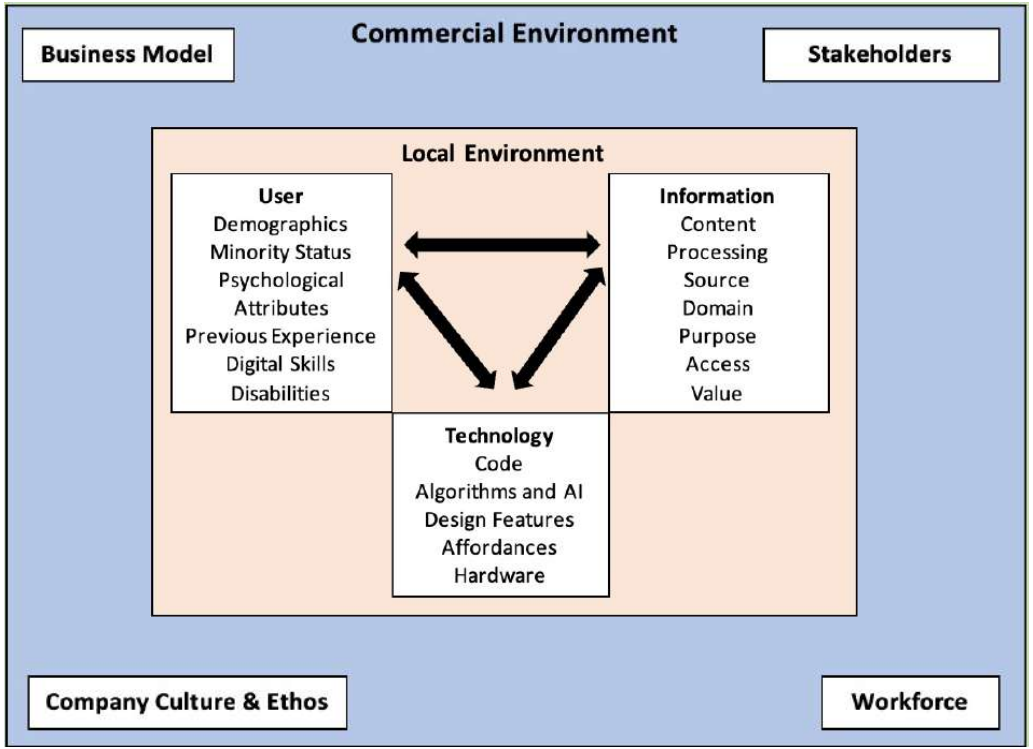
B: How do features of the user interact with features of the technology in a way that creates vulnerability?

- Decisional interference risks: *how is the technology impacting the user?*
- Improper usage risks: *how is the user using or exploiting the technology?*
- Accessibility risks: *Is the user unable to interact with the technology or features of the technology?*

C: How do features of the information interact with features of the technology in a way that creates vulnerability?

- Storage risks: *how is data being stored by the technology?*
- Processing risks: *how is data being processed by technology?*

The Commercial Environment

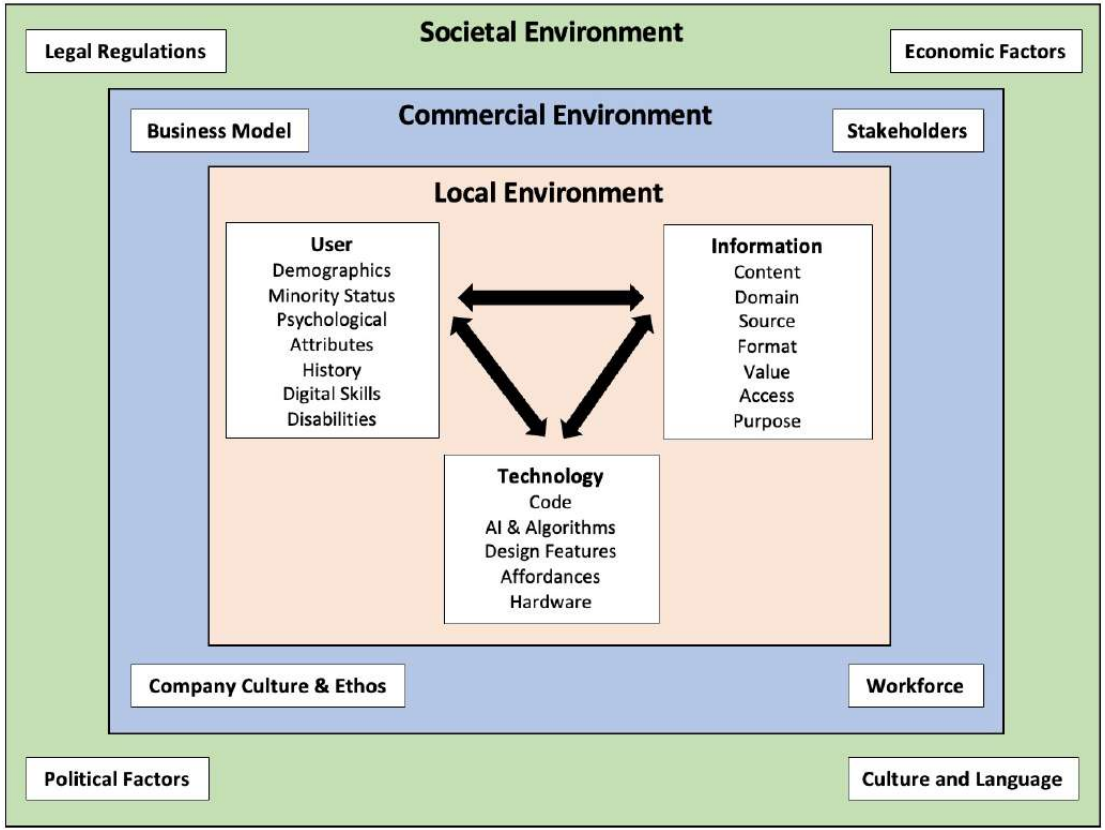


The local environment exists by virtue of the commercial environment.

