Protech.

A systematic literature review to understand: 1) the current artificial intelligence (AI) solutions for the detection and removal of child sexual abuse material (CSAM); and 2) current and ongoing research into bias in CSAM data sets, recommendations, and how to mitigate and minimise the effects of bias.

Introduction

The volume of child sexual abuse material (CSAM[1]) available online, and the demand for it worldwide, has reached unprecedented levels. For example, the National Centre for Missing and Exploited Children's (NCMEC) CyberTipline[2] system received over 32 million reports in 2022, and the majority of these came from Electronic Service Providers (ESP) (NCMEC, 2023). The Canadian Centre for Child Protection (CP3) reports that as of January 2023, their Project Arachnid system had over 22 million removal notices for CSAM to over 1000 ESP worldwide (CP3). In 2021, the Internet Watch Foundation (IWF) reported that the EU has become the largest host of CSAM globally, with 90% of webpages assessed in 2020 hosted in Europe (IWF, 2021); and in 2022 the IWF dealt with 255,588 confirmed cases of images containing, linking to or promoting CSAM that contained CSAM, had links to CSAM, or were advertising it (IWF, 2023). Child sexual abuse has a devastating and lifelong impact on the victims and survivors and the perpetual circulation of the explicit record of the sexual abuse, captured in CSAM, contributes to the re-victimisation and re-traumatisation of the victims and survivors (Wortley and Smallbone, 2006).

It is now generally agreed that an approach that prioritises the prevention of harm through early evidence-based interventions for individuals at risk of offending and at risk of recidivism, together with digital activities that tackle the supply of CSAM, is critical for sustainable long-term prevention (Scherrer & van Ballegooij, 2017). This presents an opportunity for different organisations to come together to collaborate and collectively tackle the online sexual abuse and exploitation of children. An example of this is the IWF CAID Taskforce, a collaboration between the IWF, the UK government (via the Child Abuse Images Database – CAID) and Thorn (a company that develops technology to tackle CSAM), with almost two million CSAM images being graded and hashed[3] over a two-year project from 2021 to 2022 (IWF, 2022).

In terms of the technological challenges facing responses designed to tackle CSAM, some examples include the use of anonymisation tools; private peer-to-peer (P2P) networks; the

[1]This is used as an umbrella term to encapsulate all other terms used to describe child sexual abuse materials, which include sexually explicit photographs, videos, text / written messages, cartoons or animations that depict harm to children. CSAM is usually associated with online offences but can also be used during contact offences.
[2]The centralised reporting system for the online CSAE in the US.
[3]This refers to the process of giving CSAM a digital fingerprint (IWF, 2023).

Funded by the European Union

development of ways to accurately determine the age of victims; the automatic classification of CSAM (which given the nature of the content, is challenging for humans, exacerbated by the growing and relentless volume of CSAM); the availability of datasets rightly held by law enforcement agencies; and the potential bias within these datasets (Bursztein, Bright, Clarke et al., 2019; Porto, Lima, Franco et al., 2020; Roselli, Matthews and Talagala, 2019).

Artificial Intelligence (AI), and the decisions it makes, are pervasive throughout our daily lives and social systems, such as in employment, dating, advertising, and judicial systems (Mehrabi et al., 2022). In simple terms, machine learning (ML), a subset of AI, is used to develop AI that is able to resolve complex problems, such as accurately recognising CSAM. ML algorithms essentially enable AI, which repeatedly receives large amounts of data, to process and recognise patterns within the data, until the AI learns to understand the data environment to deliver on its task (Alelyani, 2021; Du, Hargreaves, Sheppard, 2020). This is done by utilising statistical learning, which allows the AI to make predictions based on the data. However, Ahmed et al. (2021:557) point out the need for caution, because "*massive dependency on such systems can lead us to think that our system is robust and fool-proof.*" Therefore, the datasets that are used to train the AI are crucial, and there is a recognition of the need to create tools that can tackle potential bias in datasets, to ensure they are as fair as possible in order to increase the accuracy of AI responses (Alelyani, 2021; Pagano, Loureiro, Lisboa et al., 2023).

The purpose of this review is to develop knowledge and understanding of the current landscape of available AI solutions for the detection and removal of CSAM, and research into bias in CSAM data sets, recommendations, and how to mitigate and minimise the effects of bias.

Methods

A 'question-led[4]' (Government Social Research Unit, 2008) approach was adapted for this systematic literature review and was utilised to identify, assess and synthesise relevant literature in order to provide a descriptive overview and narrative synthesis to address two research questions. A systematic literature review is recognised as a comprehensive method to provide an objective analysis of the current knowledge on a topic, and it is recommended that a standalone literature review, such as this, should be systematic, explicit in reporting methods, comprehensive in scope, and therefore reproducible (Okoli, 2015). It was the chosen method here as it allows us to identify, assess and synthesise available literature in an efficient and timely manner (Garritty, Gartlehner, Kamel, et al., 2020), and also accommodates the research constraints regarding the time available (three months) and resource constraints (small research team), and the requirement of a short report (20 pages).

[4]Criteria for this are means the two research questions are the drivers of the searches and the sifts; the questions should represent a statement that can be investigated, and they should be clear and answerable (https://webarchive.nationalarchives.gov.uk/ukgwa/20140402163101/http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/how-to-do-a-rea)

Funded by the European Union

The purpose of this review is to identify, evaluate and synthesise peer reviewed research papers and grey literature, relevant to two research questions, and as such, two distinct searches were undertaken:

Search 1 was to find literature relevant to the research question:

1. *What artificial intelligence/machine learning technology is currently being used to tackle (detect and remove) CSAM from the internet?*

Search 2 was to find literature to answer this research question:

2. *What can we learn from current and ongoing research on bias in CSAM datasets, and how can we best mitigate and minimise the effects of bias?*

Search strategy

A search strategy was designed to identify peer-reviewed research papers, books, book chapters and grey literature that related to current research regarding AI solutions for the detection and removal of CSAM, and managing bias in CSAM datasets. Given the interdisciplinary nature of this research topic and following the expert advice of a university librarian, five databases were searched to maximise relevant returns: ACM Digital, IEEE, Web of Science, ProQuest Social Science Premium, and EBSCO. The searches took place in April 2023. Key search terms formed the basis of the separate search strings, which were developed for both searches (see Appendix 1). For both searches, literature was included if it was written in English, published between 2018 and 2023[5], and peer-reviewed (see Appendix 2 for inclusion criteria). Grey literature was sought through the database searches and from the researchers' networks, technology partners involved in the PROtech project, and from tech industry websites. Grey literature was included if deemed relevant to either of the research questions.

Study selection

Results from the two database searches were subject to a three-stage screening process: title and abstract, full texts against inclusion criteria, and a risk of bias (RoB)/quality assessment (QA) for research evidence or reviews. The first stage was the title and abstract screening. One researcher screened the relevant titles and then the abstracts against the inclusion criteria. This resulted in 74 potentially relevant papers, which were saved in RefWorks, checked for duplicates, and then exported to Microsoft Excel to progress through a full-text screening for inclusion in the review.

The second stage of screening was the full-text screening against the inclusion criteria above. The same process and criteria were used for both Search 1 and Search 2 (see Appendix 3 for reasons for exclusion). Two researchers worked on this screening process, which resulted in 45 papers progressing to the third screening stage, an RoB/QA (see details below). This resulted in the inclusion of 31 papers. Throughout this process, the two researchers discussed all papers they were unsure about, and the final decision was made

[5]This timeframe was selected given the research questions onus on current research in this space.

Funded by the European Union

by the principal researcher based on judgements regarding relevance. There was also an inter-rater reliability check at this point, whereby a third senior researcher screened 25% (n = 8) of the papers against both the inclusion criteria and the RoB/QA criteria and disagreed about the inclusion of one paper. This was resolved through a discussion, and the paper was excluded, leaving a total of 30 papers to be included in the review.

Table 1: Screening Record of papers after searches and de-duplication

After Title & Abstract sift	After full text sift against inclusion criteria	Totals included in review after Risk of Bias QA*	Peer-reviewed papers	Grey literature
Search 1: Current AI solutions to detect and prevent CSAM online				
51	27	13	4	9
Search 2: Current and ongoing research into bias in CSAM data sets, recommendations, and mitigation on how to minimise the effect of bias**				
23	18	17	5	12
TOTALS				
74	45	**30**	9	21
*After third researcher review.				

As a result of the paucity of research results related to AI ***and* CSAM datasets, we revised research question 2 and relaxed the inclusion criteria to include papers that discussed bias in datasets generally. They were included if deemed relevant to the purpose of the review, with the aim to assist with developing and improving CSAM datasets. The remaining papers discussed various aspects of bias in datasets and AI and ML more generally.

Risk of bias/quality assessment

The third assessment stage for both searches was a risk of bias/quality assessment (RoB/QA), and papers that passed this stage were included in the narrative synthesis. For use in this review, we utilised a novel RoB assessment tool which we had previously developed, informed by the Cochrane Risk of Bias tool (Higgins and Green, 2008). The tool focuses on issues in the design and execution of studies that could affect the results, in line with Cochrane guidance (Higgins et al., 2019). The RoB/QA tool focuses on seven areas that may result in risk of bias: 1) appropriate aim and statement of findings 2) appropriate sample selection/strategy 3) sample size, 4) execution of the study, 5) appropriate analyses, 6) clear and coherent reporting of the results, 7) conclusions are supported by results. A pre-agreed

Funded by the European Union

and pragmatic threshold[6] was set for inclusion into the narrative synthesis. Each study was rated on each criterion, but we allowed for a 'n/a' option on some of the criteria to account for the grey literature and non-research papers such as systematic literature reviews, (see Appendix 4 for full details of the RoB tool).

