Technology and Regulation

Rethinking Safety-by-Design and Techno-Solutionism for the Regulation of Child Sexual Abuse Material

Author(s)	M.R. Leiser & Andrew D Murray
Contact	m.r.leiser@digidata.uk, a.murray@lse.ac.uk
Affiliation(s)	Mark Leiser is founder of DigiData, Ltd. (United Kingdom) and Visiting Professor at the Riga Graduate School of Law (Latvia). Andrew D. Murray is Professor of Law, The LSE Law School, London School of Economics.

Keywords	child sexual abuse material, CSAM, cyber-regulatory theory, safety-by-design, regulatory theory, Online Safety Act, CSAM Proposal
Published	**Received**: 19 Nov 2024 **Accepted**: 17 Apr 2025 **Published**: 04 Jun 2025
Citation	M.R. Leiser & Andrew D Murray, Rethinking Safety-by-Design and Techno-Solutionism for the Regulation of Child Sexual Abuse Material, Technology and Regulation, 2025, 137-171 • 10.71265/nga7v921 • ISSN: 2666-139X

Abstract

This article examines the rise of technological solutions to digital regulatory challenges, with a focus on Child Sexual Abuse Material (CSAM) and the imposition of obligations on platforms to mitigate risks while safeguarding fundamental rights. This leads to new regulatory designs, such as "safety-by-design," which is favoured by European regulators due to its cost-effectiveness and efficiency in assigning responsibilities to online gatekeepers. We examine the European Union's CSAM Proposal and the United Kingdom's Online Safety Act, two ambitious initiatives that aim to utilise technology to combat the dissemination of CSAM. This proposal mandates platforms to perform risk assessments and implement mitigation measures against the hosting or dissemination of CSAM. In cases where these measures fail, a detection order can be issued, requiring platforms to deploy technical measures, including artificial intelligence (AI), to scan all incoming and outgoing communications. This approach, while well-intentioned, is scrutinised for its potential over-reliance on technology and possible infringement of fundamental rights. The article examines the theoretical underpinnings of "safety-by-design" and "techno-solutionism," tracing their historical development and evaluating their application in current digital regulation, particularly in online child safety policy. The rise of safety-by-design and techno-solutionism is contextualised within the broader framework of cyber regulation, examining the benefits and potential pitfalls of these approaches.

We argue for a balanced approach that considers technological solutions alongside other regulatory modalities, emphasising the need for comprehensive strategies that address the complex and multifaceted nature of CSAM and online child safety. It highlights the importance of engaging with diverse theoretical perspectives to develop effective, holistic responses to the challenges posed by CSAM in the digital environment.

1. Introduction

A contemporary trend in platform governance is to impose obligations on private actors to mitigate the risks associated with providing their services. This leads to new regulatory design trends and legal requirements such as digital "safety-by-design" (hereafter, SbD).[1] As it reduces public enforcement costs by imposing obligations on private actors, this approach has found favour among regulators. The EU's Digital Services Act[2] (hereafter DSA), Australia's e-safety law (e-Safety Act)[3] and the UK's Online Safety Act (OSA)[4] rely heavily on technical solutions to solve or mitigate some of the more challenging aspects of digital platformisation.[5] A common theme is the reliance on techno-solutionism – the belief that technology, particularly digital solutions, can address various social, economic, and political challenges. This often oversimplifies complex problems by presuming that there is a straightforward technological fix.

A striking example is the EU's recent proposal to prevent the dissemination of child sexual abuse material (CSAM) (including grooming solicitation): the CSAM Proposal.[6] The proliferation of CSAM is undoubtedly harmful and requires regulation.[7] The CSAM proposal is an ambitious attempt to deploy technology to assist in stopping the considerable harm associated with child abuse.[8]

The CSAM proposal requires platforms to assess and mitigate risks associated with hosting, disseminating, or facilitating CSAM, as well as grooming.[9] A detection order can be issued when those risk mitigation measures are unsuccessful.[10] This requires the platform to deploy various technical measures, including AI provided by a newly created EU Center, to scan all messages (thus, in theory, building a backdoor into encryption) and private communications.[11] This is justified by claims that AI and 'indicators' will get more competent at identifying attempts at grooming via messaging systems, and technical solutions are required to protect children from abuse.[12] Critics point out that the mass infringement of privacy and data protection

1. C. M. Kelty, 'Beyond implications and applications: the story of "safety by design"' (2009) 3 *NanoEthics* 79.
2. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1.
3. Online Safety Act 2021 (Cth) No 76 https://www.legislation.gov.au/Details/C2021A00076 accessed 10 March 2025..
4. Online Safety Act 2023 https://www.legislation.gov.uk/ukpga/2023/50 accessed 10 March 2025; see also Aotearoa New Zealand Code of Practice for Online Safety and Harms (25 July 2022) https://thecode.org.nz/wp-content/uploads/sites/38/2023/06/THE-CODE-DOCUMENT-FINAL.pdf accessed 12 May 2025
5. European Commission, 'The Digital Services Act package' (Digital Strategy, 2024) https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package accessed 12 May 2025; eSafety Commissioner, 'Online Safety Industry Standard 2024' (Digital Policy Alert, 2024) https://digitalpolicyalert.org/change/10098 accessed 12 May 2025; Ofcom, 'Britain sets first codes of practice for tech firms in online safety regime' (Reuters, 16 December 2024) https://www.reuters.com/world/uk/britain-sets-first-codes-practice-tech-firms-online-safety-regime-2024-12-16/ accessed 12 May 2025.
6. European Commission, Proposal for a Regulation of the European Parliament and of the Council on preventing and combating child sexual abuse (December 16, 2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A209%3AFIN accessed 10 March 2025.
7. M. Leiser & S. Witting, '2nd expert workshop on the CSAM Proposal at VU-Amsterdam' [Unpublished workshop report, March 2023]. Vrije Universiteit Amsterdam.
8. *Ibid*.
9. Article 3(1) CSAM Proposal
10. Article 7(1) CSAM Proposal
11. Recital 26, CSAM Proposal; European Digital Rights, 'A Safe Internet for All: Upholding Private and Secure Communications' (Position Paper, 2022) 51 https://epicenter.works/fileadmin/import/edri-position-paper-csar.pdf accessed 10 March 2025.
12. J. Street et al, 'Enhanced Online Grooming Detection Employing Context Determination and Message-Level Analysis' (arXiv, 12 September 2024) https://arxiv.org/abs/2409.07958 accessed 10 March 2025.

rights is disproportionate to the harm and that the technology is not good enough to accurately determine unknown CSAM and what amounts to the solicitation of a child.[13] Consequently, law enforcement may chase false positives or allow false negatives unchecked.[14] Others have passionately argued that the economic drivers and cultural considerations behind the production of CSAM require a far more holistic response than a technology-based solution alone.[15]

This article addresses the following question:

> Are the EU's CSAM Proposal, the United Kingdom's Online Safety Act, and other safety-by-design approaches to protecting children overly techno-deterministic? Can a more normative approach be determined by examining Cyberpaternalist, Network Communitarian, gatekeeper, and risk-based regulation theories?

