

INNOVATION THOUGHT LEADERSHIP



PROTECTING CHILDREN AGAINST ONLINE HARMS USING AI

DR S. PAGE | CTO FUTURES

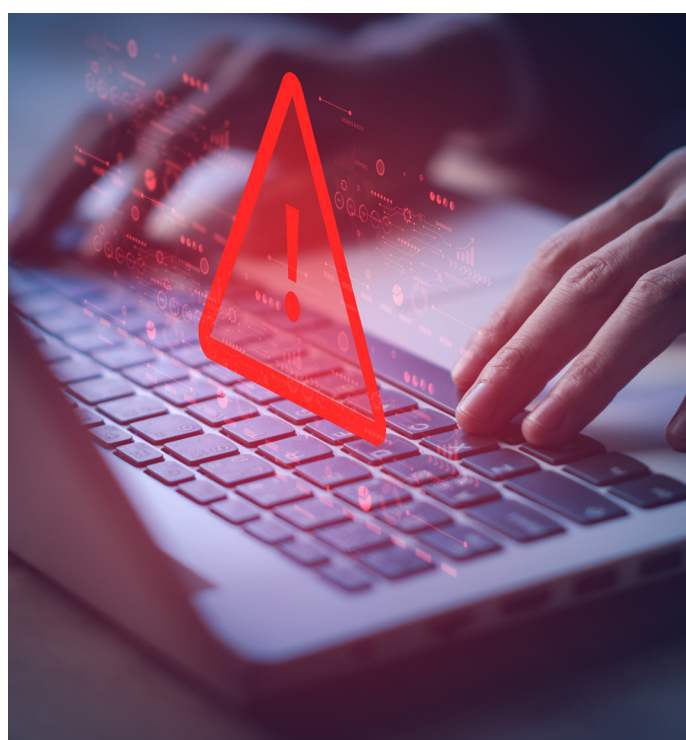
On the 19th September 2023, in what has been described as a “momentous day for children”¹, the Online Safety Bill passed after many years of discussion, marking the start of a new era for managing online harms. Whilst this is a key moment in the mission to bring about stronger protection for some of the most vulnerable in society, there is still a great challenge ahead. To set the context, NSPCC research published in August this year showed that while the Bill was being discussed in parliament over the last 5 years, there has been a staggering 82% increase in online grooming crimes and a 66% increase in child abuse image crimes.

Roke is proud to be actively engaged in the fight against Child Sexual Exploitation (CSE) through its delivery of the pioneering Vigil AI CAID Classifier to the law enforcement community. The Vigil AI CAID Classifier is a powerful technology which uses AI to detect and categorise the severity of Child Sexual Abuse Material (CSAM), including material that is previously unknown (i.e., undiscovered) and deepfake CSAM. In this whitepaper we explore the role of technology like Vigil AI in combatting CSE online harms and supporting the goals of the Online Harms Act, as well as the next set of challenges that Roke is preparing to help overcome.

CONTENTS



- 1 THE PROBLEM
- 2 USING AI TO DETECT CSAM
- 3 APPLICATIONS AND BENEFITS – POLICING
- 4 APPLICATIONS AND BENEFITS – INDUSTRY TECH PLATFORMS
- 5 MOVING FORWARD



THE PROBLEM



In 2023, it is estimated that over 1.8 trillion images have been created globally – approximately 5 billion per day, representing a 65% increase over 2021.

A large proportion of this media content is uploaded and stored on global digital platforms, from social media to file storage. The rise of generative AI and so-called “deepfake” imagery and video, which we will discuss more later in this paper, will have a further and likely significant impact on these statistics – it is thought that generative AI has already created over 15 billion images, more than photographers have taken in over 150 years. In addition to the increase in production of imagery and video, the use of live streaming services is also increasing, with over 7.5 billion hours of livestream video being watched in the third quarter of 2023, with over 150 million live streaming watchers in the US alone.