Grey Literature

The topic of technological developments in AI *and* CSAM datasets specifically is currently very under-researched, and as a result, both searches returned more grey literature than peer-reviewed literature. It is not standard practice to QA grey literature, but in this review it was screened against the inclusion criteria. The grey literature was not assessed using the RoB/QA tool, but it was screened by the research team to check whether it was relevant to the purpose of the review, offered knowledge and/or ideas that may stimulate thought and discussion, and inform practice relating to the research questions (Adams et al., 2017). We recognise the subjectivity of this assessment, and as such, the research team discussed the papers they were unsure about and made inclusion/exclusion decisions together. If agreement was not achieved, a third senior researcher made the final decision, based on their review of the literature against our requirement that it was relevant, offered knowledge and could inform practice relating to the research questions (Adams et al., 2017).

Search Results

Of the 30 papers that were included in the review, 13 related to search 1 and 17 related to search 2. For the search 1 results, 4 were peer-reviewed papers and 9 were forms of grey literature; for search 2, 5 papers were peer-reviewed and 12 were forms of grey literature. All papers were published between 2018 and 2023. Additional grey literature (n=7) obtained through internet searches of technology sites or via technology partners was also included (please see References section: *Industry websites and grey literature*).

Narrative synthesis of findings

This section provides a high-level review and synthesis of the literature in order to address each research question. Much of the literature included in this review, especially pertaining to research question 2, is highly specialised and involves complex information which is very specific to computer sciences. It is beyond the remit and expertise of this review to evaluate the models and tools presented in the papers, nor is it our intention to give a detailed description of the science behind the AI and ML processes and tools. Rather, we aim to provide an accessible overview to non-specialists about AI and ML in relation to CSAM, biases in datasets and some ways they can be tackled. Given the lack of information specific to CSAM datasets, we have drawn on our technological partners to assist in the application of these techniques to CSAM datasets.

Research question one is addressed first, with a focus on presenting what AI and ML technology is currently used to tackle CSAM. Research question 2 is then addressed in two

[6]NB Threshold for exclusion =For research papers = 'When a study has 2 x No assessments OR, 1 x No AND 2 Can't Tell'. For review papers when n/a is applied = 'When a study has 1 x No assessments OR, 1 x No AND 1 Can't Tell' (to compensate for the fewer criteria).

Funded by the European Union

parts: 1. to present an overview of the current landscape regarding bias in datasets, and 2. ways to tackle and mitigate its effects.

Research Question 1: What artificial intelligence/machine learning technology is currently being used to tackle (detect and remove) CSAM from the internet?

The purpose of AI is to use machines to replicate, or imitate, human intelligence, such as decision-making and reasoning, by processing information (data) using ML amount of online images and videos of children being sexually abused.

Thirteen papers were included to address this research question. Five proposed a tool, a method or an approach to tackling CSAM; six were reviews or papers, generally exploring the current landscape relating to the detection of CSAM; and two were focused on the practice of organisations that tackle CSAM (The IWF and NCMEC).

Ngo et al. (2022) conducted a comprehensive survey of 35 current research papers relating to online CSAM. Although this paper does not give details of the contents of each of the papers reviewed, it is useful to this review as it provides an additional layer of insight into the current landscape. Ngo et al. (2022) organised the papers into categories based on each one's dominant focus, such as the tasks addressed, the data sets used, and the AI techniques applied. In terms of the tasks addressed, most of the reviewed papers focused on Text Analysis, followed by image and video. In relation to the techniques applied to detect CSAM, most papers focused on statistical applications and on machine learning/deep learning techniques, followed by Natural Language Programming[7] (NLP).

The challenges in tackling CSAM

There are several challenges facing law enforcement and organisations who work in the child protection space. Bursztein et al. (2019) conducted a longitudinal measurement study of the online distribution of CSAM, in collaboration with the NCMEC. Their results demonstrate: the exponential growth in CSAM to almost one million reports per month, which far outstrips any capacity to tackle CSAM; and the rise in a diverse range of sharing platforms and evolving technologies, such as peer-to-peer, Tor (the dark web), mobile phones and social media platforms, all of which have accelerated the creation and sharing of CSAM within a global context.

Given these challenges, Lee et al. (2020) highlight another pressing issue and the more hidden side of CSAM – the re-victimisation of the children who are sexually abused every time an image of this abuse is shared and re-shared. They also point out that the work of law enforcement and organisations in this space is crucial to explicitly challenge the normalisation and minimisation of CSAM, which also represents a dehumanisation of the victims. Like many of the papers included in this review, Lee et al. (2020) recognise that automated detection techniques are the most effective and efficient methods to use because

[7] Concerned with giving computers the ability to understand text and spoken words in much the same way human beings can - https://www.ibm.com/topics/)

Funded by the European Union

they are faster and can also reduce the harm to the professionals who must review and categorise the images.

Overview of approaches and examples of their implementation to tackle CSAM

Lee et al. (2020) present a comprehensive review of the current challenges regarding the detection and removal of CSAM from the internet. They identify four broad categories of techniques to detect CSAM: hash value systems and image hash databases; web-crawlers; detection based on filename; and visual detection. This section synthesises the included papers within these broad categories, as far as possible, as well as including additional techniques, and provides examples of how these methods are currently being implemented.

Hash values systems and image hash databases

Image hashing is the most common and successful method of detecting known CSAM circulating on the internet (Lee et al., 2020). Images are given a unique hash value which serves as an identifying fingerprint and allows re-circulated images to be identified and removed from the internet (for a more detailed account of the hash process, see Lee et al., 2020, p.6). However, this method has some limitations, such as that only CSAM which has been hashed can be identified by the database, it cannot find new material, and it is easy for perpetrators to bypass this detection method by making slight changes to the file or image (Lee et al., 2020).

The IWF frequently utilises this method to tackle CSAM (IWF, 2023) and it refers to targeting an illegal file/image and giving it a unique hash value, which is used to identify it. The IWF in the UK, and NCMEC in the USA, have created image hash databases of these hashed illegal images which enable them to identify them if they circulate on the internet and then remove them. Hash values focus on the location of the images. This system relies on organisations like the IWF and NCMEC to work in conjunction with law enforcement and tech companies. An example is the Image Detection and Filtering process (IDFP), developed by AOL (America Online, an internet access platform) in 2004, which can scan AOL emails and generate hash values for any transmitted images. These are then compared with an internal database which contains the hash values of CSAM. The content is removed and, in the USA, NCMEC's CyberTipLine is notified of the username, their email and postcode. Following this, NCMEC notify law enforcement who can request the user's full name from the ISP.

Struppek et al. (2022) conducted a comprehensive empirical analysis on Apple's NeuralHash[8], a perceptual hashing system that maps an image to a numerical hash system, to detect CSAM on users' devices before files are uploaded to Apple's iCloud. They found that the perceptual hashing systems like NeuralHash are not robust as the hash values could be manipulated, allowing the detection system to be exploited, from hiding CSAM to 'framing innocent users'. However, we have been advised by an expert at the IWF to express caution in these assertions as 'framing innocents' is a very complicated process that also requires access to the CSAM hashes and a trained specialist (IWF expert, personal correspondence, 2023).

[8]A perceptual hashing function, based on the features of an image as opposed to its pixel values (https://blog.quastor.org/p/apples-perceptual-hashing)

Funded by the European Union

Web-crawlers

Web-crawlers can play an important role in the detection of CSAM. They can search websites, which are often connected by perpetrators' networks, looking for CSAM and guided by keywords that have been incorporated into their learning and design. Research evaluating the effectiveness of web-crawlers have found them to be successful at detecting CSAM if they are given appropriate criteria, such as images and keywords (Westlake et al., 2012; Westlake et al., 2017). However, it is crucial that the web-crawler can differentiate between what is and is not CSAM in order to be fully effective, and this relates to the quality and accuracy of the training data, which is discussed later.

Two examples of web-crawlers that are currently being used to detect CSAM are the IWF Hash List (IWF, 2021), and Project Arachnid, housed within the Canadian Centre for Child protection[9] (CP3). Both web-crawlers use Microsoft PhotoDNA (described below) and create a list of hashes on their CSAM database which assists technology companies in detecting and removing CSAM from their websites (IWF, 2021; Lee et al., 2020). The purpose of the crawlers is to prevent re-victimisation of children and they can search millions of webpages a day, find old CSAM that has been widely distributed and remove it from the Web. The Project Arachnid web-crawler is also able to access the darknet and can identify links between the darknet and the open web. On average, the Arachnid crawler detects over 100,000 potential CSAM images a month.

Detection based on filename and metadata.

This system proposed by Panchenko et al. (2012, cited in Lee et al., 2020) allows CSAM to be detected on peer-to-peer networks. It is based on ML that can identify queries and/or filenames that contain CSAM by extracting metadata from both CSAM and non-CSAM files.