Part I examines the resurgence of this safety-by-design and techno-solutionism, offering a nuanced understanding of both. We begin by examining their development, setting the stage for an in-depth analysis of their underlying principles and emerging significance in contemporary digital safety discourse. This section culminates in a critical evaluation of the rationale and role of safety-by-design, as well as the re-emergence of techno-solutionism, particularly in safeguarding children online. Part II focuses on their application within legislative frameworks. We scrutinise the techno-solutionism inherent in the European Union's CSAM Proposal, followed by an assessment of the UK's Online Safety Act. This comparative analysis extends to evaluating the EU's approach, probing whether it represents a techno-solutionist misstep. In part III, we draw upon theoretical perspectives to provide a more balanced understanding of how we might regulate CSAM. We engage with Lessig's perspectives on code as a means of control and its implications for children's safety, alongside Murray's concept of Networked Communitarianism. This section also discusses risk regulation, critically examining how a focus on potential risks might inadvertently compromise child protection efforts. We conclude by synthesising the insights gleaned, underscoring the complexities and multifaceted nature of online child safety policy. We aim to contribute to the ongoing discourse by highlighting the need for a balanced, multi-dimensional approach that recognises the limitations of techno-solutionism and advocates for more socially informed and comprehensive strategies for protecting children in the digital age.

2. The Revival of Safety-by-Design and Techno-Solutionism in Online Child Safety Policy

2.1 The rise of safety-by-design and techno-solutionism

In 1998, Professor Lawrence Lessig introduced his "New Chicago School".[16] This was not a novel approach to studying law and regulation. Instead, it was a successor model to the traditional Chicago School, which focused on regulators other than the law and intended to understand the structures of regulation outside the law's direct effect.[17] Central to his New Chicago model were his modalities of regulation – the argument that four constraints on behaviour can be used directly or indirectly to regulate the actions or activities of individuals, governments, corporations or markets intentionally or unintentionally.[18] Among Lessig's four constraints was "architecture" or "the world as I find it, understanding that as I find it, much of this world has been made."[19] As Lessig notes, the existence of walls constrains his (and the state's) ability to surveil an

13. European Digital Rights, n.11 above, 16).
14. *European Digital Right, 1.11 above,* 32-36.
15. CRIN & defenddigitalme, Privacy and Protection: A Report on the Rights of Children in the Digital Age (October 12, 2023) https://home.crin.org/readlistenwatch/stories/privacy-and-protection; European Data Protection Supervisor, 'EDPS Seminar on the CSAM proposal: "The Point of No Return?"' (Seminar Report, 10 November 2023) https://www.edps.europa.eu/system/files/2023-11/23-11-10_report_from_edps_seminar_on_csam_en.pdf accessed 10 March 2025.
16. L. Lessig, 'The New Chicago School' 27 *Journal of Legal Studies* 661 (1998).
17. *Ibid*, 661.
18. As Lessig notes, "I mean the constraining effect of some action, or policy, whether intended by anyone or not. In this sense, the sun regulates the day, or a market has a regulating effect on the supply of oranges." *Ibid*.
19. *Ibid*, 663.

individual. At the same time, if it "takes 24 hours to drive to the closest abortion clinic [this] is a constraint on a woman's ability to have an abortion."[20]

Lessig acknowledged that this modality had been long recognised and cited Bentham's Panopticon, Goffman's Frame Analysis, and Michel Foucault's works as examples of how lawyers, in general, and the Chicago School, in particular, have accounted for the impact of geography or architecture on understanding law and regulation.[21] What then was different in Lessig's "New Chicago School"? Lessig's 'New Chicago School' explored how law regulates behaviour indirectly by targeting other modalities of regulation.[22]

Lessig's contribution was to think about the competing modalities differently. He argued that the "old" model viewed law as conflicting with markets, norms, or architecture, but he encouraged lawyers and policymakers to think of the modalities as instruments of regulation that regulators might deploy. Thus, the "new" in the "New Chicago School" referred to thinking more in terms of regulatory hybrids that employ two or more modalities, rather than viewing them in competition. Lessig built out his "New Chicago Model" in his 1999 paper *The Law of the Horse: What Cyberlaw Might Teach"*[23] and his book of the same year, *Code and Other Laws of Cyberspace*.[24] Both conveyed the same message: architecture, in the form of code, is compelling in the digital environment.

Lessig observed that when architecture moves from the physical to the digital, its regulatory malleability and effects are magnified. He noted that code may displace law, where digital software code sets standards independent of, or even in conflict with, legal values, or code may enhance law, where digital software code is deployed by statute in a hybrid regulatory form. He expands upon this in his book, discussing how lawmakers can cooperate with software engineers to "encode" legal regulatory rules.[25]

The concept was not novel. Langdon Winner, in his 1980 paper, *Do Artifacts Have Politics?* wrote extensively on "technical arrangements as forms of order",[26] while Joel Reidenberg wrote about the power of digital code as a regulatory mechanism in his 1998 paper, *Lex Informatica*.[27] However, Lessig's catchy aphorism that in the digital world, "Code is Law", allowed the message to reach a wider audience and the era of digital regulation through code-based design was ushered in.[28]

This model, what we might call regulatory enforcement by design or the deployment of law through code, was an early form of algorithmic regulation[29] and has become widespread. Regulators have used so-called "by-design" systems to enhance our online experience, empower users, and ensure our safety and security. The privacy-by-design structure outlined in data protection law is arguably the most familiar "by-design" system for lawyers. Privacy by design (PbD) started in the mid-1990s as Privacy Enhancing Technologies (PETs). The European Commission adopted PETs following the 1995 *Communication on Promoting Data Protection*.[30] By 2007, the Commission advised that "PETs should be developed and more widely used."[31] By the time the Commission began drafting the General Data Protection Regulation (GDPR), a move from PETs to PbD could be identified. Koops & Leenes observed that "privacy by design is now gradually replacing the concept of PETs."[32]

