REPHRAIN

Protecting citizens online

REPHRAIN:

Scope of Research Challenges for the next stage of Protecting Citizens Online Research Programme

Marvin Ramokapane, Partha Das Chowdhury, Andrés Domínguez, Claudia Peersman and Awais Rashid

Version 1.1 - May 2021

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

Document History

Version	Date	DESCRIPTION AND COMMENTS
1.0	20/04/2021	Initial Document
1.1	07/05/2021	Changes after community review • Added: **"(e.g., unauthorised disclosure and identity theft), and ethical issues around profiling for targeted advertising, disinformation, among others [15, 23, 48, 38]"** in Page 10 Section 3.1 *"Understanding citizens' needs and empowering them in ever changing threat contexts"*. • Added: A new reference [38] titled **"To Believe or Not to Believe: an Epistemic Exploration of Fake News, Truth, and the Limits of Knowing"** to Page 10 Section 3.1 *"Understanding citizens' needs and empowering them in ever changing threat contexts"*. • Added: **"(i.e., age, behaviour, dis/ability, class, gender, race, and socio-economic status)"** in Page 10 Section 3.1 *"Understanding citizens' needs and empowering them in ever changing threat contexts"*. • Added: **"(including primary users and other stakeholders, e.g., those with safeguarding responsibilities)"** in Page 10 Section 3.1 first challenge "Usable harm mitigation tools". • Added: **"(including evolving threats)"** in Page 11 Section 3.1 "Novel tools and methods for alerting users about imminent and potential harm". • Added: **"(and between data controllers as data is shared or transferred across controller boundaries)"** in Page 12 in Section 3.2 second challenge "Maximising users' agency through effective mechanisms for data transparency, traceability, assurance and portability". • Added: **"(e.g., 'legitimate interest' is ill-defined and it is not readily possible for the user to ascertain the origins or future trajectories of the data)."** in Page 13 Section 3.2 in the challenge "Innovative methods for informed consent in seamless human-technological interactions in shared spaces". • Added: **"(users are key stakeholders in this regard)"** in Page 13 in the first paragraph of Section 3.3 *"What does good or balanced look like?"*.

REPHRAIN: Scope of Research Challenges for the next stage of Protecting Citizens Online Research Programme

Version 1.1

May 11, 2021

1 Introduction

1.1 REPHRAIN

The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) is a UKRI-funded research centre that forms part of the wider UKRI Strategic Priority Fund programme on protecting citizens online. It brings together the UK's substantial academic, industry, policy and third sector capabilities to address the current tensions and imbalances between the substantial benefits to be gained by full participation in the digital economy and the potential for harm through loss of privacy, insecurity, disinformation and a myriad of other online harms. REPHRAIN's research is rooted in an ethos of interdisciplinary research – alongside principles of responsible innovation and creative engagement – to develop new insights that allow the socio-economic benefits of a digital economy to be maximised whilst minimising the online harms that emerge.

REPHRAIN was launched on the 1st of October 2020 and involves an initial consortium of five academic institutions: University of Bristol, University of Bath, University of Edinburgh, University College London and King's College London as well as expertise and support from 23 non-academic organisations from industry, government, policy, law enforcement and third sectors. REPHRAIN's researchers address issues of privacy, online harms and adversarial influence from a range of disciplinary perspectives including: Computer Science, International Relations, Law, Psychology, Management, Design, Digital Humanities, Public Policy, Political Science, Criminology and Sociology.

The work of REPHRAIN focuses on three core missions:

Mission 1 emphasises the requirement to deliver privacy at scale whilst mitigating its misuse to inflict harms. It aims to reconcile the tension between data privacy and lawful expectations of transparency by not only drawing heavily on advances in privacy-enhancing technologies (PETs), but also leveraging the full range of socio-technical approaches to rethink how we can best address potential trade-offs.

Mission 2 emphasises the need to minimise harms whilst maximising the benefits from a sharing-driven digital economy, redressing citizens' rights in transactions in the data-driven economic model by transforming the narrative from privacy as confidentiality only to also include agency, control, transparency and ethical and social values.

Mission 3 focuses on addressing the balance between individual agency and social good, developing a rigorous understanding of what privacy represents for different sectors and groups in society (including those hard to reach), the different online harms to which they may be exposed, and the cultural and societal nuances impacting effectiveness of harm-reduction approaches in practice.

An initial set of 25 inaugural projects are addressing various research challenges to deliver innovative research outcomes towards these missions.

These missions are supported by four engagement and impact objectives that represent core pillars of REPHRAIN's approach: (1) design and engagement; (2) adoption and adoptability; (3) responsible, inclusive and ethical innovation; and (4) policy and regulation. Combined, these objectives will deliver co-production, co-creation and impact at scale across academia, industry, policy and the third sector.