Visual detection and facial recognition techniques

Visual detection methods are useful at recognising new CSAM content. Challenges in this area are related to skin tone, facial features, age and gender, and the accurate classification of adult/non-adult. Lee et al. (2020) summarise some visual detection methods that have had some success in identifying CSAM, for example an application that can improve the existing skin tone through a filter which has shown a 74.19% accuracy rate (Sae-Bae et al., 2014, cited in Lee et al., 2020). Other research has investigated the use of Convolutional Neural Networks[10] (CNN) for age classification (see Lee et al., 2020, for a more detailed summary) and has demonstrated that the use of CNN/deep learning is effective for CSAM detection.

Facial recognition is a common technique designed to map facial features to enable identification. It also relates to using faces in images/videos to estimate age and gender, which can be particularly crucial when trying to detect CSAM and progress CSAM investigations. Although aspects of facial analysis have improved, such as automated age estimation (Anda et al., 2019), facial recognition remains a challenging area due to

[9] See https://www.projectarachnid.ca/en/

[10] A network architecture for deep learning which learns directly from the data.

Funded by the European Union

influential factors such as lighting, the quality of the image, make-up, and context (Anda et al., 2019). Porto et al. (2020) suggest utilising anthropological knowledge to support CSAM detection, specifically the creation of an automatic classifier based on use of photo anthropometric[11]analysis of the human face to develop an artificial neural network which can estimate sex and age (Porto et al., 2020).

URL list blocking

The IWF produces and regularly updates a URL list that identifies webpages that are hosting CSAM, and this is shared with Internet service providers (ISP) to support them to check their websites. Martellozzo and DeMarco's (2020) explored the IWF's process of removing CSAM from the internet to see if this was useful and effective for law enforcement. Using a mixed methods approach, IWF employees and key stakeholders were interviewed to understand their perceptions and experiences of this and found that while IWF employees and stakeholders recognise the value of the URL list, there was a consensus that it should not be a priority anymore because of the changing nature of behaviours relating to CSAM production and distribution. Instead, the Hash List was considered to be a more effective way forward. However, this research was conducted in 2017 and may no longer represent the views of current IWF employees. Both Hash and URL lists are different mechanisms appropriate to different harms, they are complementary, and both continue to be important to the IWF's mission (IWF expert, personal correspondence, 2023).

Key word database

In the last decade the IWF has also developed a list of the language that perpetrators of CSAM use to communicate and distribute CSAM, which is shared with its partners (e.g. Microsoft, Google and Facebook and law enforcement) to help detect CSAM. This is important because perpetrators of CSAM abuse create unique codes and special words or phrases as a way to hide the content of the messages or sites, on a range of platforms such as gaming and chat forums (IWF, 2021[12]). Westlake et al. (2012) tested the use of keywords by combining them with image hashes to identify CSAM and found that using both techniques was more successful than the techniques used alone, and that a selection of keywords reduces the potential for false negatives.

Lee et al. (2020) also identify 'CSAM detection as a service' provided by technology companies. For example, Microsoft's PhotoDNA[13] application was developed in 2009 and helps find and remove CSAM images. This application is now widely used and has recently been developed to include the identification of known CSAM videos. Google also has a tool to assist in the detection of CSAM, which uses deep learning technology[14] which uses deep neural networks to detect CSAM images and videos. This tool can also target new, previously unfound content, which helps protect against the revictimisation of the children.

Tools, techniques and methods to tackle CSAM

[11]This is a technique that measures the face and is able to compare them with another image of a face.

[12] See https://www.iwf.org.uk/our-technology/our-services/keywords-list/

[13] See https://www.microsoft.com/en-us/photodna/cloudservice

[14] A method in AI that teaches computers to process data in a way that is inspired by the human brain.

Funded by the European Union

Some of the literature proposes new and developing ways to potentially detect CSAM, most of which only have some provisional early testing, and is summarised in Table 2 below.

Table 2: Summary of new and developing methods to tackle CSAM.

Author & Title	Description	Area of detection	Authors' evaluation
Scholey, A., and Zadeh, P. B. eds. 2022. A Digital Forensics Live Suspicious Activity Toolkit To Assist Investigators With Sexual Harm Prevention Order Monitoring.	"A proof-of-concept digital forensic investigation toolkit is proposed to assist Public Protection Officers (PPO) in the analysis of digital devices belonging to these individuals. The proposed Live Suspicious Activity Toolkit (LiSA -T) will enable a time-efficient, up to date assessment of any suspicious activity and behaviour on a Windows 10 computer. LiSA-T will evaluate and then report the findings, assisting the PPO with an informed decision as to whether to involve the Digital Forensic Unit (DFU), to further examine a device in a more in-depth forensic manner." (p.1).	Up-to-date assessment of suspicious sexual activity by those under a Sexual Harm Prevention Order (SHPO).	Still under development, easy to use, easy to deploy and portable. A key feature of LiSA-T is that it looks for CSAM wherever it may be stored on the system. LiSA-T checks for all USB storage devices that have been connected to the system
Murcia Triviño, J., et al. (2019). C3-Sex: a Chat bot to Chase Cyber perverts	A smart chatbot to interact with suspects in order to profile their interest regarding 'child pornography'. It is based on an Artificial Conversational Entity (ACE) that connects to different online chat services to generate posts and start conversations that exhibit the behaviour of someone interested in acquiring and distributing CSAM. The bot should be able to manage conversations and will be able to analyse the conversation in order to profile the suspect. C3-Sex is composed of four models: Retrieval-based, Generative, Emotional classification and	Online perpetrators of CSAM	Able to maintain long conversations with suspects. Builds a friendly relationship. Suitable for supporting LEA. Need to improve interactions to make more human-like.

Funded by the European Union

	Opinion classification. Together, these models constitute a solution able to keep conversations with suspects and profile them to identify 'perverts'.		
Moreira, D. C. & Fechine, J. M. (2018) A Machine Learning-based Forensic Discriminator of Pornographic and Bikini Images.	A method to improve a known technique for detecting nudity and ability to distinguish between pornographic images and bikini images, and potentially applicable to CSAM detection. It uses skin detection and facial recognition in ML.	Differentiate between images of bikinis and pornographic images.	Outperformed existing algorithms.
Anda, F., Lillis, D., Kanta, A., Becker, B.A., Bou-Harb, E., Le-Khac, N.-A., Scanlon, M., 2019. Improving Borderline Adulthood Facial Age Estimation through Ensemble Learning	Improves the accuracy of underage estimation in conjunction with a deep learning model (DS13K) that has been fine-tuned on the Deep Expectation (DEX) model.	Facial age	Achieved an accuracy of 68% for the age group 16 to 17 years old. Comprehensive performance evaluation of offline and cloud based facial recognition models.
Kobayashi, H., Kadoguchi, M., Hayashi, S., Otsuka, A., and Hashimoto, M. eds. (2020).	A proposed method to develop an expert system with a mechanism to automate crime category classification and threat level assessment, using the information collected by crawling the dark web. In a test, using a 'bag of words' from 250 posts on the dark web, the system counted the frequency of	Online criminal and harmful content	Experimental result (with 1500 dark web posts) shows 76.4% recall rate for 6 criminal category classifications. Under development:

Funded by the European Union

An Expert System for Classifying Harmful Content on the Dark Web.	the words and classified sample posts into 6 criminal categories including CSAM.		more crime categories; improve displays and logic.

Some considerations of ways to further tackle CSAM

Razi et al. (2021) argue that taking a human-centred approach to AI and ML is important in the detection of online sexual risks and advocate for 'human-centred machine learning' (HCML). This is a new field of computer science which draws from the social sciences so that AI and ML can respond to societal needs and learn from human insight and ethical considerations. The authors highlight this approach as particularly important in developing a more victim and survivor-centred approach to the detection of sexual risk and provide a conceptual framework for use in reviewing literature using a human-centred lens.

Bursztein et al. (2019) point out the changes in the distribution of CSAM over the last 25 years, from File Transfer Protocols (FTP) and email being the earliest distribution tools in the late 1990s to early 2000's, which shifted to chatrooms and messenger in the mid-2000s, then to gaming, SMS and mobile phone to the dark web (Tor) and URLs around 2015 onwards (see Bursztein et al., 2019, p.2605). It is fair to assume that this will continue to evolve as new technologies develop. As a result, the authors recommend continued and co-ordinated global action, and present some new ways to develop and utilise deep learning technology to automate detection of CSAM in order to tackle it, especially new content. For example, they suggest deep learning as a way to scale the review process by using two types of automated clustering: scene clustering involves grouping reported images/videos to look for the same scenes within them (e.g., backgrounds); facial clustering by developing existing facial recognition technology to further assist in identifying the same victims and perpetrators across images/videos. The authors also recommend using deep learning to help prioritise the 'most actionable' content, and to reduce harm to professionals who have to review CSAM.

Research Question 2: What can we learn from current and ongoing research about bias in CSAM datasets, and how can we best mitigate and minimise the effects of bias?

The initial intention of this review was to develop knowledge and understanding of the bias in CSAM datasets and ways to mitigate this. However, our search results revealed a paucity of literature in this space, particularly regarding CSAM datasets. Only one paper (Laranjeira et al., 2022) focused specifically on CSAM datasets but this was more about overcoming the challenges for researchers to access them.