Commercial factors may impact the individual factors of the local environment, or they may impact the interactions between them.

We can break each commercial factor down into its components, for example looking at the impact of each team in a company's workforce, or each revenue stream in the business model.

The Societal Environment

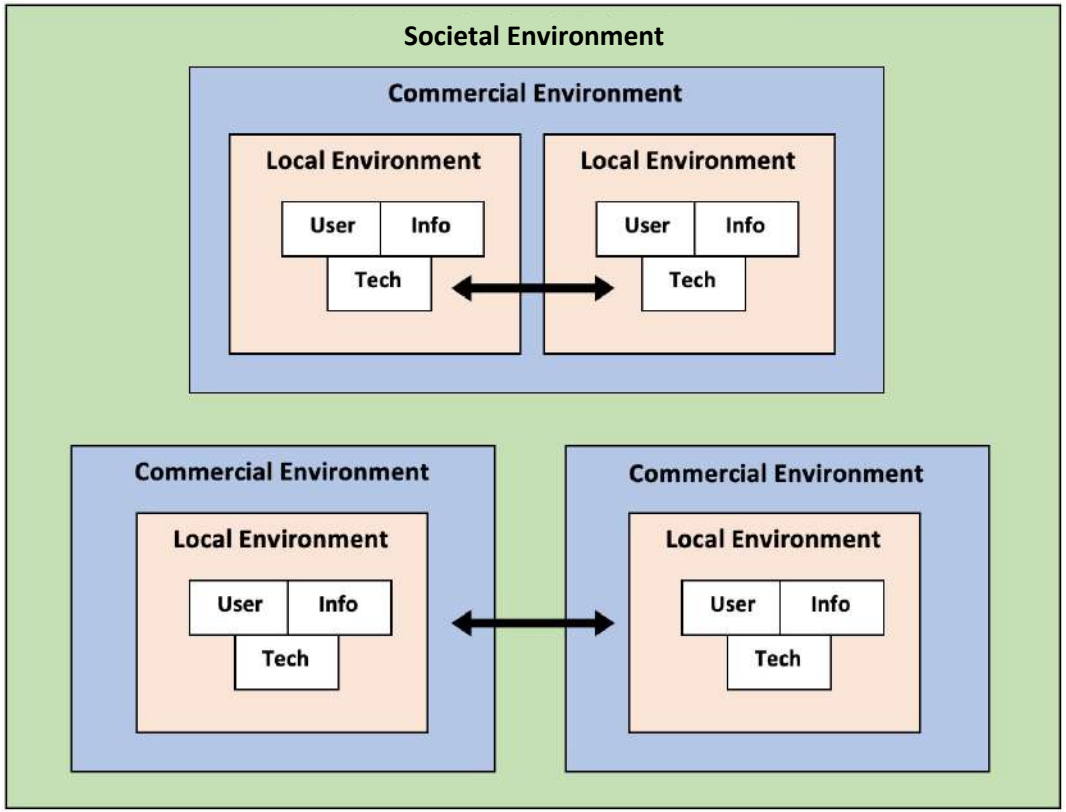


The commercial and local environments exist within an overarching societal context.

The societal context impacts the abilities and aims of the commercial environment, which in turn impacts the local environment.

There are also arguments that this relationship is also reversed: the local technological environment influences social and cultural aspects of a society (Postman, 1988).

Interacting commercial or local environments



There may also be non-hierarchical interactions between environments.

For example, the company Facebook (or Meta..) owns Oculus, Instagram, Whatsapp and Facebook. Vulnerabilities are likely to arise from the interaction between these local environments.

There may also be cross platform interactions; how does the local environment of Twitter interact with the local environment of TikTok? How does the interaction between commercial environments lead to vulnerability?

Purpose of this framework

- Identify the interactions between components in the context that may create or cause vulnerabilities
  - This allows us to pinpoint where mitigation strategies could be employed
- Illuminate the possible role of higher order factors, such as culture or legal regulations, in influencing the vulnerability of technological systems
- Anticipate future risks or vulnerabilities in new technologies
  - By defining the features of each environment, we can identify unforeseen risky interactions between system components.
- Move away from presuming direct causal relationships between risk and harm; applying a systems-level approach to the study of online harm.

National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online

REPHRAIN
Protecting citizens online

Thanks for listening.

Happy to take any questions, comments, criticism or compliments ☺

ac974@bath.ac.uk
aj266@bath.ac.uk
ls487@bath.ac.uk
dae30@bath.ac.uk
pssds@bath.ac.uk

@aliciagcork
@joinson
@lauragesmith
@davidaellis

National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online

REPHRAIN
Protecting citizens online

To learn more about REPHRAIN, our future plans and how to get involved:

 www.rephrain.ac.uk

 @REPHRAIN1

 rephrain-centre@bristol.ac.uk

We would love to hear from you. Thank you!