1.2 The Scope Document

REPHRAIN is the first phase of the UKRI Strategic Priority Fund programme on Protecting Citizens Online. A second phase is planned to fund further research nodes connected to REPHRAIN and addressing research topics that complement its research activities and contribute towards the three missions. Furthermore, REPHRAIN also has a *Capability Fund* that will be funding further research projects on relevant topics through *Strategic Funding Calls*. Therefore, since its launch, the Centre team has been engaging in a series of scoping activities – involving consultation workshops as well as literature reviews – with two key aims:

- To develop the initial Map of privacy, online harms and adversarial influence online – an online resource (to be regularly updated) that provides stakeholders with a shared understanding of the landscape and establishes a baseline of current state-of-the-art;
- To identify priority areas for the first REPHRAIN Strategic Funding Call and provide input to the UKRI call on Protecting Citizens Online for the next stage research projects related to REPHRAIN.

1.3 Community consultation on the REPHRAIN Scope Document

The purpose of this Scope document is to seek community input and feedback on the topics identified as input to the UKRI call on Protecting Citizens Online projects. Topics that were deemed suitable for the REPHRAIN Strategic Funding call have not been included here. Details on those topics will be included in the REPHRAIN Strategic Funding Call to be launched on 19 April 2021.

We, therefore, invite stakeholders from academia, industry, government, policy, law enforcement and third sector organisations to comment on the Scope.

In the same fashion as the development of the Scope is based on community consultations, we are keen to iterate and update it based on community review and feedback. Comments should, therefore, be constructive. If in doubt, colleagues commenting should consider the type of constructive comments that would help improve a paper submitted to a peer reviewed conference or journal paper.

Comments should cover the following key questions:

Question 1: What are the strengths of the current set of topics? Examples responses include but are not limited to: coverage of key research challenges; interdisciplinary nature of the problems being considered; transformative nature of potential research outcomes. In a nutshell, we are interested in understanding what you appreciate about the proposed topics.

Question 2: What are the improvements that can be made to the current scope? Example responses include but are not limited to: comments on coverage of key research challenges; approaching the problem from a different disciplinary standpoint; improving relevance to long term needs of society. However, comments mustn't simply state that something is imperfect. We would like to receive proposals on how to improve upon any issues highlighted. Proposals may include suggested rewordings, additional topic areas to be covered and so on. However, suggestions for additional topics must include a core reference that provides a synthesis of state-of-the-art and justifies the inclusion of the topic (e.g., through a peer-reviewed paper that provides a systematic literature review). In sum, we are interested in understanding how the topics may be improved through specific updates.

Question 3: Are there other comments not covered by Questions 1 and 2? We welcome other comments on the Scope document as colleagues may see fit. Similar to questions 1 and 2, we request that these are constructive and make concrete proposals for improvement.

Comments can be sent by email to rephrain-centre@bristol.ac.uk either as a free form text or an annotated PDF document. Comments can also be submitted online.

Given the timescales for the UKRI call, all comments must be received by 17:00 BST on 04 May 2021. The REPHRAIN team will give all comments full consideration and update the Scope as suitable.

2 Scoping research methodology

The centre team conducted scoping workshops with various stakeholder communities: academia, industry/practice and policy. The purpose of these workshops was to elicit views of various communities on the current research problems. This was complemented by a literature review to establish the current state-of-the-art and identify where the gaps in the literature align with stakeholder input from the workshops. The detailed methodology is discussed below.

2.1 Literature review

We collected and reviewed key publications providing insights on privacy, online harms and adversarial influence as well as those discussing state-of-the-art methods and tools used to detect, investigate and reduce online harms. Our target paper selection focused on the last decade to ensure that we maintained focused on up to date research advances as well as gaps in state-of-the-art. However, exceptions were made for classical papers published before the last decade where they provided insights into underpinning issues or longstanding research problems in this area. The initial paper collection identified a broad range of papers based on title and abstract. This selection was narrowed down through further reading of key sections from the papers within the corpus to ensure relevance to REPHRAIN's missions. Additional paper collection was performed after the scoping workshop to further understand issues raised in the workshops. In the end, our scoping phase draws upon 67 papers in total from 37 venues/journals.

2.2 Scoping workshops

2.2.1 Launch event discussion sessions

As part of the Centre launch event on 29 October 2020 – attended by 40 attendees from 21 academic and 19 non-academic organisations – the Centre team organised a set of discussion sessions to inform the design of the scoping work. Participants of these discussions were asked about the types of online harms they expected the centre to address, methods and tools for reducing online harm, data sharing issues, and important artefacts that ought to form part of the REPHRAIN Toolbox—synthesising outcomes from the various projects within the Centre. The insights from these discussions were used to identify the questions to be explored within the scoping workshops.