Funded by the European Union

The part of the review will firstly outline the issues and challenges surrounding CSAM datasets that inhibit our knowledge and understanding about biases that may be specific to them. We then synthesise the content of the papers by giving an overview of the issues involved in dataset bias, describing types of bias and exploring some causes of bias in datasets. We present existing or proposed ways to mitigate and minimise the effects of bias in ML models and AI.

CSAM datasets

CSAM databases are in the possession of law enforcement agencies, for example Interpol's Child Sexual Exploitation database[15], which contains more than 4.3 million items of CSAM, and the UK government's Child Abuse Image Database[16] (CAID). The only non-law enforcement agency to access CAID is the IWF, who are working with the database to find and remove CSAM from the internet (IWF, 2022). Laranjeira et al. (2022) highlight the lack of information regarding the biases that might exist in CSAM datasets, and this paucity was certainly supported by our database searches. Researching and gaining understanding of CSAM datasets is challenging because of the sensitive and illegal content, and they are largely inaccessible to researchers. We were assisted by an expert from the IWF who was able to provide insight into how bias may affect a CSAM dataset specifically (IWF Expert, personal correspondence, 2023), and we draw on this insight throughout this section.

Laranjeira et al. (2022) recognise the challenges in researching and understanding CSAM datasets and propose an analysis template that is designed to enable the safe publicising of some characteristics of CSAM datasets for researchers, while not re-victimising the children abused in the images. Their paper reports on a case-study whereby they tested the validity of the analysis template using the Region-based annotated Child Pornography Dataset[17] (RCPD). This contains over 2000 samples among regular and CSAM images, produced in partnership with Brazil's Federal Police.

The authors chose the following attributes for automated extraction[18]: demographics (age, gender, etc.); pornography; context (objects, scenes); quality (e.g., luminance); and metadata (e.g., file extension, image resolution). In terms of bias, the researchers found that the RCPD indicates that CSAM data shared online has a different distribution from reported cases of physical abuse in Brazil in terms of race. They acknowledge that one of the main challenges for CSAM is distinguishing it from legal images of children and adult pornography, and they consider it essential to balance available benchmarks in both dimensions.

A range of documentation practices in the literature was explored, showing that extracting sparse attributes and presenting them as aggregated statistics is both valuable and safe. In discussing their proposal, the researchers acknowledge that the automatic extraction of features has the downside of relying on measures that are uncertain and prone to errors; however, they recognise that it produces cheap annotations without adding to the burden of

[15]https://www.interpol.int/en/Crimes/Crimes-against-children/International-Child-Sexual-Exploitation-database

[16]https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759328/CAID_Brochure_May_2018_for_gov_uk.pdf

[17]A benchmark produced by the Federal Police of Brazil, due to its extensive set of labels (Laranjeira et al., 2022).

[18] This refers to information automatically being extracted from web pages as part of the ML process.

Funded by the European Union

law enforcement agents providing benchmarks for the research community. The paper shows that context information from scene and object classification can be correlated with CSAM labels, highlighting that residential scenes and child-like visual cues often occur in scenes of child abuse, and are therefore valuable dimensions for detection.

Bias in datasets – an overview of the issues

Bias is a broad, multi-disciplinary concept that can have multiple meanings. None of the included papers provide a definition of bias in relation to AI, ML and datasets. Kaminska's (2022) definition below is chosen for this review as it offers a succinct explanation of bias in relation to datasets:

> *"Bias in data is an error that occurs when certain elements of a dataset are overweighted or overrepresented. Biased datasets don't accurately represent ML model's use case, which leads to skewed outcomes, systematic prejudice, and low accuracy."* (Kaminska, 2022).

A fundamental and central concern is that the outcomes of AI models, in the real world, are free from bias and fair, and all of the included papers are concerned with these issues (for example, Alelyani, 2021; Mehrabi et al., 2021; Chakraborty, Majumdar and Menzies, 2021; Milaninia, 22020), which link to concerns about the trustworthiness of AI (for example see Chen, 2023; Liu et al, 2022). In relation to CSAM datasets and in real life terms, if the AI cannot do this accurately, it is possible that a victim of CSAM abuse will continue to be re-victimised as that image remains on the internet (IWF expert, personal correspondence, 2023).

AI and ML models can perpetuate and exacerbate biases that are inherent in datasets, which can result in outcomes that may discriminate against certain groups (Jiang and Nachum, 2019; Mehrabi et al., 2021). Often, biases are not identified until after AI models are deployed (Wang et al., 2021) and so understanding, identifying and tackling bias in datasets is crucial to trustworthiness, fairness and non-discriminatory datasets. Given this, some authors point out that researchers and engineers have an ethical duty to ensure AI can make fair decisions (Chakraborty, et al., 2021; Pagano et al., 2022).

Bias and fairness are often conceptualised as a binary, with bias and unfairness at one end and fairness and trustworthiness at the other. For example, Mehrabi et al. (2021:2) focus on fairness and safety in ML and define bias inversely, and in relation to fairness, as the "*absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics*". Bias or unfairness in CSAM datasets could lead to the absence of detection of certain groups of children being sexually abused. For example, if the AI has not been exposed to children with different skin tones, those children in CSAM may not be identified. The authors also identify challenges that need to be addressed in the ML space, such as gaining consensus on a what is understood by the concept of fairness in machine learning in order to try to make different systems more unified and comparable and using an agreed definition of fairness to search for, and more easily recognise and agree on, unfairness.

Funded by the European Union

Lui et al. (2022) and Chen and Sundar (2023) focus on the trustworthiness of AL. Lui et al. (2022) frame the trustworthiness of AI from three interconnected perspectives: a technical perspective, in that it is accurate, robust and transparent; a user perspective, in relation to safety, privacy and accessibility; and from a social perspective, meaning AI should be ethical, fair and law-abiding. Linked to trustworthiness is the need for an ethical framework, the idea that AI is beneficial, responsible and fair.

In a similar vein, Chen and Sundar (2023) discuss the trustworthiness of AI in terms of data transparency and highlight the 'white-box' or 'black-box' types of ML model. White-box ML models are likely to be easier to explain and simpler in terms of functionality, outcomes and structure. Black-box ML models are more mathematical, harder to understand their structure and functioning and explain to non-specialists. As such, black-box models may tend to lack transparency, which may impact on levels of trustworthiness (Chen and Sundar, 2023).

An overview of types of bias

There are many types of biases identified in the included papers and the purpose of this section is to provide an overview of some of them that may relate to CSAM datasets.

Pagano et al. (2022), in their systematic review of bias in ML models, provide three broad classifications of types of bias: data bias (present in datasets), algorithmic bias (caused by algorithms), and user bias (caused through user interaction). Mehrabi et al. (2021) also highlight the interactive nature of bias, describing the 'feedback loop phenomenon' (iterated algorithmic bias) between data, algorithms and users, whereby bias in data and algorithms can impact on the users and can confirm and perpetuate existing biases.

Alasadi et al. (2022) discuss how bias can arise in face-matching[19] algorithms, something which could occur in CSAM datasets. Their work relates particularly to intersectional bias, which is the result of the simultaneous presence of several demographic variables such as age, race and gender. Cabrera et al. (2019) also focus on identifying intersectional bias in ML and highlight how the intersectional nature of these potential biases can further increase the challenges of tackling bias, which is important to mitigate to ensure the AI does not imitate real-world intersectional biases and discrimination. This type of bias is also likely to effect CSAM datasets when training the machine to identify images that have multiple attributes such as age, gender, skin colour and sexual activity (IWF expert, personal correspondence, 2023).

Bias can be introduced to ML at almost every stage of its development and there is a general consensus that bias in ML is primarily caused by the nature and quality of the training data. If the training data used for ML is not representative, or is mis-labelled, it creates a potential cause of bias in datasets, increasing the likelihood of biased outcomes of the AI (Chakraborty et al., 2021; Chen and Sundry, 2023). Chen & Sundry (2023) highlight the importance of the quality of labelling in training datasets, as a way to tackle labelling bias, measured by how accurate the label matches the raw data. Ahmed, et al. (2021) point out that some labels will

[19] Refers to the task of matching a low-resolution face image of a person with a high-resolution face image of the same person (Alasadi et al. (2022:150)

Funded by the European Union

be more 'favourable' or 'unfavourable', which affect the direction of the outcome for some groups.

Some of the papers highlight and discuss the importance of recognising the influence of human bias within ML datasets (Ahmed et al., 2021; Alelyani, 2021; Milaninia, 2021). Alelyani (2021) argues that datasets contain bias because they reflect human behaviour, practice, experience, and actions, and machine learning models are therefore inclined to be biased due to the bias in the training datasets. The author also makes a philosophical point that it is humans who also identify, or socially construct, what actually is a bias, both in the real-world and in terms of machine learning, and this in itself could be biased or inherited bias.

Humans are integral to the process of machine learning in CSAM datasets. It is humans who decide what information is needed and who then teach the machine by telling it what to look for and classify as CSAM. It is also humans who review the outcomes of the ML and interpret how the machine developed its logic to produce its results or decisions (IWF expert, personal correspondence, 2023). Milaninia (2021) identifies the most common human biases that appear in datasets as: implicit[20], selection[21], reporting[22], group attribution[23], and automation[24] bias. These biases exist in the real world and can be transferred to AI and ML.