With this ever-growing increase in content generation comes an ever-increasing CSE threat and many argue that the true scale of the internet’s CSAM problem is largely unknown. For law enforcement, the dangers of online harms are not new; policing groups all over the world have been investigating online crimes since the advent of the internet.

A sub-set of this policing community specialises in countering CSE, a set of crimes which often includes the production of CSAM and the publishing and sharing of this content online. Thousands of officers all over the world are, at this very moment, reviewing “cybertips”², investigating potential CSE crimes, working tirelessly to safeguard victims and bringing offenders to justice, supported by a collection of dedicated industry partners.

The rise of generative AI, deepfakes, and more recently, CSAM deepfakes has compounded the problem that society faces. As many in law enforcement and the child protection industry feared, open-source generative AI models are now being used to create new CSAM content – sometimes using a base-image of a real person or child which the model transforms into sexualised content.

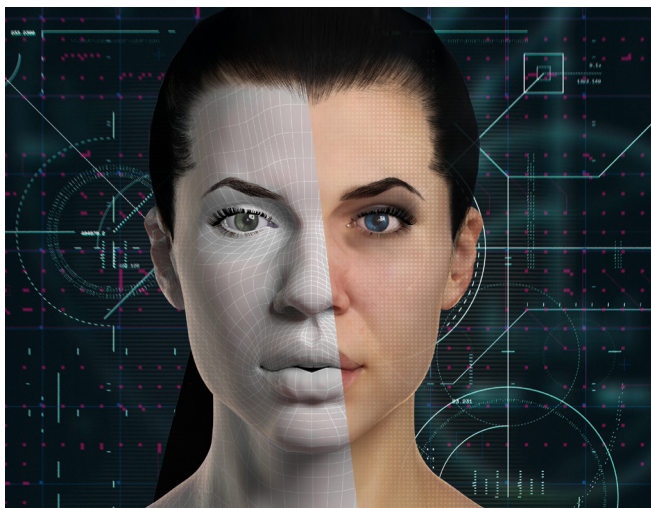
¹<https://www.nspcc.org.uk/about-us/news-opinion/2023/2023-09-19-the-online-safety-bill-has-been-passed-in-a-momentous-day-for-children/>

²Cybertips are reports provided by members of the public or technology platforms signalling the potential existence of CSAM – these reports are typically managed by regional NGOs as the US-based National Centre for Missing and Exploited Children (NCMEC) or the Internet Watch Foundation (IWF) in the UK before being passed on to policing

Over the last few months, the news has been peppered with reports of increasing evidence of deepfake CSAM production³, most recently including a worrying report of children using generative AI models to create CSAM of other children using “de-clothing” apps⁴.

A recent Internet Watch Foundation (IWF) report⁵ regarding generative AI CSAM confirms that what many regarded as a potential future problem, is, very sadly, already a reality. The obvious negative effect is that the volume of CSAM will increase significantly, and likely lead to an even greater volume of industry referrals. Arguably, the greatest impact of deepfake CSAM lies in victim identification.

Law enforcement agencies around the world have specialist Victim Identification Departments (VID) who meticulously review and analyse previously unknown CSAM for clues to identify the location, offender and most importantly the victim.



If generative AI CSAM becomes indistinguishable from real CSAM, this will result in law enforcement spending valuable time identifying victims who do not exist and will inevitably reduce how quickly they can identify and rescue real victims from physical harm.

A major first step in UK society's attempt to combat online harms is the recent passing of the Online Harms Bill, which went on to receive royal ascent in October 2023, becoming part of UK law. The new Act creates important new obligations for tech platforms to actively manage the risk of online harms. Specific attention is drawn to illegal content, such as CSAM and terrorism-related material.

Here, the Act obliges organisations to ensure their platforms are not a conduit for CSAM; severe fines may face those that do not meet the required standard. The Act makes provision for “accredited technology”, for which the Vigil AI CAID Classifier provided by Roke is a potential candidate, being deployed where platforms require advanced solutions to help them identify illegal content.