20. *Ibid*, 661 (1998), 663 citing *Casey v. Planned Parenthood* 505 U.S. 833.
21. *Ibid*, 665.
22. *Ibid*, 666.
23. L. Lessig, 'The Law of the Horse: What Cyberlaw Might Teach' 113 *Harvard Law Review* 501 (1999).
24. Basic Books, 1999.
25. *Ibid*, 53-4.
26. L. Winner, 'Do Artifacts Have Politics? 109 *Daedalus* 121 (1980).
27. J.R. Reidenberg, 'Lex Informatica: The Formulation of Information Policy Rules through Technology ' (1997-1998) 76 *Texas Law Review* 553.
28. n.24, above.
29. K. Yeung, 'Algorithmic regulation: A critical interrogation' (2018) 12 *Regulation & Governance* 505.
30. European Commission, *Communication on Promoting Data Protection*, Registratiekamer et al. (1995).
31. European Commission, *Communication on Promoting Data Protection by Privacy Enhancing Technologies (PETs)* COM (2007) 228 final.
32. B-J. Koops & R. Leenes, 'Privacy regulation cannot be hardcoded. A critical comment on the "privacy by design" provision in data-protection law,' [2014] 28 *International Review of Law, Computers & Technology* 159.

Data Protection by design was formalised in Article 25 of the GDPR. This requires data controllers to implement appropriate technical and organisational measures, such as pseudonymisation and data minimisation, to protect the rights of data subjects. Further, by Art.25(2), data controllers must "implement appropriate technical and organisational measures for ensuring that, by default, only personal data necessary for each specific purpose of the processing are processed". These measures reflect the philosophy that the most secure way to ensure the privacy of data subjects is to minimise the collection, storage, and processing of data by design.

Privacy by design is not without its critics. In 2013, at the outset of the EU's data protection reform programme, Koops & Leenes argued that the "design" element of PbD was problematic as "although simple and particular rules may be suitable for hard-coding in IT systems...many legal requirements (including data protection requirements) have been formulated in such a way as to allow flexible application in practice", as such they argued the encoding of these elements in design was likely to be problematic.[33] This reflects broader concerns about encoding flexible legal values and principles in code-based design and the risk of techno-determinism, which will be discussed below.

2.2 Understanding Safety-by-Design

Like PbD, Safety-by-design (SbD) is another "by design" solution. SbD is not a new principle nor unique to the digital environment, as it originated in workplace design and safety and can be traced back to the British Industrial Revolution. SbD is best defined as replacing reactive regulatory mechanisms such as legal liability or regulatory enforcement with proactive risk limitation models through anticipation and prevention of harm that might occur rather than trying to implement remedies after the harm has occurred. In the digital technology field, the Australian eSafety commissioner describes it as "focusing on the ways technology companies can minimise online threats by anticipating, detecting and eliminating online harms before they occur."[34]

Possibly the first SbD regulation was the Factories Act 1844, which provided that all dangerous machinery was to be securely fenced off, failure to do so was regarded as a criminal offence; and that no child or young person was to clean mill machinery while it was in motion.[35] Historically, SbD tended to focus on marginal risks where the benefit of regulation outweighed implementation costs. Examples include the successful UK road safety regulatory programme, which included mandatory seat belt installation and use from 1965 to 1983.[36]

Wider adoption of SBD in road safety may be seen in the 1997 Euro NCAP scheme. Although a voluntary scheme, the New Car Assessment Programme established design safety standards for in-car passengers and pedestrians in case of a collision. The NCAP scheme was, in turn, the model for the European Whole Vehicle Type Approval scheme launched in 2007 that tasked member states with ensuring vehicles sold into the EU market were safe.[37] The vehicle safety framework was substantially reviewed and updated in 2019 with the passage of the Regulation on type-approval requirements for motor vehicles and their trailers.[38] This added a substantial number of additional safety-by-design requirements, including that vehicles be equipped with tyre pressure monitoring systems, intelligent speed assistance, facilitation for the installation of alcohol interlocks, driver drowsiness and attention warnings, advanced driver distraction warnings,

33. *Ibid*, 166-7.
34. Australian eSafety Commissioner, 'Safety by Design": https://www.esafety.gov.au/industry/safety-by-design accessed 10 March 2025.
35. UK Parliament, 'Later factory legislation': https://www.parliament.uk/about/living-heritage/transformingsociety/livinglearning/19thcentury/overview/laterfactoryleg/ accessed 10 March 2025.
36. Exact figures on lives saved by this simple safety-by-design feature vary. In 2008, the Royal Society for the Prevention of Accidents suggested that more than 50,000 lives had been saved in the UK by seatbelts in the 25 years since mandatory seatbelt laws were introduced; See D. Batty, 'Seatbelt law anniversary marked with grim statistics' *The Guardian* 31 January 2008: https://www.theguardian.com/uk/2008/jan/31/transport.world accessed 10 March 2025.
37. Framework Directive for the approval of motor vehicles and their trailer Dir.2007/46/EC (Repealed and replaced by Reg. (EU) 2018/858).
38. Regulation (EU) 2019/2144.

advanced emergency braking systems, emergency lane-keeping systems and frontal protection systems designed to protect pedestrians.

There is a long and convincing history of safety-by-design in the road safety arena. Although it is impossible to state with certainty how many lives have been saved by improved vehicle safety standards one 2015 survey estimated that 613,501 lives had been saved in the US alone by vehicle safety technologies between 1960 and 2012[39], while a UN report just on seatbelt use estimates that millions of lives have been saved globally.[40]

Of course, road safety is but one area where safety-by-design systems have been implemented. A highly successful safety-by-design programme in the construction industry was launched in the UK in 1995 with the implementation of the Construction (Design and Management) Regulations 1994.[41] These required a health and safety file be kept and be made available for inspection by health and safety inspectors, that a health and safety plan be developed and applied during all elements of construction and most importantly, that project designers in designing the project "includes among the design considerations adequate regard to the need to avoid foreseeable risks to the health and safety of any person at work carrying out construction work or cleaning work in or on the structure at any time."[42]