Ahmed et al. (2021) position the bias of humans as a central and urgent concern that needs to be understood and addressed if we are to eliminate bias in ML and AI. The authors recognise the possibility of humans potentially feeding their own biases into AI models through the training data as a result of assumptions, mis-labelling or inadequate sampling. Labelling data relates to 'tagging' images in a CSAM dataset, which involves the viewing and assessing of the images by humans and training the machine to recognise and classify the images correctly as CSAM or not (IWF expert, personal correspondence, 2023).

Causes of bias

Ahmed et al. (2021:557) summarise the fundamental, but simplistic, issue at the root of any discussion regarding the cause of bias in dataset, "*garbage in, garbage out*". Drawing on the section above and the literature included in this review, the causes of bias in datasets can be distilled to the fact that ML models are shaped by the training data used to teach them, which in most cases is collected, curated and tested by humans. Chakraborty et al. (2021) investigated whether the root causes of bias stemmed from prior decisions made regarding what data is chosen (by humans) and the label assigned to them (by humans). This opens the door to, at the very least, subjective judgement, and at worst implicit or explicit biases. Mehrabi et al. (2021:5) reiterate that most AI models and algorithms are inherently data-driven and can only learn and develop using data, and as a result "*data is tightly coupled to the functionality of these algorithms and systems.*" A further issue is obtaining the content

[20]Occurs when assumptions are made based on one's own mental models and personal experiences*.
[21]Occurs when training data sets fail to reflect their real-world distribution*.
[22]Occurs when how people write about actions ,outcomes or properties is not a reflection of their real-world distribution*.
[23]A tendency to project norms or beliefs of a few individuals to an entire group to which they belong*.
[24]Refers to the human tendency to favour results generated by automated or computer systems over those generated by non-automated systems*
*Milaninia, 2021

Funded by the European Union

for training sets, and this is particularly challenging in terms of CSAM datasets (Laranjeira et al., 2022).

Ways to mitigate against and minimise bias in datasets

Central to both the cause and the mitigation of bias in datasets is the efficacy of both the training dataset and the training process. Most sources of bias can be distilled, as described above, as stemming from the quality and representativeness of the training data, the accuracy and the quality of the labels assigned to the data and algorithms, and the potential impact of human bias.

As stated earlier, as bias can occur at every stage of ML, so too can the process of mitigating against and minimising bias. Mehrabi (2021) suggests that bias in algorithms can be targeted throughout the life-course of the processing stages involved in ML: the pre-processing stage, the in-processing stage, and the post-processing stage of ML. Techniques within these three stages are able to transform the data to remove erroneous AI in the pre-processing stage, modify the learning algorithms in the in-processing stage, and compare and assess the outcomes in the post-processing stage.

Roselli et al. (2019) suggest several ways to manage bias in AI datasets, such as vetting the training data to ensure it is accurate and relevant and seeking to identify bias by evaluation of the datasets, which can involve specific toolkits (see Table 1), and matching the training to the production needs, which can be done by monitoring both the input and the output of the AI.

Sevetlidis et al. (2022) propose a method to mitigate bias caused by overfitting in datasets. Overfitting[25] is an unwanted ML behaviour which occurs when the ML model can only respond accurately to the training data but is unable to do so against new data, which defeats its purpose. To tackle the bias caused by overfitting, they suggest expanding the training dataset with more diverse samples and propose a novel pipeline which gathers images from the internet to construct an image dataset with low bias. They created a data collection of more than 880,000 previously unseen images and tested the proposed method, which classified the images with 72% accuracy.

Dixon et al. (2018) demonstrate how imbalances in training data can lead to unintended bias, and they present a first step towards mitigating and increasing fairness in text classifications. They tested strategies to mitigate bias, such as adding new data to the most disproportionate data bias. For example, classifications of text as 'toxic' were disproportionately high, so the researchers introduced more 'non-toxic' data into the ML to re-balance the dataset. Chakraborty et al. (2021) also developed a tool to mitigate bias by rebalancing biased labels in datasets (see Table 3).

To mitigate against bias in a CSAM dataset, the training dataset has to be representative of the relevant population, and this is complex to achieve, especially when there are numerous combinations of underlying attributes (such as gender, age, skin colour) all at play (IWF

[25] https://www.ibm.com/topics/overfitting

Funded by the European Union

expert, personal correspondence, 2023). The training process is an iterative one and requires layers of training, testing and then human assessment and interpretation of what the machine got wrong. This becomes the next iterative cycle and then the machine is re-trained to correct its errors. Another reason for continued testing of ML in a CSAM database is because the content of CSAM changes, for example, there has been a significant increase in so called 'self-generated images'(IWF, 2023) and in order for AI to be able to recognise these as a particular form of CSAM, the training model has to be updated to reflect the wider societal changes. The IWF mitigate potential human bias by ensuring images are reviewed by multiple people and quality assessment and assurance of CSAM reviewers' decisions (IWF expert, personal correspondence, 2023).

Funded by the European Union

A range of tools, methods or approaches were presented in the literature which offer potential solutions and ways to tackle and mitigate a range of types of bias. These are summarised in Table 3 below.

Table 3: Summary of current or potential methods to mitigate bias in [CSAM] datasets.

Author& Title	Type	Description	Type of bias	Authors evaluation
Wang et al., 2021. A bias-detection tool: REVISE (REvealing VIsual biaSEs)	A bias detection tool	A tool that assists in the investigation of a visual dataset, searching for potential biases along three dimensions: (1) object-based (relates to the size, context, or diversity of the depicted objects), (2) person-based (relates to metrics focus on analysing the portrayal of people, and (3) geography-based (relates to the representation of different geographic locations).	Object; Person; Geography	Effective at surfacing and helping mitigate different kinds of biases in visual datasets. Does not claim that REVISE will identify all visual biases.
Alassadi et al., 2022. A Generative Approach to Mitigate Bias in Face Matching using Learned Latent Structure.	A generative AI framework for countering bias.	A new framework for the fair matching of low-resolution and high-resolution facial images. Framework consists of two parts: a variational auto-encoder (VAE) that converts the images into their more generic underlying representation, and second, a neural network architecture that uses the above representations to undertake a multi-label classification.	Bias in face matching algorithms: Age and gender.	Considers that the proposed model can improve classification accuracy and reduce categorical bias compared with a baseline classical VAE architecture. Acknowledges that the binary nature of the variables is a limitation. The VAE can manage multiple sensitive attributes simultaneously.
Alelyani, 2021. 'Detection and Evaluation of Machine Learning Bias.'	A technique to detect and evaluate bias in ML models.	Proposes a wrapper bias detection technique based on a novel alternation function to detect and evaluate bias in ML in relation to human cognitive bias. The function is used to find the impact of potentially biased attributes on an ML model's prediction.	Gender and race.	The technique was shown to be able to detect and evaluate bias, but did not aim to justify the model. Results can be investigated further when a real-world model is built.
Yohannis & Kolovos, 2022. 'Towards Model-based Bias Mitigation in Machine Learning.'	A model-based approach to facilitate bias measurement and mitigation.	A toolkit, FairML, that implements a model-based approach to automate bias mitigation in ML. Enables the generation of bias mitigation code in a concise, declarative way.	General.	Used examples provided by IBM Fairness AI 360 as the baseline to assess the expressiveness of the tool, which was shown to reproduce all examples covering a range of datasets, classifiers and bias mitigation algorithms/ metrics. Shown to require fewer lines of code to produce comparably accurate measurement values to the ones produced by a baseline code. User evaluation by data scientists with experience in bias identification and mitigation was not sought.
Sevetlidis et al., 2022. 'Tackling Dataset Bias With an Automated Collection of Real-World Samples.'	A novel pipeline that automatically creates new thematic datasets with low bias.	The pipeline consists of several modules that collaborate to construct a dataset. It automatically constructs non-biased datasets by sampling many different sources. The pipeline uses a list of keywords to query Web-based resources, collect samples, and discard duplicate or irrelevant data which degrade the performance of classification tasks .	Dataset.	The model was demonstrated to build a dataset rapidly, scale well to multiple classes, and permit looser sample inclusion criteria by imposing a hard rule on excluding duplicate and irrelevant content. A different approach is needed where a dataset might contain classes of broad interest.
Cabrera et al. (2019).	FAIRVIS Visual Analytics	A visual analytics system for discovering intersectional bias in machine learning models.	Intersectional attributes bias	Discovered biases in two real datasets. Demonstrates how interactive visualization may help data scientists and

Summary

This review has presented a high-level overview and synthesis of the literature used to address the two research questions to be addressed by this review. Our aim has been to

Funded by the European Union

make this accessible to non-specialist and to assist in the development of the Protech project and not to explain the complex and specialist compute science behind the AI. Below we provide a brief summary of the findings for both the research questions.

Research question 1 summary:

- The exponential growth in CSAM far outstrips any capacity to tackle CSAM and several challenges face agencies who work in child protection.
- The re-victimisation of the children who are sexually abused every time an image shared and re-shared is a hidden side of CSAM.
- Automated detection techniques are effective and efficient methods to use to tackle CSAM and can reduce the harm to professionals who review and categorise images.
- A range of tools and techniques are being used to tackle CSAM.
- Taking a human-centred approach to AI and ML is important to be ethical and prevent harm.
- Ways to create and distribute CSAM will continue to evolve as new technologies develop.