This is an important aspect to the Act which recognises that there is a need to measure and assess the performance and capability of counter-CSAM solutions (e.g., image and video content classifiers) in order that tech platforms can get access to operationally viable and effective solutions. This in turn helps to ensure law enforcement is not overburdened with erroneous reports which can waste critical time.

³E.g. <https://www.bbc.co.uk/news/uk-england-cambridgeshire-67145583>

⁴<https://www.bbc.co.uk/news/technology-67521226>

⁵ Find link

2

USING AI TO DETECT CSAM



Technology has provided the infrastructure and opportunity for bad actors to commit online harms. It is imperative that technology is also harnessed to mitigate and combat these harms.

In the CSE domain, a number of technologies are used to detect CSAM and/or suspicious communications. The most established of such technologies is “hashing” - hashing is a term used to describe a range of methods for creating a unique signature for an image (or video) using a mathematical process. Once hashed, content can be compared or checked against databases of known CSAM.

A range of hash databases exist in policing, government entities and NGOs (e.g. NCMEC) across the world, as well as a number managed by industry, for example the Tech Coalition hash-set. Microsoft’s PhotoDNA hashing technology is the most commonplace in the private sector; PhotoDNA is a “perceptual” hash, meaning that it can detect matches or “hits” if the image has been slightly modified – e.g. cropped, rotated or resized.

Hash-based CSAM detection is fast and effective and is a powerful tool in combatting CSE. However, it is fundamentally limited as it can only detect content that has already been identified and hashed. This of course means that it cannot and will never identify new or unknown content, sometimes referred to as “first generation” content. Given the increase in content production and associated CSAM production it is easy to see how this technology alone is not and will not be an effective tool for combatting CSE in the online domain, even if the additional challenges of generative AI are put aside.

Thankfully, other solutions do already exist, most notably those making use of machine learning or AI-powered content analysis and classification, such as the Vigil AI CAID Classifier.



The Vigil AI CAID classifier is the result of a pioneering collaboration between UK industry and the UK Home Office Child Abuse Image Database (CAID) team. Originally conceived in 2016, it was, to the Author's knowledge, the world's first operationally viable CSAM classifier, able to not only detect CSAM but also to categorise or "grade" the severity of the content, according to established government standards.

The classifier is built by training state-of-the-art deep learning algorithms on the UK Home Office Child Abuse Image Database (CAID), a highly respected dataset which is meticulously analysed by highly trained police officers, providing some of the finest grained CSAM data in the world. Roke's team has researched and developed highly specialised training approaches for the classifier, resulting in very high levels of accuracy. The Home Office has deployed this technology in 40 police forces in the UK helping hundreds of officers rapidly analyse content from investigations and identify victims and offenders. The ability for the classifier to grade severity not only assists with evaluating risk but helps to manage the psychological welfare of the officers working on the cases, many of which can go on to suffer from PTSD, by helping to prepare them for exposure to disturbing content.

The Vigil AI CAID Classifier is pioneering in a number of respects. In the UK, Vigil is the first ever example of an industry-government engagement that has led to a fielded product trained on this type of data.

It has required significant levels of innovation from technical, policy and government partnering perspectives. From a data science perspective, training machine learning classifiers on imagery without looking at the underlying data and understanding its patterns is highly unconventional; it is a well-established fact that understanding your data is very often the most important aspect of building any machine learning solution.

To address this challenge, the Vigil team built novel bespoke tooling which allows it to work hand-in-hand with operational users (who are able to and trained to study the data) to capture the critical behaviours of the model as it was trained in order to provide a closed-loop on model performance. Fast forward 7 years to 2023 and this type of tooling is more commonplace, but in early 2016 this broke new ground. Adding to this challenge is the complexity of training on bulk, highly sensitive, illegal and legal data. A significant amount of work has gone into sourcing appropriate training datasets (by Policing and the Home Office as well as Vigil AI), and ensuring the data is protected both in training and in deployment of the technology.