The original 1994 Regulations have since been replaced by a more stringent set of safety measures in the Construction (Design and Management) Regulations 2015. These replace and enhance the original provisions by requiring that "when preparing or modifying a design, the designer must take into account the general principles of prevention and any pre-construction information to eliminate, so far as is reasonably practicable, foreseeable risks to the health or safety of any person."[43] The key move in 2015 is to replace a duty to avoid risks with one to eliminate them where possible. These regulations are credited with considerable safety improvements in UK construction, with around ten fatalities per 100,000 workers in 1994[44] being reduced to 2.1 per 100,000 workers in 2022/23.[45]

Modern safety-by-design is based on risk regulation, which is founded on the development of the risk society. Previously, risks were often addressed in isolation, such as workplace, food, or driving risks, and were dealt with directly. Ulrich Beck is regarded as the godfather of the risk society. According to Beck, risk has a different significance in contemporary everyday life. He argues that human activity and technology in "advanced modernity" produce side-effect risks that need specialised expertise to assess and recognise. He claims that these are collective, global, and irreversible in their impact, and thus have a greater effect than historical risks.[46]

Hood, Rothstein, and Baldwin explain that risk should not be seen in isolation: "As well as a 'risk society,' we are also said to live in a 'regulatory state'".[47] This leads to a balancing of principles between risk tolerances and the anticipation and intrusiveness of regulation. The balancing between these principles may be inconsistent; as Hood, Rothstein and Baldwin note,

> Some domains, notably road safety, are dominated by a 'cost-benefit-analysis culture' in which the costs of additional safety measures are weighed against probable benefits using explicit value-of-life calculations...[however] smoking tends to be less heavily regulated than vehicle

39. National Highway Traffic Safety Administration, *Lives Saved by Vehicle Safety Technologies and Associated Federal Motor Vehicle Safety Standards, 1960 to 2012* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812069.pdf.
40. United Nations, 'UN rights chief warns EU online child safety plan risks "systemic" surveillance' (UN News, 14 June 2023) https://news.un.org/en/story/2023/06/1137412 accessed 10 March 2025.
41. SI 1994/3140.
42. Reg.13(2).
43. SI 2015/51, Reg.9(2).
44. House of Commons Committee of Public Accounts, *Health, and Safety Executive: Improving Health and Safety in the Construction Industry* HC 627 (December 14, 2004).
45. Health and Safety Executive, *Work-related fatal injuries in Great Britain 2023* https://www.hse.gov.uk/statistics/assets/docs/fatalinjuries.pdf accessed 10 March 2025.
46. U. Beck, *Risk Society* (trans. M. Ritter), Sage 1992.
47. C. Hood, H. Rothstein & R. Baldwin, *The Government of Risk: Understanding Risk Regulation Regimes*, OUP, 2001, 4.

> emissions although it is normally assumed to be a much bigger killer, and domestic accident risks are much more lightly regulated than occupational risks, even though the former claim ten times more lives a year than the latter in the UK.[48]

Disproportionate risk regulatory responses can often be traced to specific examples of harm or moral panics, such as the BSE crisis in the UK in the early 1990s[49] or the Dangerous Dogs Act 1991.[50]

How does one design an effective and balanced SbD system? As Hood, Rothstein, and Baldwin outline, there are various techniques for modelling risk, from the cost-benefit analysis used in road safety to "quantified risk assessment," in which risks are expressed in elaborate numerical terms. Forms of SbD regulation or management vary from a culture of inter-agency bargaining (who pays, how much and when) to wholly qualitative 'seat of the pants' approaches to standard-setting in cases such as the regulation of guns or activity holiday centres.[51]

The development of a risk-regulatory regime is equally heterogeneous, with approaches varying from "specialist 'risk bureaucracies', in the sense of state agencies staffed by specialists in risk management, to more generalist agencies, self-regulatory arrangements, or the law courts, which may rely on lay reporting about hazards rather than specialised monitoring."[52] To design an SbD system, the risk first has to be quantified by a model, and a regulatory regime must be put in place (to approve or oversee the design). The risk identified in the CSAM proposal is set out in the explanatory memorandum: "At least one in five children falls victim to sexual violence during childhood. A 2021 global study found that more than one in three respondents had been asked to do something sexually explicit online during their childhood, and over half had experienced a form of child sexual abuse online". Children with disabilities face a higher risk: up to 68% of girls and 30% of boys with intellectual or developmental disabilities will be sexually abused before their 18th birthday.[53] This quantifiable risk falls into the higher categories of Hood, Rothstein, and Baldwin's modelling. Has there been a complete cost-benefit analysis, or is this more in the style of Hood, Rothstein, and Baldwin's 'quantified risk assessment' model? It seems to fall in part between both. While the costs to the EU in setting up and supporting the new EU Centre are closely modelled, the compliance costs for Coordinating Authorities and, most importantly, for Information Society Service Providers are less well modelled.

Design solutions are assumed to be more efficient and to engender greater trust. Indeed, the CSAM proposal adopts such language, noting that Information Society Service Providers shall benefit from legal certainty in harmonised rules and higher levels of trust in their services where safer-by-design methods are adopted. In addition, users will benefit from a more structured approach to preventing and detecting abuse, as well as higher levels of trust in services that utilise safer-by-design methods. The benefit of "by design' systems in a regulatory sense is compliance or enforcement through self-execution, which is assumed to be more efficient in terms of cost and outcomes. Lessig recognised this benefit, noting that "what distinguishes the architectural constraints from other constraints is how they are experienced. They are experienced as conditions on one's access to areas".[54]

Murray and Scott's paper Controlling the New Media: Hybrid Responses to New Forms of Power explained the impact of this.[55] As the authors explain, ordinary systems of cybernetic regulation rely on three elements: a standard-setter, an information gatherer, and a method of behaviour modification. For a hierarchical system (in Lessig's language, Law), the elements are:

48. *Ibid*, 7.
49. J. Eldridge & J. Reilly, 'Risk and Relativity: BSE and the British Media' in N. Pidgeon, R.E. Kasperson & P. Slovic (eds) *The Social Amplification of Risk*, CUP, 2003.
50. S. Hallsworth, 'Then they came for the dogs!' (2011) 55 *Crime, Law, and Social Change* 391.
51. Hood, Rothstein & Baldwin, n.47 above, 7.
52. *Ibid*.
53. European Commission, n.6 above.
54. Lessig, n.23 above, 509.
55. A. Murray & C. Scott, 'Controlling the New Media: Hybrid Responses to New Forms of Power' (2002) 65 *Modern Law Review* 491.

(1) a Formalised Rule
(2) Monitoring (by, e.g. Law Enforcement), and
(3) Enforcement (by courts/prisons, etc).