Research question 2 summary:

- Researching and gaining an understanding of CSAM datasets is challenging because of the sensitive and illegal content.
- Bias is a broad, multi-disciplinary concept that can be classified in different categories, but a central concern is that the outcomes of AI models, in the real world, are free from bias and fair.
- AI and ML models can perpetuate and exacerbate biases inherent in datasets, which can result in outcomes that may discriminate against certain groups.
- There is a general consensus that bias in ML is primarily caused by the nature and quality of the training data.
- It is important to recognise the influence of human bias within ML datasets which reflect human behaviour, practice, experience and actions.

Limitations

There are some limitations to this review, which faced inherent constraints in its size, timescale, and the fact that CSAM datasets, by the nature of their content, cannot be publicly available to researchers. As a result, some other relevant literature may have been missed. There was a paucity of peer-reviewed research papers, and the majority of the results were drawn from grey literature that could not be rigorously quality assessed, and primarily proposed new, untested models or strategies to tackle CSAM.

Conclusion

The demand and volume for CSAM worldwide has reached unprecedented levels, and the burden placed on law enforcement and organisations is untenable. As the use of AI to tackle CSAM is increasing, caution must also be used to prevent and mitigate bias in CSAM datasets. We reviewed literature to address two research questions designed to give greater knowledge, understanding and insight in the current AI technologies currently used, and potential biases in CSAM datasets and ways to tackle it. This review has given an insight into

Funded by the European Union

the key issues relating to current AI tackling CSAM and ways to mitigate bias in datasets. The findings also highlight the need for agencies from different sections to work collaboratively to develop and train AI tools to tackle CSAM.

Funded by
the European Union

References

General references

Beier, K.M., Grundmann D., Kuhle L.F., Scherner G., Konrad A., and Amelung T., 2015. The German Dunkelfeld project: a pilot study to prevent child sexual abuse and the use of child abusive images. *The Journal of Sexual Medicine.* [e-journal] 12(2):529-542. doi:10.1111/jsm.12785.

Canadian Centre for Child Protection (CP3) 2023. Parents' perspectives on how child sexual abuse material impacts the entire family. System Failures, Resilience, and Recommendations for Change. Available at: https://www.protectchildren.ca/pdfs/C3P_FamilySurveyReport_en.pdf [Accessed 19 May 2023].

Cubitt, T., Napier, S., and Brown, R., 2021. Predicting prolific live streaming of child sexual abuse. *Trends & issues in crime and criminal justice*. [e-journal] 634 pp1-21 Available at: https://search.informit.org/doi/abs/10.3316/agispt.20211101056192 [Accessed on 15th May, 2023].

Das., S, Dey., A, Pal., A, and Roy., N., 2015. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications.* [e-journal] 115(9) Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=68508ffc9f75462fd31de620d03093b214734011> [accessed on 15th May 2023].

Du, J., Hargreaves, C.,Sheppard, J., Anda, F., Sayakkara, A., Le-Khac, N., and Scanlon, M., 2020. SoK: Exploring the State of the Art and the Future Potential of Artificial Intelligence in Digital Forensic Investigation. *The 15th International ARES Conference on Availability, Reliability and Security.* Virtual Event Dublin, June 25–28, New York, Association for Computing Machinery.

Elliott, I.A., Findlater, D. & Hughes, T., 2010. Practice report: A review of e-Safety remote computer monitoring for UK sex offenders. *Journal of Sexual Aggression,* [e-journal]16(2), pp 237-248, DOI: 10.1080/13552601003686870

European Union Agency for Law Enforcement Cooperation, 2020. *Exploiting isolation: Offenders and victims of online child sexual abuse during the COVID-19 pandemic.* [pdf] European Union Agency for Law Enforcement Cooperation. Available at: https://www.europol.europa.eu/cms/sites/default/files/documents/europol_covid_report-cse_jun2020v.3_0.pdf [Accessed on 15th May 2023].

Garritty C, Gartlehner G, Kamel C, King V.J, Nussbaumer-Streit B, Stevens A., et al., 2020. *Cochrane rapid reviews. Interim guidance from the cochrane rapid reviews methods group. [pdf]* Cochrane rapid reviews members group. Available at: http://methods.cochrane.org/sites/methods.cochrane.org.rapidreviews/files/uploads/cochrane_rr_-_guidance-23mar2020-final.pdf [Accessed on 15th May 2023].

Government Social Research Unit., 2008. *Government Social Research Rapid Evidence Assessment Toolkit.* [Online]. Available at:

Funded by the European Union

https://webarchive.nationalarchives.gov.uk/ukgwa/20140402164155/http:/www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment. [Accessed on 13 April 2023].

Hern, A. 2022. Sites reported record 29.3m child abuse images in 2021. *The Guardian,* [online] 24th March 2022, Available at:<https://www.theguardian.com/technology/2022/mar/24/sites-reported-record-293m-child-abuse-images-in-2021#:~:text=A%20record%2029.3m%20items,a%2035%25%20increase%20from%202020.> [Accessed on 16th November, 2022].

Higgins, J., & Green, S., 2008. *Cochrane handbook for systematic reviews of interventions*. [pdf] Chichester, UK: John Wiley & Sons Ltd. Available at: https://www.radioterapiaitalia.it/wp-content/uploads/2017/01/cochrane-handbook-for-systematic-reviews-of-interventions.pdf [Accessed on 15th November, 2022].

Higgins, J.P., Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, et al., 2019. *Cochrane Handbook for Systematic Reviews of Interventions* [e-book] John Wiley& Sons and The Cochrane Collaboration. Available at: https://books.google.co.uk/books?hl=en&lr=&id=cTqyDwAAQBAJ&oi=fnd&pg=PR3&dq=Cochrane+Handbook+for+Systematic+Reviews+of+Interventions&ots=tvnJw6AJig&sig=649X8IoRpSHlm5xVjalEEqLF5d0&redir_esc=y#v=onepage&q=Cochrane%20Handbook%20for%20Systematic%20Reviews%20of%20Interventions&f=false [Accessed on 15th March, 2023].

INHOPE, 2021. *Annual Report, 2020.* [online] Available at: https://inhope.org/media/pages/the-facts/download-ourwhitepapers/c16bc4d839-1620144551/inhope-annual-report-2020 [Accessed on 19h March, 2023].

Insoll, T., Ovaska, A. & Vaaranen-Valkonen,N., 2021. *CSAM Users in the Dark Web: Protecting Children Through Prevention. ReDirection Survey Report IWF.* [pdf] Suojellaan Lapsia ry: Protect Children. Available at: https://www.suojellaanlapsia.fi/en/post/csam-users-in-the-dark-web-protecting-children-through-prevention [Accessed on 15th May 2023].

Internet Watch Foundation, 2021. *Annual Report: International Trends* [online] Available at: https://annualreport2020.iwf.org.uk/trends/international/geographic [Accessed on 11th May 2023].

Internet Watch Foundation, 2022. *Annual Report: The Internet Watch Foundation.* [online] Available at: https://annualreport2021.iwf.org.uk/. [Accessed 9th May 2023].

Internet Watch Foundation., 2023. *A deep dive into the digital and emergency happening #BehindTheScreens, in childrens' bedrooms. The Annual Report 2022.* [online] Available at: https://www.iwf.org.uk/about-us/who-we-are/annual-report-2022/ [Accessed 8th May 2023].

Funded by the European Union

Kaminska, J., 2022. *8 types of data bias that can wreck your machine learning models.* [online] Available at: https://www.statice.ai/post/data-bias-types. [Accessed 9th May 2023].

National Centre for Missing and Exploited Children (NCMEC) 2023. 2022 CyberTipline Reports by Electreonic Service Providers (ESP). NCMEC. Available at: https://www.missingkids.org/content/dam/missingkids/pdfs/2022-reports-by-esp.pdf [Accessed 19 may 223].

Okoli, C., 2015 A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems* [e-journal] 37(43). DOI: 10.17705/1CAIS.03743

Panchenko, A., Beaufort, R., Fairon, C., 2012. *Detection of child sexual abuse media on p2p networks: normalization and classification of associated filenames.* [pdf] Proceedings of the LREC Workshop on Language Resources for Public Security Applications. Available at: https://www.researchgate.net/profile/Zygmunt-Vetulani/publication/233897801_Language_Resources_for_Public_Security_Applications_Needs_and_Specificities/links/53e3fe640cf2fb74870f6e14/Language-Resources-for-Public-Security-Applications-Needs-and-Specificities.pdf#page=32 [Accessed 9th May 2023].

Prichard, J; Wortley, R; Watters, P; Spiranovic, C; Hunn, C; Krone, T., 2022. Effects of automated messages on Internet users attempting to access 'barely legal' pornography. *Sexual Abuse: a journal of research and treatment* Vol. 34(1) 106–124, DOI: 10.1177/10790632211013809

Scherrer, A. & van Ballegooij, W., 2017. *Combating sexual abuse of children Directive 2011/93/EU*: European Implementation Assessment. [pdf] EPRS: European Parliamentary Research Service. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2017/598614/EPRS_STU%282017%29598614_EN.pdf [Accessed 14th May 2023].