The experiences and lessons learnt on this journey and the approach that has been developed pave the way for other similar technology to be created using highly sensitive datasets and exploited to combat growing online harms. Roke is now drawing on the expertise it has built working on Vigil to support the genesis of AI technology in other highly regulated markets.

3

APPLICATIONS AND BENEFITS – POLICING



“Before Vigil it was a case of going through hundreds or thousands of images manually; now the classifier helps us get to the images we need to prioritise which is really valuable.” (UK Police User, 2023)

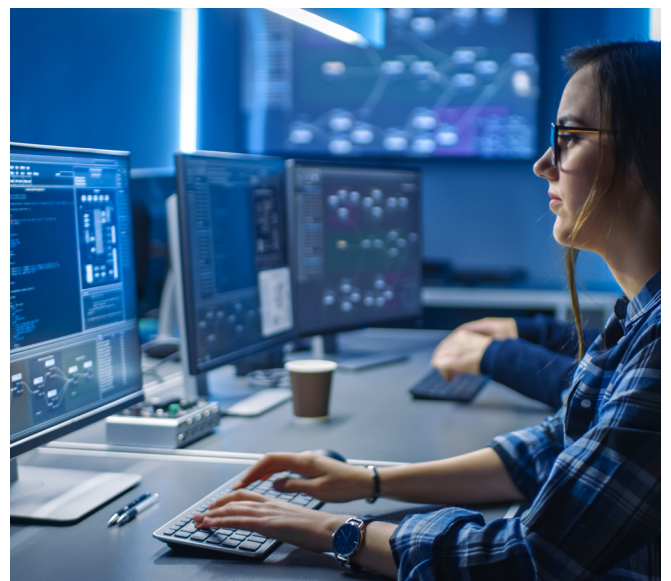
The classifier is used in three primary use-cases in policing. The first is in **Rapid Evidence Gathering** -here, the police may have someone in custody on suspicion of CSE related offences and there can be a need to very quickly establish if there is any evidence of any CSAM crimes and potentially victims that need rescuing on any seized devices.



However, seized devices can contain 100s of thousands or millions of images on them. Using the Vigil AI CAID Classifier to rapidly scan images on these devices allows officers to rapidly determine if there is any content of concern. The second and most common use-case is in **Forensic Analysis and Victim Identification**.

In this use-case the classifier is used to very quickly identify and remove benign content and legal 18+ pornography, leaving the investigator with content that is likely to be CSAM. It is also used to “pre-categorise” the data into levels of severity – this helps with risk assessment and gaining insights into the offender’s sexual interests, as well as supporting the “grading” process which ultimately has an influence on sentencing⁶.

Users are able to sort or filter very large collections of content so that certain types of CSAM content are brought right to the top of the stack. Finally, the classifier is used in **Large Scale Dataset Analysis** and in supporting bulk media categorisation.



⁶Note, Vigil is deployed as an operator assist tool – all images are manually verified by trained police officers

3

By using the classifier in this way, a number of important outcomes are being achieved:

- **FASTER DATA ANALYSIS AND REDUCED BACKLOGS**

Officers are reporting up to 50-80% time savings on cases where the classifier is used – leveraging its ability to bring important content to the top of the stack

- **VICTIMS BEING IDENTIFIED AND RESCUED EARLIER**

Ultimately, the classifier supports more rapid victim identification. In a recent case in 2023, the classifier identified an image that had so far been undiscovered, and this image led directly to the identification of the victim

- **SUPPORT TO PSYCHOLOGICAL WELFARE**

As discussed above, the classifier’s ability to grade imagery by severity can support officers with preparing themselves for what they are about to view

**“It allows officers to make those sort of fast time decisions about safeguarding rather than a couple of days later or weeks later when they have had time to sort through large numbers of images”
(UK Police User, 2023)**



APPLICATIONS AND BENEFITS – INDUSTRY TECH PLATFORMS

4



Outside of law enforcement, Vigil can provide benefits to a wide range of technology platforms including content service providers, moderation service providers, social media platforms, gaming and file-hosting platforms, or indeed any platform that facilitates the hosting or streaming of end-user generated content.