Table 1: Topics explored with participant group in the workshops

Participant Group	Topics Explored	Rationale
Academia	• Prevalent online harms • Impact of online harms on individuals and communities • Approaches and tools to mitigate harm • Key research gaps	• To understand the online harm research landscape, identify key research works, and research gaps in the area.
Industry	• Prevalent online harms • Approaches and tools to mitigate online harms • Technical challenges around tackling online • Challenges around regulations	• To understand the online harms from an industry view point and the state-of-practice, approaches and tools to reduce harm. To also understand the challenges introduced by having to comply with regulations.
Law enforcement & Policy makers	• Prevalent online harms • Approaches and tools to detect, reduce, and disrupt online harms • Challenges around law enforcement	• To understand trending online harms from law enforcement viewpoint, state-of-practice tools and approaches, their strengths and weaknesses. Also, to identify gaps in the regulatory landscape.

2.2.2 Workshop design

The scoping workshops aimed to engage experts from academia, industry, law enforcement, policymakers/regulators, and safety tech developers. The main reason behind this recruitment strategy was to gain insights from various backgrounds and identify where the tensions and the trade-offs exist. The workshop script focused on identifying the key resources in the area, developing an understanding of the area of harm on which participants were working, the current state-of-the-practice tools and methods in the area, regulations, and the expectations of the Centre's role in reducing online harm. While the workshops aimed at filling these gaps, each workshop was tailored to each particular field, as shown in Table 1.

2.2.3 Recruitment and invitations

Workshop invitations were extended to different experts and community members from the centre's established networks and communication channels. All interested participants were required to complete and return the consent form to participate in the scoping workshops[1]. 54 participants responded, but only 40 could attend the workshops. There was a largely even distribution of representation from academia and non-academia. Table 2 shows the number of sessions and participants who took part in the sessions.

Table 2: Workshop sessions and number of participants.

Participant Group	No. of workshop sessions	No. of participants
Academia	3	21
Industry	4	10
Law enforcement & Policy makers	2	9

2.2.4 Procedure

Each workshop was held online using Microsoft Teams and scheduled for approx. 2.5 hrs with a 30 min break in between. Each session was facilitated by at least two members of the research team. Eight sessions were audio-recorded, while two sessions were not due to some participants declining the option. Two further members of the team transcribed the sessions where recordings were not

[1]In line with REPHRAIN research procedures on ethics, the Scoping research was reviewed and approved by University of Bristol Faculty of Engineering Research Ethics Committee and the REPHRAIN Ethics Board.

possible. As part of the workshop, participants had access to an online whiteboard application, *Padlet*[2], to note and share important resources. Workshop sessions ran from the 5th of February to the 3rd of March 2021.

2.2.5 Data analysis

After audio recordings were transcribed, three researchers started the coding process to create a codebook (i.e., Thematic Analysis [13]). To develop the initial codebook, the researchers individually coded three transcripts and then consolidated the most salient codes and themes into a shared codebook. This process aimed to identify the relationship between codes and themes, noting and grouping similar concepts. This process initially generated more than 300 codes and 31 different themes. The researchers then employed an *arguing to consensus* technique [27] to resolve discrepancies between codes and themes. In the end, the number of broader themes were reduced to 5 with 16 sub-themes. The final themes and sub-themes were clearly defined and described with examples where applicable in the final codebook. The researchers then used the codebook to code the rest of the transcripts. In a scenario where there was a disagreement or an identification of a new code or theme of interest, the researchers discussed it further and identified where it could be grouped.

2.2.6 Workshop findings and literature review synthesis

Following the thematic analysis, the centre team discussed and identified themes that reflected potential research challenges to be included in the funding calls. For each theme, the researchers further merged similar challenges and split up broader challenges into more fine-grained ones. Then, the three researchers contrasted these challenges with those identified from the literature. In the case where the challenge was missing from the collected literature, the researchers undertook a further literature search to confirm if the challenge was just missed in the initial search or if it was not recognised in literature. Moreover, if the challenge was not recognised by existing literature, the researchers identified publications that highlighted existing work or future paths of inquiry to identify a clear justification of why the challenge needs to be explored. Additional challenges from the literature review which were not covered in the codebook were also discussed and put under relevant themes as appropriate.

Finally, the set of topics to be considered for the UKRI call were separated from those to be prioritised in the REPHRAIN Strategic Funding call.