Ratner, C., 2021. When "sweetie" is not so sweet: artificial intelligence and its implications for child pornography. *Family Court Review,* [e-journal] Vol. 59 No. 2, DOI: 386–401, doi: 10.1111/fcre.12576

Quayle, E., M. Vaughan, and Taylor. M. 2006. Sex offenders, Internet child abuse images and emotional avoidance: The importance of values. *Aggression and Violent Behavior* [online] 11 (1): 1–11. DOI: https://doi.org/10.1016/j.avb.2005.02.005

Quayle, E., 2020. Prevention, disruption and deterrence of online child sexual exploitation and abuse. *ERA Forum.* 21, 429–447. DOI: https://doi.org/10.1007/s12027-020-00625-7

United Nations., 2022. *The Core Principles of the AI for Safer Children Initiative*. [pdf] Available at: https://unicri.it/sites/default/files/2022-06/AI4SC%20Principles_0.pdf [Accessed 27th April 2023].

Vess, J., 2008. Sex offender risk assessment: Consideration of human rights in community protection legislation. *Legal and Criminological Psychology,* [online] Available at: https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1348/135532507X269804?casa_toke

Funded by the European Union

n=2Sr5NVJKup4AAAAA:We4eMNaffqbFrHvIvQOCRBHk0XVN4R1kPU4Zef5rr6IDLjT0ND3bZIPH9rMk21CZvmNF1suR7ojz_cjC [Accessed 30th April 2023].

Westlake, B., Bouchard, M. and Frank, R. 2012. Comparing Methods for Detecting Child Exploitation Content Online.In: Richard Frank School of Criminology And International Cybercrime Research Centre Simon Fraser University. *2012 European Intelligence and Security Informatics Conference.* Odense, Denmark. 22-24 August 2012. IEEE.

Westlake, B., Bouchard, M., Frank, R., 2017. Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation. *Sex. Abuse* [e-journal] 29 (7). DOI: https://doi.org/10.1177/1079063215616818.

Winder, P., Marsden, L. and Rotundo, R., 2023. *Automated Content Classification (ACC) Systems*. [pdf] Winder.AI. Available at:https://www.ofcom.org.uk/__data/assets/pdf_file/0016/252151/ACC-Phase-2-Report.pdf [Accessed 30 April 2023].

Wortley, R.K., Smallbone, S., 2006. *Child Pornography on the Internet.* [pdf] US Department of Justice, Office of Community Oriented Policing Services, Washington, DC. Available at:https://popcenter.asu.edu/sites/default/files/child_pornography_on_the_internet.pdf [Accessed 8 May 2023].

We Protect., 2021. *We Protect Global Threat Assessment.* [online] Available at:<https://www.weprotect.org/global-threat-assessment-21/#report> [Accessed 17th November 2022].

Websites:

Definition of Bias:

Statice., 2022. *8 types of data bias that can wreck your machine learning models.* [website] (date of update: May 18, 2022) Available at: <https://www.statice.ai/post/data-bias-types#:~:text=What%20are%20biases%20in%20machine%20learning%20and%20artificial%20intelligence%3F,systematic%20prejudice%2C%20and%20low%20accuracy [Accessed 17th October 2022].

References of the search results of Research Questions 1 and 2.

References of Studies for Research Question 1

Anda, F., Lillis, D., Kanta, A., Becker, B.A., Bou-Harb, E., Le-Khac, N. A., Scanlon, M., 2019. Improving Borderline Adulthood Facial Age Estimation through Ensemble Learning, in: Association for Computing Machinery Proceedings of the *14th International Conference on Availability, Reliability and Security, ARES '19.* , Canterbury, CA, United Kingdom, August 26 - 29, 2019 New York, NY, USA, Association for Computing Machinery.

Funded by the European Union

Bursztein, E., Clarke, E., DeLaune, M., Elifff, D.M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K., Bright, T., 2019. Rethinking the Detection of Child Sexual Abuse Imagery on the Internet, in: Liu & White,*The World Wide Web Conference, WWW '19.* San Francisco CA USA, May 13 - 17, 2019. New York, NY, USA, Association for Computing Machinery.

Kobayashi, H., Kadoguchi, M., Hayashi, S., Otsuka, A., and Hashimoto, M., 2020. In: IEEE. An Expert System for Classifying Harmful Content on the Dark Web. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI).* Arlington, VA, 9th November 2020. IEEE Press.

Laranjeira da Silva, C., Macedo, J., Avila, S., dos Santos, J., 2022. Seeing without Looking: Analysis
Pipeline for Child Sexual Abuse Datasets. In: *2022 ACM Conference on Fairness, Accountability,*
and Transparency, FAccT '22. Seoul, Republic of Korea, June 21-24th 2022. New York, NY, USA,
Association for Computing Machinery.

Lee. H-E., Ermakova, T., Ververisa, V., Fabiana, B., 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation* [e-journal] 34, 301022. DOI: https://doi.org/10.1016/j.fsidi.2020.301022

Martellozzo, E. and DeMarco, J., 2020. Exploring the removal of online child sexual abuse material in the UK: Processes and practice. *Crime Prevention and Community Safety* 22:331–350. DOI: https://doi.org/10.1057/s41300-020-00099-2

Moreira, D. C., Fechine, J. M., 2018. A Machine Learning-based Forensic Discriminator of Pornographic and Bikini Images. In: *2018 International Joint Conference on Neural Networks.* Rio de Janeiro, Brazil, 8th –13th July 2018. IEEE.
Mukherjee, S., Sederholm, T., Roman, A.C., Sankar, R., Caltagirone, S., Ferres, J.L., 2021. A

Machine Learning Pipeline for Aiding School Identification from Child Trafficking Images, in:
Proceedings of the Conference on Information Technology for Social Good, GoodIT '21. Rome, Italy,
9th-11th September 2021. New York, NY, USA, Association for Computing Machinery.

Triviño, J.M., Rodríguez, S.M., López, D.O.D. and Mármol, F.G., 2019, August. C3-sex: A chatbot to
chase cyber perverts. *In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl*
Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl
Conf on Cyber Science and Technology Congress. Fukuoka, Japan, 5th –8th Aug. 2019. IEEE.

Funded by the European Union

Ngo, V.M., Dang, C.N., Thorpe, C. & McKeever, S. 2022, Investigation, Detection and Prevention of Online Child Sexual Abuse Materials: A Comprehensive Survey. In: *2022 RIVF International Conference on Computing and Communication Technologies (RIVF).* Ho Chi Minh City, Vietnam, 2022. Dublin, Dublin Technological University.

Porto, L.F., Nascimento Correia, L. L., Franco, A., Pianto, D., Machado Carlos, E. P., and de Barros, V.F., 2020. Estimating sex and age from a face: a forensic approach using machine learning based on photo-anthropometric indexes of the Brazilian population. *International journal of legal medicine,* [e-journal] 134 (6), pp.2239-2259. DOI: 10.1007/s00414-020-02346-5.

Razi, A., Kim, S., Alsoubai, A., Stringhini, G., Solorio, T., De Choudhury, M., Wisniewski, P.J., 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online
Sexual Risk Detection. *Proceedings of the ACM on Human-Computer Interaction. 5(CSCW2) pp1-*
38. DOI: https://doi.org/10.1145/3479609

Scholey, A., and Zadeh, P. B. eds. 2022. A Digital Forensics Live Suspicious Activity Toolkit To Assist Investigators With Sexual Harm Prevention Order Monitoring. In: *2022 IEEE Conference on Dependable and Secure Computing (DSC).* Edinburgh, 22th –24th June 2022. IEEE.

Struppek, L., Hintersdorf, D., Neider, D., Kersting, K., 2022. Learning to Break Deep Perceptual
Hashing: The Use Case Neural Hash, in: *2022 ACM Conference on Fairness, Accountability, and*
Transparency, FAccT '22. Seoul Republic of Korea June 21 - 24, 2022. New York, NY, Association for Computing Machinery.

Industry websites and grey literature:

Safer, 2020. Protect your platform with the industry-leading CSAM detection platform built by experts in child safety technology. [online] Available at: <https://safer.io/how-it-works/> [Accessed on 16th February 2023].

Protecting Children Google, 2023. Fighting child sexual abuse online. [online] Available at: < https://protectingchildren.google/#introduction> [Accessed on 18h February 2023].

Todorovic, N,., and Chaudhuri., A., 2018. Using AI to help organizations detect and report child sexual abuse material online. *Google in Europe.* [blog] 3rd September, 2018. Avalible at: <https://www.blog.google/around-the-globe/google-europe/using-ai-help-organizations-detect-and-report-child-sexual-abuse-material-online/> [Accessed on 4th February 2023].

Funded by the European Union

Protecting Children Google, 2023. Discover our child safety toolkit. [online] Available at: <https://protectingchildren.google/tools-for-partners/#learn-about-our-tools> [Accessed on 8th February 2023].

Amazon AWS, 2023. What is Deep Learning? [online] Available at: <https://aws.amazon.com/what-is/deep-learning/#:~:text=Deep%20learning%20is%20a%20method,produce%20accurate%20insights%20and%20predictions> [Accessed on 8th May 2023].

IBM, 2023. What is natural language processing? [online] Available at: <https://www.ibm.com/topics/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can-> [Accessed on 6th May 2023].

Microsoft, 2023. Help stop the spread of child exploitation images and protect your business. [online] Available at: <(https://www.microsoft.com/en-us/photodna/cloudservice).> [Accessed on 6th January 2023].

References of Studies for Research Question 2

Ahmed, S., Athyaab, S.A., Muqtadeer, S.A., 2021. Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach, In: *proceedings of the 6th international conference on inventive computation technologies (icict 2021).* Coimbatore, India, 20th –22nd January 2021. IEEE.