The classifier has been designed to have a very simple interface, requiring only access to the media to be classified, and providing a set of simple and clearly understandable output scores for each image or video - notably CSAM vs non-CSAM and, optionally, severity gradings and scores for legal adult pornography etc.

It is further designed for scalability, available as a “containerised” solution which allows platforms to plug it into scalable cloud architectures that automatically spin-up and shutdown computing nodes depending on load. As a result, the classifier can be used across platforms of any scale.

The design of the classifier allows for processing of millions of images a day on a single computer, and integration can be approached from a number of angles from full-scale processing of all content to more targeted detection schemes. This allows platforms to exploit Vigil without a material impact on their systems and the end-user experience

The benefits to platforms of using technology like Vigil in addition to hash-based detection are many, and there are strong parallels with the law enforcement applications.

Vigil helps platforms **secure their own systems and prevent them from being used as a conduit for CSAM** which in turn helps to **protect their brand, and reduce the risk of costly investigation, fines and legal prosecution.**

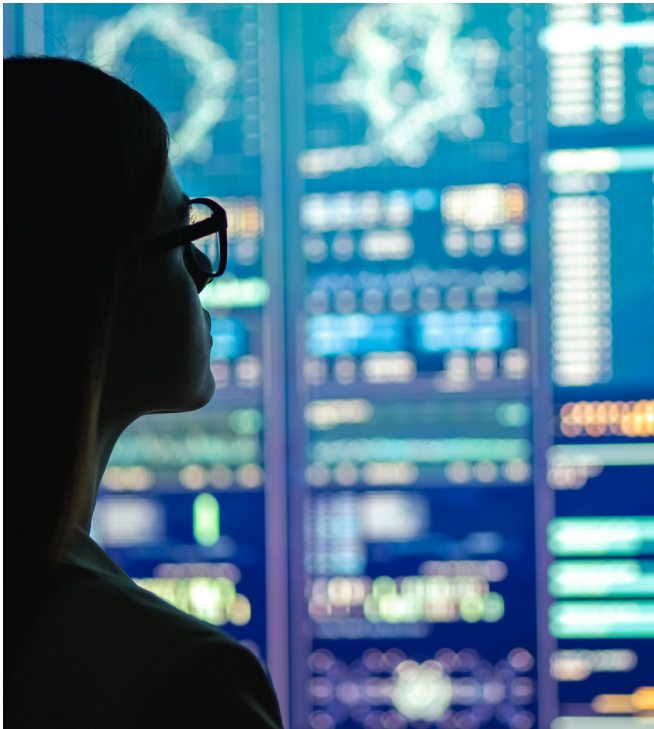
In the worst cases, a failure to combat CSE related crimes on a tech platform could potentially lead to a platform being shutdown, as is believed to be the case with the recent closure of the Omegle platform⁷ in 2023, and in the near shutdown of the Kik platform in 2019 where the platform was allegedly linked to CSE crimes⁸.



⁷<https://www.washingtonpost.com/technology/2023/11/09/omegle-chat-app-shutdown/>

⁸ <https://www.vice.com/en/article/43k4dw/kik-had-a-huge-child-predator-problem-now-its-shutting-down>

4



As with policing, Vigil helps identify illegal content of CSAM rapidly, bringing it to the moderation team's attention fast and allowing for rapid reaction and reporting. For platforms with moderation backlogs it can **bring the most severe content to the top of the stack, so that it can be prioritised accordingly** so that users, groups or parts of the platform can be stopped.

The ability of the classifier to grade severity can also be used to route images through different moderation system pathways, for example to different teams who may have had different training or employment terms, or simply to help manage exposure.

There has been a raft of class-action lawsuits⁹ over the last few years from employees working for moderation companies or in moderation teams – often outsourced to foreign countries with lower-paid workforces; here, numerous claims have been made about exposure to horrific content including CSAM, beheadings and suicide related material, without adequate, if any, psychological welfare and protection measures.