3 Proposed Topics for the UKRI Call

Our analysis identified three overarching research themes, each with three interdisciplinary challenges to be addressed (see Table 3). After identifying these themes, we also analysed how they relate to the current REPHRAIN missions

[2] www.padlet.com

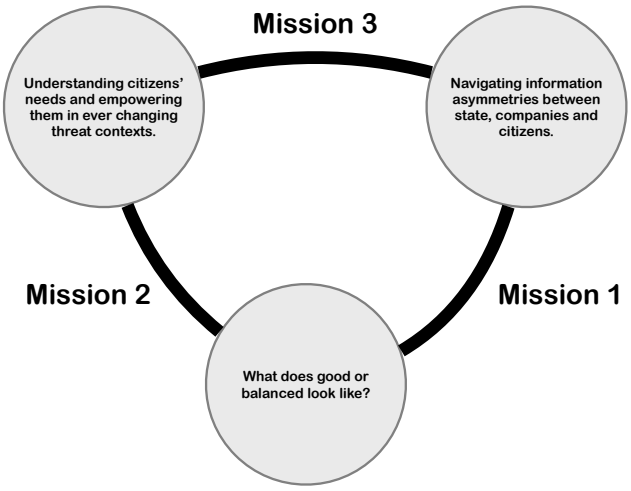

Figure 1: Themes from the scoping research and their relationship to REPHRAIN missions

Table 3: Overarching research themes and interdisciplinary challenges

Overarching Research Theme	**Research Challenges**
Understanding citizens' needs and empowering them in ever changing threat contexts.	- Usable harm mitigation tools. - Novel tools and methods for alerting users about imminent and potential harm. - User-oriented tools to prevent harm through impersonation.
Navigating information asymmetries between state, companies and citizens.	- Methods, tools and techniques for responsible data usage and dissemination. - Maximising users' agency through effective mechanisms for data transparency, traceability, assurance and portability. - Innovative methods for informed consent in seamless human-technological interactions in shared spaces.
What does *good* or *balanced* look like?	- Balancing privilege management and anonymous credentials. - Forensic readiness without violating data privacy rights. - Designing and validating metrics for assessing the effectiveness of harm mitigation mechanisms.

and would add complementary strengths to the existing programme of research on Protecting Citizens Online. The relationships between the themes from the scoping research and the REPHRAIN missions are shown in Figure 1.

We next discuss each theme and its constituent challenges.

3.1 Understanding citizens' needs and empowering them in ever changing threat contexts

A common assumption is that citizens are completely informed and technically capable before giving and/or withdrawing consent when engaging in online activities. This is further complicated by the need to trust data controllers, which, in turn, leads to insecurities (e.g., unauthorised disclosure and identity theft), and ethical issues around profiling for targeted advertising, disinformation among others [15, 23, 48, 38]. Those aiming to exploit users' data or inflict harms regularly adapt their approaches as privacy and harm reduction measures are put in place. Such a highly dynamic threat landscape calls for novel interdisciplinary approaches, technologies and methods that consider the diverse realities (i.e., age, behaviour, dis/ability, class, gender, race, and socioeconomic status) and changing needs of citizens with regards to privacy and harm migitation. Three fundamental interdisciplinary research challenges need to be addressed.

- **Usable harm mitigation tools.** Existing studies [30, 42] concerning users have identified lack of usability and inclusion of users in the design process as some of the main reasons why users fail to protect themselves or use privacy enhancing technologies. Prior works argue that most systems are not designed with the user in mind, system designers often assume users are capable of understanding and using systems or the privacy measures in place [3]. While growing literature has explored ways in which systems or security measures can be made usable, introduction of new technologies, tools, and mechanisms still poses threats to users' privacy. Producing knowledge about – and in cooperation with – users, the challenges they face, their coping strategies, and other dynamics of citizens and digital technologies intersecting with each other is crucial for the development and adoption of usable PETs and empowering users online.

  A related key challenge is understanding adversarial threats the way users understand them. Existing adversarial models are often theoretical and do not match actual privacy requirements of real-world applications [41]. This often results in solutions that do not address the needs of real users and their real problems. There is a need to focus and include users in the design of harm mitigation tools. Solutions should therefore build on a nuanced up-to-date understanding of users and non-users (including traditionally excluded groups), their complex realities and changing behaviours in response to technological innovations. Interdisciplinary research spanning usability, psychology, HCI, design and beyond is needed to enrich our understanding and develop novel *usable* harm mitigation tools. Moreover, diverse methodological perspectives can help to take into account diverse actors (including primary users and other stakeholders, e.g., those with safeguarding responsibilities), and empower people to protect themselves from online harm and participate safely online.

- **Novel tools and methods for alerting users about imminent and**

potential harm. The development of underground forums [7] with information about users like their passwords, credit cards, images and other details is a cause for concern both among the law enforcement and citizens. A related and very problematic issue is of sharing images without the knowledge and consent of the victim. Disinformation, hate speech are on the rise with fatal consequences. Recent research points to the ease with which potential perpetrators can start operating in the underground market [11]. However, there are no easy methods for users to ascertain if their information is being used/sold in the underground markets without their knowledge and consent. Nor do they have, at their disposal, effective methods, tools and techniques to evaluate the potential risks (including evolving threats) and take mitigating actions. Novel tools and methods are needed to alert potential victims when and if their sensitive information is involved in breaches or underground data markets and how to effectively – without expending substantial cost in time or other resources – mitigate the harms that may result.