There is a clear break between monitoring and enforcement, and a choice can be made not to enforce the rule.[56]

However, as Murray & Scott point out, "architecture-based regimes may be self-executing [about] monitoring and behaviour modification"[57] This brings two effects. The first is the *ex-ante* nature of enforcement. Whereas Lessig's other modalities (Law, Norms and Markets) rely on *ex-post* enforcement (police and prosecutors decide whether to bring charges; courts and juries decide whether to convict; social observers choose whether to admonish or ostracise; markets decide whether to accept or recognise your transaction/offer) with by-design architecture enforcement is "baked in". The locked door will not permit passage even if you have a legal right to gain access; Robert Moses bridges prevented buses from reaching the beaches for whatever reason, and speed bumps are equally effective against teenagers exceeding the speed limit and the police vehicle pursuing them. By-design solutions are inscrutable, objective, and self-executing.[58] These are generally viewed as positive values, but as we shall see, they also have weaknesses.

2.3 Techno-Solutionism and Determinism Explained

Often attributed to Evgeny Morozov, the term 'technosolutionism' represents an intellectual paradigm predicated on the presumption that technological advancements possess the inherent capacity to resolve multifarious social, political, and environmental quandaries, frequently circumventing a comprehensive engagement with the intricate and foundational aspects of these conundrums.[59] Technosolutionism posits that innovations in computing, artificial intelligence, and data analytics, among other fields, are equipped to offer prompt and effective solutions to challenges that, in reality, necessitate a more nuanced, multifaceted, and interdisciplinary approach to addressing them.

Critiques of technosolutionism centre on its propensity to engender a reductive interpretation of complexities, thereby neglecting the essential human, cultural, and societal dimensions integral to these issues. Furthermore, an overreliance on technological solutions is identified as a potential catalyst for unforeseen repercussions, which may, in certain instances, exacerbate the very dilemmas they aim to alleviate.[60] For example, O'Neil argues that the discriminatory and harmful outcomes of algorithms used in finance and criminal justice are due to reliance on technology.[61]

Among the proponents of a technologically optimistic perspective is Kurzweil, renowned for his foresight in technological evolution and transhumanist ideals. Kurzweil posits that the relentless pace of technological advancement will be instrumental in overcoming formidable obstacles such as disease and poverty.[62] Diamandis espouses a sanguine outlook on emerging technologies, such as robotics and AI, envisioning their pivotal role in resolving global crises, including hunger and energy scarcity.[63] Meanwhile, Kai-Fu Lee advocates for the beneficial applications of AI across various sectors, including healthcare and education, while concurrently emphasising the necessity of judicious development and governance.[64] These figures

56. As happened in the case of the "Colston Four" in January 2022. See I. Hare, 'Public order, public protest, and public monuments' (2023) 54 *Victoria University of Wellington Law Review* 183.
57. Murray & Scott, n.55 above, 501.
58. R. Brownsword, 'Code, control, and choice: why East is East and West is West' (2005) 25 *Legal Studies* 1.
59. E. Morozov, 'To save everything, click here: the folly of technological solutionism' 4 *Journal of Information Policy* 173-175 (2014); See also L. Mumford, *The Myth of the Machine* (Harcourt, Brace & World, 1967).
60. For examples on how algorithmic bias and data-driven solutions can perpetuate racial inequalities, see R. Benjamin, 'Race after technology' in W. Longhofer & D. Winchester (eds) *Social Theory Re-Wired* (Routledge, 2023); For a critique of technosolutionism during the Covid-19 pandemic, see S. Milan, 'Techno-solutionism and the standard human in the making of the COVID-19 pandemic' 7 *Big Data & Society* 2053951720966781 (2020).
61. C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy* (Crown, 2017).
62. R. Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (Duckworth, 2006).
63. P.H. Diamandis & S. Kotler, *Abundance: The future is better than you think* (Simon and Schuster, 2012).
64. K. F. Lee, *AI superpowers: China, Silicon Valley, and the new world order* (Houghton Mifflin, 2018).

seldom epitomise an unadulterated, unquestioning adherence to technosolutionism. They frequently emphasise the importance of responsible development, the integration of ethical considerations, and the need to address potential adverse effects that accompany the acclaimed advantages. The nature of technosolutionism often lies in the balance between recognising technology's potential and acknowledging its limitations.

Technodeterminism posits technology as the primary driver of societal change. Renowned for his aphorism, "The medium is the message," McLuhan argued that communication technologies fundamentally alter how we perceive and interact with the world.[65] This concept asserts that technological advancements and innovations shape society's structure, values, and behaviour, often deterministically, meaning that these changes are seen as inevitable outcomes of technological progress. Technodeterminism is frequently invoked as a critical framework rather than a definitive school of thought in academic discourse. It embodies a generalised notion that technology fundamentally steers, or in some instances dictates, the trajectory of human society whilst often marginalising the significance of social, political, and cultural determinants.

In a technodeterminist view, social, cultural, and historical contexts are often considered secondary to the influence of technology. It suggests that technology development follows a fixed path with predictable societal outcomes. For example, the invention of the printing press is often cited as a technodeterminist catalyst for spreading literacy and democratising knowledge. At the same time, the Internet is seen as a transformative force for global communication and information exchange.

Critics of technodeterminism argue that this view oversimplifies the relationship between technology and society. They suggest that it neglects the role of human agency, social structures, and cultural contexts in shaping and being shaped by technological development.[66] The interaction between society and technology is typically more complex and bidirectional; societal needs and values can influence technological innovation just as much as new technologies can impact society. Haraway, for example, scrutinises the confluence of technology, science, and culture in her work, drawing attention to issues of power disparities and the potential of technology to perpetuate extant inequalities.[67] Feenberg delves into the social and political ramifications of technological evolution, underscoring its influence on labour, knowledge, and societal structures.[68] Meanwhile, Turkle examines the repercussions of technology on human behaviour and interpersonal relationships, exploring the risks associated with overdependence on technology and its impact on identity and social connections.[69]

Technodeterminism is often employed as a heuristic device to challenge reductive perspectives on the role of technology in society, advocating for a more nuanced and multifaceted analytical approach to addressing society's problems. A comprehensive understanding of technodeterminism requires an appreciation of its diverse interpretations and an acknowledgement of the complex interplay between technology and socio-political and cultural factors. While acknowledging the valuable insights from exploring technodeterminism's critiques of societal oversimplification regarding technology's influence, a closer examination reveals its limitations.