Alelyani, S., 2021. Detection and Evaluation of Machine Learning Bias. *Applied sciences* [e-journal]. *1(14), DOI:* https://doi.org/10.3390/app11146271

Alasadi, J., AlHilli, A., Atrey P. K. and Singh, V. K. eds. 2022. A Generative Approach to Mitigate Bias in Face Matching using Learned Latent Structure - *2022 IEEE Eighth International Conference on Multimedia Big Data*. Naples, Italy, 5th – 7th December 2022. IEEE.

Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern and D. H. Chau, FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In: *2019 IEEE Conference on Visual Analytics Science and Technology (VAST),* Vancouver, BC, Canada, 20th – 25th October, 2019. IEEE.

Chakraborty, J., Majumder, S., Menzies, T., 2021. Bias in Machine Learning Software: Why? How? What to Do*?, in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*.

Funded by the European Union

Athens,
Greece, 23rd – 28th August. New York, NY, USA, Association for Computing Machinery.

Chen, C., Sundar, S.S., 2023. Is This AI Trained on Credible Data? The Effects of Labeling Quality
and Performance Bias on User Trust, in: P*roceedings of the 2023 CHI Conference on Human Factors*
in Computing Systems, CHI '23. Hamburg, Germany, 23rd– 28th April. New York, NY, USA. Association for Computing Machinery.

Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L., 2018. Measuring and Mitigating Unintended
Bias in Text Classification, i*n: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and*
Society, AIES '18. New Orleans LA USA February 2th – 3th, 2018. New York, NY, USA, Association for Computing Machinery.

Jiang, H. and Nachum, O., 2020, June. I*dentifying and correcting label bias in machine learning.* [pdf]
In International Conference on Artificial Intelligence and Statistics. Available at: <http://proceedings.mlr.press/v108/jiang20a/jiang20a.pdf> [Accessed on 1st April 2023].

Laranjeira da Silva, C., Macedo, J., Avila, S., dos Santos, J., 2022. Seeing without Looking: Analysis
Pipeline for Child Sexual Abuse Datasets, in: *2022 ACM Conference on Fairness, Accountability, and*
Transparency, FAccT'22. Seoul Republic of Korea June 21 – 24, 2022. New York, NY, USA, Association for Computing Machinery.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., Tang, J., 2022. Trustworthy AI:
A Computational Perspective. *ACM Transactions on Intelligent Systems and Technology* [e-journal]
14(1) pp 1– 59. DOI: https://doi.org/10.1145/3546872

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computer Surveys.* [e-journal] 54(6), pp1-35. DOI:https://doi.org/10.1145/3457607

Milaninia, N. 2020. Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications. *International Review of the Red Cross,* [e-journal] 102 (913), pp.199-234. DOI: 10.1017/S1816383121000096.

Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Cruz, G.O.R., Peixoto, R.M., Guimarães, G.A.D.S., Santos, L.L.D., Araujo, M.M., Cruz, M., de Oliveira, E.L.S. and Winkler, I., 2022. *Bias and*

Funded by the European Union

unfairness in machine learning models: a systematic literature review. [pdf] Cornell University, USA. Available at: < https://doi.org/10.48550/arXiv.2202.08176> [Accessed on 3rd February 2023]

Roselli, D., Matthews, J., Talagala, N., 2019. Managing Bias in AI, in: *Companion Proceedings of*
The 2019 World Wide Web Conference, WWW '19. San Francisco, USA, May 13th– 17th 2019. New
York, NY, USA, Association for Computing Machinery.

Sevetlidis, K. R. ., Pavlidis, G., Mouroutsos, S., and Gasteratos, A. 2022. *T*ackling Dataset Bias
With an Automated Collection of Real-World Samples. *IEEE Access.* [e-journal]. 10 pp.126832
126844. DOI:10.1109/ACCESS.2022.3226517

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A. and Russakovsky, O., 2022. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision,* [e-journal] 130 (7), pp.1790–1810. DOI: 10.1007/s11263-022-01625-5.

Yohannis, A., Kolovos, D., 2022. Towards Model-Based Bias Mitigation in Machine Learning, in:
Proceedings of the 25th International Conference on Model Driven Engineering Languages and
Systems, MODELS '22. Montreal Quebec Canada, October 23th – 28th 2022. New York, NY, USA,
Association for Computing Machinery.

Funded by the European Union

Appendices

Appendix 1: Search strings:

(Combinations of)

Search 1

"AI" OR "artificial intelligence" OR "machine learning" AND "child sexual abuse material" OR CSAM OR "child sexual exploitation" OR ~~CSEM~~ OR "indecent image" OR "IIOC" OR "child* porn*" AND detecting OR removing OR tackling AND "CSAM datasets" OR "child* sex* abus* material*") AND ("indecent image*") OR IIOC OR "child* porn*" AND dataset*

Search 2:

Bias* AND "CSAM datasets" OR "child* sex* abus* material* datasets" OR "indecent image*" OR IIOC OR "child* porn*" AND dataset* AND "bias in machine learning" OR "bias in artificial intelligence" AND "online child sexual abuse" AND "~~recommendations OR mitigations AND minim~~* OR reduc* AND "effect* of bias" AND datasets"

Appendix 2: Table 1: Eligibility criteria (Inclusion/exclusion)

Criteria and parameters	Include	Exclude
Publication type	Peer-reviewed articles Systematic reviews Books or edited chapters Grey literature from academic, private, public and third-sector organisations AI relevant websites	Non-peer-reviewed papers Media articles; opinion pieces, book reviews.
Published in English language	Reference to any country as long as it is relevant to purpose of review	Non-English language
Publication date	2018 to 2023	Outside of these dates
Search 1: Criteria is focused on the content and focused on the intersection of (any) AI technology AND any aspect of CSAM (detection, prevention etc).	Reference to AI technology to tackle any forms of CSAM	Must contain BOTH
	AI technologies / solutions to tackle any aspect of CSAM (prevention; production; sharing; use for grooming etc)	

Funded by the European Union

Search 2: Current and ongoing research into bias in CSAM data sets, recommendations and mitigation on how to minimalise the effect of bias	Original criteria (C1 to C3) **C1.** Reference to any forms of AI / ML **AND** (any) forms of online CSAM **C2.** Bias / fairness / trust in datasets AI / Machine Learning / Deep learning **C3**.Ways to mitigate, tackle, reduce the effects of bias in CSAM datasets.	If do not meet C1 And/or C2 And/or C3
Criteria adjusted due to paucity of literature that contained BOTH	Adjusted criteria (C1 to C3) **C1**. Bias / fairness / trust in datasets AI / Machine Learning / Deep learning **C2**. Ways to mitigate, tackle, reduce the effects of bias in CSAM datasets. **C3.** Reference to any forms of AI / ML **AND** (any) forms of online CSAM	If do not meet C1 And/or C2 And/or C3

Appendix 3: Examples of reasons for exclusions

Reasons for exclusion against inclusion criteria	• Little reference to CSAM specifically • The article is about toys; little mention of prevention; no mention of CSAM • No reference to AI • It's about adult content, not CSAM • Limited mention of bias in dataset in terms of AI • Looks at *quality* of AI technologies, not biases • Doesn't offer a new or innovative solution
Reasons for exclusion against the Risk of Bias/ QA tool	• Reporting not consistently clear • Sample details are unclear. • Analysis is not clear

Appendix 4: Risk of Bias/Quality Assessment screening tool

PROTECH systematic literature review: Risk of Bias/Quality Assessment screening tool

Some material included in the full-text screen are not research papers (e.g. systematic literature reviews) so not all of the following criteria apply, particularly Questions 2 and 3. In these cases please use 'n/a'.

The exclusion criteria:

For research papers = 'When a study has 2 x No assessments OR, 1 x No AND 2 Can't Tell'.
For review papers when n/a is applied = 'When a study has 1 x No assessments OR, 1 x No AND 1 Can't Tell' (to compensate for the fewer criteria).

Screening papers:

1a). Is there a clear statement of the research aims?

- Yes or No.
- The answer to 1a) does not influence whether a study meets the threshold for inclusion/exclusion. The answer to this question is only required for question 1b).

1b). Is the chosen research methodology appropriate given the stated research aims?

- If the answer to 1a) is NO, score this question (1b) as Can't Tell.
- If the answer to 1a) is Yes, and you assess that the methodology is appropriate for the stated research aims then score as 'Yes'.
- If the answer to 1a) is Yes but you assess the methodology as not appropriate for the stated research aims, then score as 'No'.

2. Was the sampling strategy likely to select participants that are representative of the target population?

- Yes
- No
- Can't tell
- N/A if the peer-reviewed paper is not a research paper – i.e. a systemic review.

3. Given the research design, is the sample size appropriate?

- Yes
- No
- N/A if not a research paper

4. Has the chosen research design been conducted appropriately (e.g. ethics etc)?

- Yes
- No

[Additional guidance: Were there any issues with the data collection methods? Is it clear how it was collected? Was the setting for collection justified? Were the methods justified?].

5. Are the methods of analysis appropriate?

- Yes
- No
- Can't tell

[Additional guidance: in the case of a review, please check that analysis of the contents under review are appropriate].

6. Were results clearly presented for all of the analyses described in the methods?

- Yes
- No

7. Where applicable, are conclusions supported by results?

- Yes
- No

Funded by the European Union