- **User-oriented tools to prevent harm through impersonation.** A common instance of harm on the internet is when perpetrators gain trust of victims and exploit them through impersonation [24]. Impersonation has a clear link to damage reputations and is being used in many instances of intimate partner violence [48], sexual violence [22], child grooming [37], and revenge pornography[3]. Impersonation for financial profit is addressed through institutional approaches [16] to detect suspicious emails, links or contents. Recent research also proposes deep learning based approaches to identify suspicious web sites [33]. However, user-oriented tools are lacking in this space. Given the fast proliferation of adversarial capabilities, there is a need to develop usable and scalable tools to support citizens in assurance of communicating entities and content on the internet.

3.2 Navigating information asymmetries between state, companies and citizens

Citizens engage on a daily basis in data transactions with diverse platforms, data infrastructures, public institutions and intermediaries. Much of the underlying mechanisms and data flows taking place in cyberspace are distinctly opaque to average users making them vulnerable to unfair and illegal practices, or their personal data being unknowingly exploited [26, 14]. Moreover, vast information and power asymmetries make it difficult or too costly for individuals to make informed decisions about their data transactions and online behaviour [39]. In the context of a growing digital economy, there is a need for mechanisms, tools and methods that improve the transparency, traceability and usability of people's digital traces and enable citizens to have agency over how their data is being used, monetised or shared with third parties. Three fundamental interdisciplinary research challenges need to be addressed.

[3] https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/

- **Methods, tools and techniques for responsible data usage and dissemination.** Respect for individual privacy is a socially agreed value within specific contexts [4]. However, individuals exist as part of social groups with mutually agreed values and common interests (e.g., medical research) which may require users to share their personal data for collective benefit. Nevertheless, this may lead to privacy violations [20, 1, 44], and in turn, good privacy protections may render data unusable [40, 2]. Recent research [5] argues for a nuanced understanding of privacy or lack of it while others advance the understanding of privacy-utility trade-offs of data [2]. An often related issue is of accessing victim information repeatedly which might cause repeated trauma for victims; due care in such instances goes well beyond privacy violations. Consequently, there is a need for frameworks [43], methods, and tools for responsible data usage and dissemination [34] to address the tensions between social good and privacy.

- **Maximising users' agency through effective mechanisms for data transparency, traceability, assurance and portability.**

  The current technological landscape is highly complex, inhabited by diverse and disparate platforms, technical architectures and data-driven business models. In such a context, it is very challenging for citizens to find out how their data is handled and used by data controllers. The opacity with which digital platforms operate makes it challenging even for technically knowledgable users to manage their privacy and identify unfair and harmful practices by platforms [46]. Moreover, making computational systems more open and transparent does not directly equate to them being accountable [8]. Novel provisions such as the right for data portability or right to access have been enshrined in GDPR as a means to give people more control over their data and draw benefits from the digital economy. Moreover, it has been argued that effective data portability could also foster competition between online platforms and a healthier digital ecosystem [17, 36]. Yet, some of these rights remain difficult to implement in practice due, among other factors, to lack of robust definitions, lack of skills and tools for enforcement, and little economic incentives by big technology companies to open up their systems and enable interoperability with competitors [50]. There is a need for actionable mechanisms and effective tools to provide citizens with more agency over their personal data, trace and understand their digital footprints and verify that the contractual obligations between data subjects and data controllers (and between data controllers as data is shared or transferred across controller boundaries) are observed [6].

- **Innovative methods for informed consent in seamless human-technological interactions in shared spaces.** Rapid changes in technology and the regulations that govern it has led to more complex systems. For example, the introduction of GDPR forced a lot of service providers to

change the way they sourced consent from users. However, most of these web consent mechanisms are not usable and do not provide sufficient information to help users give informed consent [29, 31] (e.g., "legitimate interest" is ill-defined and it is not readily possible for the user to ascertain the origins or future trajectories of the data). One key challenge is delivering and sourcing informed consent where interactions with devices and technologies are increasingly implicit, e.g., where a user with wearables catalyses dynamic interactions amongst a range of surrounding devices and services. State-of-the-art technologies depend on screens and privacy policies to help users make informed consent; the limitations are well-researched and well-understood [9, 18]. Moreover, this challenge is further compounded where groups or shared spaces are concerned [53]. Interdisciplinary research is required to develop novel socio-technical mechanisms for informed consent in such seamless interactions that often involve shared devices and/or spaces.