2.4 The rationale & role of safety-by-design & the resurgence of techno-solutionism in addressing child safety online

Technodeterminism's focus on the transformative power of technology looms large when navigating the treacherous waters of online child safety. Examining the problem solely through this lens risks neglecting crucial societal dimensions and perpetuating simplistic solutions to a multifaceted challenge. Within this context, the rising tide of the SbD movement and the potential resurgence of techno-solutionism require scrutiny.

65. M. McLuhan, *Understanding media: The extensions of man* (MIT press, 1994).
66. B. Latour, *Science in Action: How to Follow Scientists and Engineers Through Society* (Harvard University Press, 1987).
67. D. Haraway, 'A Cyborg Manifesto' (1991) *Nature* 1-6.
68. A. Feenberg, *Questioning Technology* (Routledge, 2012).
69. S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other* (Basic Books, 2011).

The allure of SbD, embedded in technologies making them inherently resistant to misuse, is undeniable. Proponents envision platforms proactively filtering harmful content, algorithms detecting abuse, and design choices prioritising safe user experiences.[70] However, a critical evaluation reveals potential limitations. Oversimplifying the interplay between technology and social factors can overlook the root causes, such as poverty, inequality, and a lack of digital literacy.[71] Moreover, concerns surface regarding unintended consequences, potential infringement on privacy, and the chilling effect on expression.[72]

The question then arises: does SbD inadvertently fuel a resurgence of techno-solutionism? When relying on technological fixes, might we need to pay more attention to nuanced social and cultural considerations? After all, technodeterminism's pitfalls lie not in acknowledging technology's influence but in overemphasising its power to solve complex societal problems unilaterally.[73] A balanced approach is crucial to effectively safeguarding children online. This necessitates understanding technology's limitations while harnessing its potential in conjunction with robust social interventions, effective legal frameworks, and empowering educational initiatives. By navigating beyond the simplistic allure of techno-solutionism, we can chart a course towards a safer online environment for children that recognises the complex interplay of technology, society, and individual agency.[74]

While recognising the critical need for comprehensive child safety measures, part II of our analysis delves deeper into the potential pitfalls of over-relying on technological solutions.

3. Technosolutionism in Practice

3.1 Technosolutionism and the CSAM Proposal

Aligning with the UN Convention on the Rights of the Child (UNCRC) and recommendations of the CRC Committee, the European Commission's strategy on combating child sexual abuse (EU CSA strategy) acknowledges that progress requires a comprehensive approach combining prevention, reporting, investigation, protection, treatment, and follow-up.[75] Legislative measures are part of a broader, multi-stakeholder strategy addressing both online and offline abuse. In May 2022, the Commission introduced the Better Internet for Kids Plus (BIK+) strategy to promote children's protection, empowerment, and respect online, underscoring the need for further regulation of digital technologies, including social media, to address the harms of online child sexual abuse (OCSA).[76]

Alongside BIK+, the Commission unveiled its draft Regulation on 11 May 2022—the CSAM Proposal—which aims to establish a coherent EU framework for preventing and combating online child sexual abuse.[77] It advances the rights in Articles 19 and 34 of the CRC and Article 24 of the EU Charter, affirming children's right to necessary protection and care. The Proposal sets clear obligations for service providers with a significant risk of exposure to CSAM or solicitation, complementing the Digital Services Act (DSA)[78] by elaborating

70. A. Feenberg, *Technological Society: The Promise and the Peril* (Routledge, 2017).
71. S. Livingstone, *Parenting for a Digital Future: How Homes and Schools Can Work Together for Children's Wellbeing* (OUP, 2020).
72. D. Susser, *The Limits of Privacy* (Routledge, 2011). J. Townend, 'Freedom of Expression and the Chilling Effect' in H. Tumber & S. Waisbord, *The Routledge Companion to Media and Human Rights* (Routledge, 2017).
73. L. Winner, *Autonomous technology: Technics-out-of-control as a theme in political thought* (MIT Press, 1978).
74. M. S. Gal, 'Algorithmic Challenges to Autonomous Choice' 25 *Michigan Telecommunications & Technology Law Review* 59 (2018); A. D. Murray, Almost Human: Law and Human Agency in the Time of Artificial Intelligence (TMC Asser Press, 2021) https://www.asser.nl/upload/documents/20240926T102230-AL6-Murray-WEB%20new.pdf accessed 10 March 2025.
75. European Commission, 'Protecting Children from Sexual Abuse' (Home Affairs, European Commission) https://home-affairs.ec.europa.eu/policies/internal-security/protecting-children-sexual-abuse_en accessed 10 March 2025, 7 accessed 10 March 2025.
76. European Commission, 'EU Better Internet for Kids (BIK+) Strategy' (Communication) COM(2022) 212 final, 11 May 2022 https://digital-strategy.ec.europa.eu/en/policies/strategy-better-internet-kids accessed 10 March 2025.
77. European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Rules to Prevent and Combat Child Sexual Abuse' COM(2022) 209 final, 11 May 2022 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0209 accessed 10 March 2025.
78. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).

specific rules to combat OCSA. The CSAM Proposal clarifies providers' responsibilities in assessing and mitigating risks, detecting, reporting, removing CSAM, and preventing the solicitation of minors. While some consider it essential for holding platforms accountable, critics argue that it risks introducing mass surveillance across the EU.[79]

Conceptually, the CSAM Proposal operates as a workflow. It provides a structured framework to help platforms identify and address abusive content, supported by regulatory articles specifying the roles and responsibilities of all stakeholders. Providers must assess risks in the initial three months[80] and develop mitigation strategies.[81] The Coordinating Authority may seek a detection order from the competent judicial authority if significant risks persist. Before finalising the order, a four-week commentary phase allows providers and the EU Centre to discuss the draft and provide feedback. Providers then submit an implementation plan that details the technologies they will use, which must comply with applicable regulatory requirements. Once implemented, providers report their actions and results to a new 'EU Centre'.[82] Suppose substantial risks remain unaddressed after the initial mitigation period. In that case, the Coordinating Authority can request a detection order from the competent judicial authority of the Member State, a pivotal element of the regulatory framework.[83]

Before the detection order is finalised, a four-week commentary phase is initiated.[84] This phase provides an open platform for providers and the central EU body, the EU Centre, to engage in constructive deliberations, scrutinise the draft request, and offer informed perspectives.[85] Providers create a definitive implementation plan once the comments have been consolidated and considered.[86] This phase, as described in Recital 26, emphasises the providers' role in determining the technological strategies employed to comply with the detection orders. Interestingly, while providers are given autonomy to choose their technologies, the chosen technology must align with regulatory requirements. Providers traverse this systematic workflow and must report their findings and subsequent actions to the EU Centre.