Tech platforms seeking to take a level of ethical responsibility for the end-to-end operation of their moderation efforts can consider the use of tools such as Vigil to **manage exposure to highly damaging content**.

Ultimately, deploying technology like Vigil also helps support the mission to protect our children and **rescue victims and prosecute victims in a timely fashion**.



⁹E.g. <https://www.theguardian.com/technology/2023/aug/16/sama-ceo-regrets-firm-took-on-facebook-moderation-work-kenya-staff-allege-exposure-graphic-content>

MOVING FORWARD



Over the coming months and years, OFCOM, which is tasked with policing the Act, will need to establish the necessary processes such that solutions can be accredited, and to understand if platforms are making use of effective solutions..

There is some precedent for such activity – in the 2000s, the UK Government establish an accreditation process for CCTV Video Analytics systems, “iLids” to help critical national infrastructure and other potential end-user organisations to identify those systems that could be relied upon. Importantly, the design of these accreditation processes requires an in-depth understanding of the scientific evaluation of machine learning technology as applied to CSAM detection and also the context in which such systems will be deployed.

Aspects such as content throughput, integration pathways and suitability for cloud-based scaling (and associated cost and energy footprints) are major factors for platforms and must not be ignored. Further, security must be a critical consideration. Roke is actively engaging with the community to bring its broad consultancy expertise and detailed understanding of the CSAM detection domain to bear in this area.

Roke is continuing to develop the Vigil AI CAID Classifier to improve its accuracy and speed. Roke’s Vigil team includes highly experience machine learning engineers and software engineers who are applying state-of-the-art machine learning architectures and approaches and supporting law enforcement with exploiting this important solution.

Roke is investigating the use of homomorphic encryption with a view to supporting high-accuracy classification on End-to-End Encrypted (E2EE) communications. Roke is also developing a Vigil-based solution for addressing the growing problem of CSAM live streaming.



5


However, with a live stream, a decision should, ideally, be made in real-time as soon as the stream becomes illegal – in order that the platform can shutdown the stream or take other appropriate action. Roke has developed a prototype solution which is able to detect CSAM content in live streaming video in real-time across multiple streams whilst maintaining high levels of precision and recall and is seeking industry partners to trial and test the solution.

Society is beginning to waken to the scale and potential damage of online harms, but to many it still represents a “happens to someone else” problem.

Amidst the news reports, debates on E2EE and privacy, tech company shutdowns, and moderation class-action lawsuits lies the true, and often unspoken tragedy – the victims and their experiences, which are all too often relived through the re-sharing of CSAM content. Roke is committed to continuing its mission to develop and deploy the most powerful AI technology to help create a safer world for our children; for us, there is simply no more important application for AI.

If you would like to learn more about our work in this important area, please contact info@roke.co.uk





We believe in improving the world through innovation. We do it by bringing the physical and digital together in ways that revolutionise industries.

That's why we've fostered an environment where some of the world's finest minds have the freedom, support and trust to succeed.

Roke is a team of curious and deeply technical engineers dedicated to safely unlocking the economic and societal potential of connected real-world assets. Our 60 year heritage and deep knowledge in sensors, communications, cyber and AI means our people are uniquely placed to combine and apply these technologies in ways that keep people safe whilst unlocking value. For our clients, we're a trusted partner that welcomes any problem confident that our consulting, research, innovation and product development will help them revolutionise and improve their world.

If you're bringing the physical and digital worlds together, we'd love to talk.

Roke Manor Research Ltd
Romsey, Hampshire, SO51 0ZN, UK

T: +44 (0)1794 833000
info@roke.co.uk www.roke.co.uk

© Roke Manor Research Limited 2023 • All rights reserved.

This publication is issued to provide outline information only, which (unless agreed by the company in writing) may not be used, applied or reproduced for any purpose or form part of any order or contract or be regarded as representation relating to the products or services concerned. The company reserves the right to alter without notice the specification, design or conditions of supply of any product or service.