3.3 What does *good* or *balanced* look like?

The online world is riddled with conflicting interests, ongoing tensions and ethical dilemmas which are difficult to resolve. In the digital age, individuals often find themselves trapped between the drive for technological innovation and the need for adequate legal protections from online harm. Similarly, in certain situations, there may be a legitimate requirement to override the privacy of individuals to aid law enforcement or to fulfil a duty of care. Existing norms and regulations that guarantee rights and obligations sometimes fail to adequately address the negative side effects of emerging digital technologies without stifling innovation. Algorithmic technology and data-centric applications are prominently at the forefront of academic research both for their benefits as well as the ethical concerns that they raise. The question of what is "balanced" or "good" requires reasoned dialogue between various stakeholders (users are key stakeholders in this regard) about what are the acceptable trade-offs, compromises, and consequences, and how these may reflect in the design of systems and platforms as well as regulatory and policy interventions. This theme focuses on three key research challenges in this regard.

- **Balancing privilege management and anonymous credentials** A privilege management infrastructure (PMI) forms the basis of granting user access or prevent unauthorised users from accessing goods and services; with authentication, authorisation and audit being the principal components of a PMI. Service providers verify user credentials to authenticate and grant users access. The regulatory environment too requires the use of credentials to prevent money laundering and other identity based crimes [21]. The verification of identity attributes is also becoming increasingly important for individuals in a documented society. Refusing to share attribute information might lead to discrimination, exclusion or

disadvantage [45]. However, repeated use of credentials and collection of more data may render users vulnerable to identity theft, unauthorised disclosure, targeted campaigns [54], and other frauds [28]. The threats may come from both the service provider and the third-party that issues such credentials [52]. However, anonymity systems require users to rely on third parties to resolve a pseudonym to a long term stable identity [10]. Users may not be comfortable with external entities mediating the verification. A mediator might have access to information about a user which a user might not possess [25]. A mediator perusing the identity or attribute information of a user without their knowledge may lead to unauthorised disclosure [41]. Novel advances in PETs are required to facilitate anonymous privilege management without compromising the legitimate needs of businesses, regulators and law enforcement.

- **Forensic readiness without violating data privacy rights.** While tools for preventing inappropriate and harmful uses of the digital ecosystem are important, effective apprehension of cybercrime is equally critical. The key to this is irrefutable evidence in a court of law. However, a recent report[4] suggests that conviction rates are low owing partly to insufficient evidence. The ability to generate evidence of quality to stand judicial scrutiny is difficult on digital devices and platforms. They are highly-connected and complex with ever changing state information, which, in turn, affects the integrity of the evidence [51]. Moreover, inbuilt software is usually not engineered to be useful for gathering evidence [35]. A recent white paper from REPHRAIN advances that data science based approaches can complement existing investigation capabilities[5]. Research at the boundary of technology, law, ethics, human rights and criminology is needed to produce tools that will make devices sensitive to legitimate forensic needs without violating the data privacy rights of the subjects.

- **Designing and validating metrics for assessing the effectiveness of harm mitigation mechanisms.** There is a growing public concern on the perils of behavioural data collections and inference mechanisms used by social media platforms [19]. The ability to gather personal sensitive information increases exponentially in the era of pervasive computing [32]. The initiatives to protect citizens on the internet range from community driven[6] to regulatory oversight through ICOs and legislations[7]. There are platform driven efforts as well, for example, the bullying filter of Instagram[8]. A pertinent question is: are these interventions effective? And how to measure effectiveness? There are initiatives to formulate metrics to ascertain the effectiveness of technical protection mechanisms against

[4] https://publications.parliament.uk/pa/cm201719/cmselect/cmhaff/515/51507.htm

[5] https://cpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2021/03/White-Paper-Towards-Data-Scientific-Investigations.pdf

[6] https://ec.europa.eu/digital-single-market/en/safer-internet-day-sid

[7] GDPR, CCPA

[8] https://www.nytimes.com/2018/05/01/technology/instagram-bully-filter.html

online harm, with a case for significant improvement upon the existing efforts. For example, recent research in pervasive computing indicates the nuances involved [47]. Moreover, in the domain of hate speech there is a need for rigorously assessed interventions [12]. The assessment of interventions needs to span from regulatory [49], social awareness[9] as well as technical approaches. There remains a critical need to develop context-specific metrics that are empirically validated in order to provide sound evaluating mechanisms for mitigating particular online harms.

References

[1] John M. Abowd. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery.

[2] John M. Abowd and Ian M. Schmutte. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, 109(1):171–202, January 2019.

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman M. Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys (CSUR)*, 50(3):44:1–44:41, 2017.

[4] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and Human Behavior in the Age of Information. *Science*, 347(6221):509–514, 2015.