Once initiated, the detection phase requires providers to deploy specific technologies and methodologies calibrated to identify potential risks using specified indicators provided by the EU Center, which are explicitly tailored to detect CSAM or the solicitation of children.[87] Technological considerations emphasise the characteristics of the detection technologies and the onus on providers in their application.[88] Furthermore, the Commission is mandated to develop and refine guidelines in collaboration with Coordinating Authorities and the EU Centre.[89] A standout feature of this regulatory design is the EU Center's commitment to providing service providers with state-of-the-art detection technologies at no cost.[90] Thus, the detection orders process represents a blend of technosolutionist and regulatory measures crafted to combat the risks associated with CSAM. Proponents of the Proposal refer to its safeguards, including the proportionality of processing, the use of state-of-the-art and reliable detection technologies, and the employment of relevant key indicators for grooming detection.[91]

79. K.R. Ludvigsen, S. Nagaraja, & A. Daly, 'YASM (Yet Another Surveillance Mechanism)' https://arxiv.org/abs/2205.14601; See also, H. Abelson, et al., 'Bugs in our pockets: The risks of client-side scanning' (2024) 10 *Journal of Cybersecurity* tyad020 https://academic.oup.com/cybersecurity/article/10/1/tyad020/7590463, both accessed 10 March 2025.
80. Article 3(4), CSAM Proposal.
81. Article 4.
82. Article 9.
83. Article 7(1).
84. Article 7(3)(d).
85. Article 40.
86. Article 8(3).
87. Article 8(1), Article 10(1).
88. Articles 10(3) and 10(4).
89. Article 11.
90. Article 50(1).
91. European Parliamentary Research Service, 'Combating Child Sexual Abuse Online' (Briefing, 2022) https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/738224/EPRS_BRI(2022)738224_EN.pdf accessed 10 March 2025.

Proponents have also expressed significant displeasure with the proposal's critics, suggesting that technology companies already do much of what would become a legal obligation if the proposal were enacted.[92] Microsoft already uses classifiers to detect adult content on services like OneDrive, Teams, and Xbox Live, which can sometimes identify new CSAM material.[93] Known images of CSAM are identified through databases such as the National Center for Missing and Exploited Children and the UK's Internet Watch Foundation.[94] Technologies like Google's Content Safety API and Thorn's classifier, which they claim have a high precision rate, are used to detect new material by identifying markers of content depicting abuse.[95] There is recognition of the challenges in detecting new or 'first-generation' CSAM material, especially considering the high standards required due to privacy implications. Snap, for example, is exploring technologies that could detect new or unhashed CSEA material. Meta has collaborated with child exploitation experts, including NCMEC, to develop a taxonomy that may help identify malicious intent in sharing CSAM material.[96]

Technology to police for CSAM typically relies upon classifiers—algorithms that categorise data into various classes or related categories. These classifiers, particularly those that rely on textual data, often face formidable challenges in accurately interpreting context.[97] Differentiating between illicit and innocuous interactions necessitates a nuanced understanding that current algorithms struggle to achieve.[98] This inherent limitation often results in a high rate of false positives, generating unnecessary alerts and potentially infringing upon individual liberties.[99] The reliability of algorithmic age detection, particularly in distinguishing between juveniles aged 16 to 19, remains questionable. Particularly in this age bracket, there is a significant propensity for false-positive and false-negative outcomes.[100]

Replicating the complexities and nuances of human communication remains a pivotal obstacle for classifiers. The inherent ambiguities and emotive expressions in natural language pose significant challenges for these algorithms, potentially leading to misinterpretation and exacerbating false outcomes.[101] Moreover, while currently considered a low-probability threat owing to the requisite technical expertise, the possibility of malicious actors repurposing classifiers for unintended applications is a burgeoning concern.[102] Such

92. Thorn, 'Open Letter: Thorn and 50+ Organizations Welcome the EU's Proposal to Prevent and Combat Child Sexual Abuse' https://www.thorn.org/blog/open-letter-thorn-and-50-organizations-welcome-the-eus-proposal-to-prevent-and-combat-child-sexual-abuse/ accessed 10 March 2025.
93. For an overview of all ongoing efforts by BigTech, see Parliament of Australia, 'Law enforcement capabilities in relation to child exploitation', (November 2023), https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Law_Enforcement/ChildExploitation47th/Report/Chapter_5_-_Technology_and_child_exploitation at 5.9 accessed 10 March 2025.
94. T. Bernard, 'The Present and Future of Detecting Child Sexual Abuse Material on Social Media', *UnitaryAI* https://www.unitary.ai/articles/the-present-and-future-of-detecting-child-sexual-abuse-material-on-social-media, (October 16, 2023) accessed 10 March 2025.
95. Google, 'Protecting Children' https://protectingchildren.google/intl/en_uk/#tools-to-fight-csam accessed 10 March 2025.
96. K. Wiggers, 'Meta, Discord and Others Unveil Effort to Combat Online Child Sexual Exploitation and Abuse' *TechCrunch*, 7 November 2023 https://techcrunch.com/2023/11/07/meta-discord-and-others-unveil-effort-to-combat-online-child-sexual-exploitation-and-abuse/ accessed 10 March 2025.
97. European Parliamentary Research Service, *Proposal for a Regulation Laying Down the Rules to Prevent and Combat Child Sexual Abuse: Complementary Impact Assessment* (Study, 2023) 5 https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf accessed 10 March 2025.
98. S.K. Witting, 'Transnational by default: online child sexual abuse respects no borders' (2021) 29 *The International Journal of Children's Rights*, 731, 742; See Leiser and Witting n.7, above.
99. European Data Protection Board and European Data Protection Supervisor, Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Rules to Prevent and Combat Child Sexual Abuse (28 July 2022) https://www.edpb.europa.eu/system/files/2022-07/edpb_edps_jointopinion_202204_csam_en_0.pdf accessed 10 March 2025; See e.g., K. Hill, 'A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal, *New York Times* 21 August 2022, https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html accessed 10 March 2025.
100. Yoti, *Age Estimation White Paper* (May 2022) https://www.yoti.com/wp-content/uploads/Yoti-Age-Estimation-White-Paper-May-2022.pdf accessed 10 March 2025.
101. Abelson et al, n.79 above.
102. Abelson et al, n.79 above. See also A. Crocker, M.R. Leiser & S.K. Witting, 'Outcome Report of 2nd Expert Workshop on EU proposed Regulation on preventing and combatting online child sexual abuse' in *Outcome Report of 2nd Expert Workshop on EU proposed Regulation on preventing and combatting online child sexual abuse* VU-Amsterdam (April 2023) at 10-13.