[5] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Secrets and Likes: The Drive for Privacy and the Difficulty of Achieving it in the Digital Age. *Journal of Consumer Psychology*, 2021.

[6] Fatemeh Alizadeh, Timo Jakobi, Jens Boldt, and Gunnar Stevens. GDPR-Reality Check on the Right to Access Data: Claiming and Investigating Personally Identifiable Data from Companies. In *Proceedings of Mensch Und Computer 2019*, MuC'19, page 811–814, New York, NY, USA, 2019. Association for Computing Machinery.

[7] L. Allodi, M. Corradin, and F. Massacci. Then and Now: On the Maturity Of The Cybercrime Markets The Lesson That Black-Hat Marketeers Learned. *IEEE Transactions on Emerging Topics in Computing*, 4(1):35–46, 2016.

[9] `https://www.thinkmind.org/index.php?view=article&articleid=cyber_2018_6_20_80051`

[8] Mike Ananny and Kate Crawford. Seeing Without Knowing: Limitations Of The Transparency Ideal And Its Application To Algorithmic Accountability. *New Media & Society*, 20(3):973–989. Publisher: SAGE Publications.

[9] Pauline Anthonysamy, Phil Greenwood, and Awais Rashid. Social Networking Privacy: Understanding The Disconnect From Policy to Controls. *Computer*, 46(6):60–67, 2013.

[10] M. R. Asghar, M. Backes, and M. Simeonovski. PRIMA: Privacy-Preserving Identity And Access Management At Internet-Scale. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018.

[11] R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, and D. McCoy. Mapping The Underground: Supervised Discovery Of Cybercrime Supply Chains. In *IEEE 2019 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–16, 2019.

[12] Catherine Blaya. Cyberhate: A Review And Content Analysis Of Intervention Strategies. *Aggression and Violent Behavior*, 45:163–172, 2019. Bullying and Cyberbullying: Protective Factors and Effective Interventions.

[13] Virginia Braun and Victoria Clarke. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[14] Jenna Burrell. How The Machine 'Thinks': Understanding Opacity In Machine Learning Algorithms. *Big Data & Society*, 3(1):205395171562251, 2016.

[15] Ryan Calo. The Boundaries Of Privacy Harm Essay. *Indiana Law Journal*, 86:1131, 2011.

[16] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar. SoK: A Comprehensive Rexamination of Phishing Research from the Security Perspective. *IEEE Communications Surveys Tutorials*, 22(1):671–708, 2020.

[17] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. The Right To Data Portability In The GDPR: Towards User-Centric Interoperability of Digital Services. *Computer Law & Security Review*, 34(2):193–203, 2018.

[18] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-Scale Readability Analysis of Privacy Policies. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 18–25. ACM, 2017.

[19] Casey Fiesler and Blake Hallinan. "We Are The Product": Public Reactions to Online Data Sharing and Privacy Controversies In The Media. New York, NY, USA, 2018. Association for Computing Machinery.

[20] Ferdinando Fioretto, Pascal Van Hentenryck, and Keyu Zhu. Differential Privacy of Hierarchical Census Data: An Optimization Approach. *Artificial Intelligence*, 296:103475, 2021.

[21] Michelle Frasher and Brian Agnew. Multinational Banking and Conflicts Among US-EU AML/CTF Compliance & Privacy Law: Operational & Political Views In Context. *SWIFT Institute Working Paper No. 2014-008*, 2016.

[22] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[23] Melanie Freeze, Mary Baumgartner, Peter Bruno, Jacob R. Gunderson, Joshua Olin, Morgan Quinn Ross, and Justine Szafran. Fake Claims of Fake News: Political Misinformation, Warnings, And The Tainted Truth Effect. *Political Behavior*.

[24] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA, 2017. Association for Computing Machinery.

[25] Yuri Gurevich, Efim Hudis, and Jeannette M. Wing. Inverse Privacy. *Commun. ACM*, 59(7):38–42, June 2016.

[26] Lucas D. Introna. Algorithms, Governance, And Governmentality: On Governing Academic Writing. 41(1):17–49. Publisher: Sage Publications, Inc.

[27] Barbara Johnstone. *Discourse Analysis*. John Wiley & Sons, 2017.

[28] Fujun Lai, Dahui Li, and Chang-Tseh Hsieh. Fighting Identity Theft: The Coping Perspective. *Decision Support Systems*, 52(2):353–363, 2012.

[29] Dominique Machuletz and Rainer Böhme. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs After GDPR. *Proc. Priv. Enhancing Technol.*, 2020(2):481–498, 2020.

[30] Susan E. McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. Investigating the Computer Security Practices and Needs of Journalists. In Jaeyeon Jung and Thorsten Holz, editors, *24th USENIX Security Symposium, USENIX Security 15, Washington, D.C., USA, August 12-14, 2015*, pages 399–414. USENIX Association, 2015.

[31] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark Patterns After the GDPR: Scraping Consent Pop-ups and Demonstrating Their Influence. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020.

[32] J. P. Nzabahimana. Analysis of Security and Privacy Challenges in Internet of Things. In *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pages 175–178, 2018.

[33] A. Odeh, I. Keshta, and E. Abdelfattah. Machine Learning Techniques for Detection of Website Phishing: A Review For Promises and Challenges. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0813–0818, 2021.

[34] Nuffield Council on Bio-Ethics. The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues. Technical Report 558, University of Cambridge, 2015.

[35] L. Pasquale, D. Alrajeh, C. Peersman, T. Tun, B. Nuseibeh, and A. Rashid. Towards Forensic-Ready Software Systems. In *2018 IEEE/ACM 40th International Conference on Software Engineering: New Ideas and Emerging Technologies Results (ICSE-NIER)*, pages 9–12, 2018.

[36] Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards, and Christian Sandvig. Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook. *New Media Soc.*, 20(1):293–310, 2018.

[37] Awais Rashid, Alistair Baron, Paul Rayson, Corinne May-Chahal, Phil Greenwood, and James Walkerdine. Who Am I? analyzing digital personas in cybercrime investigations. *Computer*, 46(4):54–61, 2013.

[38] Jennifer Rose. To Believe or Not to Believe: An Epistemic Exploration of Fake News, Truth, and the Limits of Knowing. *Postdigital Science and Education*, 2(1):202–216.

[39] Alex Rosenblat and Luke Stark. Algorithmic Labor and Information Asymmetries: A case Study of Uber's Drivers. *International journal of communication (Online)*, pages 3758–3785, 2016. Publisher: University of Southern California, Annenberg School for Communication and Journalism, Annenberg Press.

[40] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential Privacy and Census Data: Implications for Social and Economic Research. *AEA Papers and Proceedings*, 109:403–08, May 2019.

[41] Theodor Schnitzler, Muhammad Shujaat Mirza, Markus Dürmuth, and Christina Pöpper. SoK: Managing Longitudinal Privacy of Publicly Shared Personal Online Data. *Proc. Priv. Enhancing Technol.*, 2021(1):229–249, 2021.

[42] Lucy Simko, Ada Lerner, Samia Ibtasam, Franziska Roesner, and Tadayoshi Kohno. Computer Security and Privacy for Refugees in the United States. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 409–423. IEEE Computer Society, 2018.

[43] L. Sion, K. Wuyts, K. Yskout, D. Van Landuyt, and W. Joosen. Interaction-Based Privacy Threat Elicitation. In *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, pages 79–86, 2018.

[44] J. Daniel Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3):477–564, 2006.

[45] Lior Jacob Strahilevitz. Toward a Positive Theory of Privacy Law. *COASE-SANDOR Institute for Law and Economics Working Paper No. 637 Public Law and Legal Theory Working Paper NO. 421, University of Chicago*, 2013.

[46] Fred Stutzman, Ralph Gross, and Alessandro Acquisti. Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality*, 4(2), Mar. 2013.

[47] M. Sun and W. P. Tay. Inference and Data Privacy in IoT Networks. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2017.

[48] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *2021 2021 IEEE Symposium on Security and Privacy (SP)*, pages 473–493, Los Alamitos, CA, USA, may 2021. IEEE Computer Society.

[49] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Measuring the Impact of the GDPR on Data Sharing in Ad Networks. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, ASIA CCS '20, page 222–235, New York, NY, USA, 2020. Association for Computing Machinery.

[50] Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin. FALCON: Honest-Majority Maliciously Secure Framework for Private Deep Learning. *Proc. Priv. Enhancing Technol.*, 2021(1):188–208, 2021.

[51] Tina Wu, Frank Breitinger, and Ibrahim Baggili. IoT Ignorance is Digital Forensics Research Bliss: A Survey to Understand IoT Forensics Definitions, Challenges and Future Research Directions. New York, NY, USA, 2019. Association for Computing Machinery.

[52] W. T. Young, H. G. Goldberg, A. Memory, J. F. Sartain, and T. E. Senator. Use of Domain Knowledge to Detect Insider Threats in Computer Activities. In *2013 IEEE Security and Privacy Workshops*, pages 60–67, 2013.

[53] Eric Zeng and Franziska Roesner. Understanding and Improving Security and Privacy in Multi-user Smart Homes: A Design Exploration and In-Home User Study. In Nadia Heninger and Patrick Traynor, editors, *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, pages 159–176. USENIX Association, 2019.

[54] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake News: Fundamental Theories, Detection Strategies and Challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 836–837, New York, NY, USA, 2019. Association for Computing Machinery.