alterations could severely compromise the reliability and effectiveness of these systems, underscoring the need for robust security measures and ethical considerations.[103]

Another critical concern is the bias inherent within training datasets and algorithms. Such biases can significantly skew classifier outcomes, necessitating vigilant human oversight and continuous efforts to mitigate bias.[104] This ensures that the outputs of these systems are fair, objective, and representative of the populations they serve. Most text-based classifiers are predominantly trained on English language models, limiting their efficacy in multilingual contexts.[105] This limitation presents a significant challenge for broader adoption and necessitates the development of culturally and linguistically diverse training datasets to ensure inclusivity and equitable outcomes.

Turning to the specific context of CSAM detection, several salient issues emerge. First, the general description of classifiers as machine learning-based algorithms for categorising data is fundamental. However, it is essential to note that classifiers based on simplistic features, such as geographic location, age, or gender, are inadequate for specifying user groups subject to detection orders. According to the European Parliamentary Research Service's Impact Assessment on the Proposal, their broad nature and the ease with which they can be circumvented might render them ineffective.[106]

Second, the advent of GPT-based Large Language Models (LLMs) has significantly expanded the capabilities and applications of machine learning, particularly in the realm of natural language understanding and generation. This technological leap forward has profound implications for how classifiers are perceived and utilised across various domains. The traditional understanding of classifiers as machine learning-based algorithms designed to categorise data remains essential. Classifiers have been widely employed in multiple applications, including spam detection in emails and the identification of potential fraud in financial transactions. These models are trained on historical data to recognise patterns and make predictions about new, unseen data.

The rise of GPT-based LLMs introduces a nuanced perspective on the capabilities and limitations of traditional classifiers. While it is true that classifiers based on simplistic features, such as geographic location, age, or gender, might have been sufficient for specific tasks, the complexity and subtlety of human language, as well as the intricacies of human behaviour, demand more sophisticated approaches. With their deep understanding of complex language and ability to generate contextually relevant responses, GPT-based LLMs showcase that classifiers can move beyond simplistic categorisations. These models can comprehend and analyse text to mimic human understanding, allowing for a more nuanced and dynamic data classification. This capability is particularly crucial when dealing with complex and sensitive issues such as content moderation and sentiment analysis, where understanding context, subtlety, and intent is paramount.

Moreover, the use of GPT-based LLMs highlights the importance and necessity of developing fair, transparent, and unbiased models. As classifiers become more sophisticated, the potential for misuse or unintended consequences increases. This evolution calls for re-evaluating how user groups are defined and categorised, especially in contexts subject to detection orders. Simplistic categorisations based on basic demographic features may not only be inapt but could lead to discriminatory practices and overlook the nuanced ways individuals interact with content and systems. While the general description of classifiers remains a cornerstone in machine learning, the emergence of GPT-based large language models (LLMs) highlights the need to shift towards more sophisticated and context-aware approaches. This evolution does not alter the fundamental nature of classifiers but expands their potential and underscores the importance of advancing these technologies responsibly and inclusively.

---

103. Abelson et al, n.79 above.

104. European Agency for Fundamental Rights, *Bias in Algorithms – Artificial Intelligence and Discrimination* (2022), 11 https://fra.europa.eu/en/publication/2022/bias-algorithm; I. Chen, F.D. Johansson & D. Sontag, 'Why is my classifier discriminatory?' *Advances in Neural Information Processing Systems* 31 https://papers.nips.cc/paper_files/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf both accessed 10 March 2025.

105. European Parliamentary Research Service, above n.97, 17 & 82.

106. Ibid, 56.

Despite the development of LLMs and the improved classification of complex descriptions, certain aspects remain challenging for algorithmic tools to address. The risks associated with CSA detection encompass concerns regarding feasibility, privacy, security, and transparency, all of which remain regardless of the technology deployed.[107] A point of contention is the narrow interpretation of the privacy indicator, which primarily focuses on the risk of abuse by the service provider.[108]

Third, experts foresee the risk of function creep, in which the CSA model could eventually be extended to domains such as counterterrorism or even political dissent.[109] Additionally, the transparency and fairness of algorithms used for content moderation are under scrutiny, particularly regarding the risk of discrimination against certain groups by these algorithms.[110]

Finally, it should be acknowledged that detection quality and its impact on Law Enforcement Agencies (LEAs) are pressing concerns. Experts anticipate that technology-led detection will compromise detection quality due to the need to identify new content and address grooming, resulting in higher error rates and negatively impacting LEAs investigative capabilities. Moreover, the potential for abuse in discrimination and predictive policing, along with the questionable feasibility of filtering reports due to the sheer volume of communications, presents significant challenges, including the risk of numerous false positives.[111]

3.2 The UK's Online Safety Act: A Technosolutionist Pipe Dream?

The UK's Online Safety Act, while aiming to address pressing concerns about online harms, risks also falling prey to the allure of technosolutionism and technodeterminism. The Act focuses on platforms that allow user-to-user speech and search engines. In this regard, it follows a similar regulatory pathway to the DSA. CSAM content is addressed both indirectly and directly, with a reliance on technological solutions to address social and cultural issues.

Indirect regulation may be found in the Act's general requirements for online safety and the duties of care owed by service providers to users. So-called user-to-user services[112] have a duty of care to ensure that users do not encounter illegal content while using the service.[113] This requires the service operator to conduct a risk assessment of unlawful content and fulfil their duty to ensure the safety of users from illegal content.[114]

In fulfilling the safety duty, heavy reliance is placed upon technology, with platform operators being required to prevent access to illegal content and/or mitigate the risk of the platform being used to commit illegal acts or to expose users to illegal acts by the design and operation of the platform, including "regulatory compliance and risk management arrangements" and "design of functionalities, algorithms and other features."[115] This approach emphasises immediate responsiveness, algorithmic moderation, and policing tools, which have caused several commentators to express concerns.[116] Beyond this general duty to online safety, there is also a specific duty of care for children's safety, which applies to any platform or search

107. EPRS, n.97 above, 10–11.
108. Ibid, Annex 8, 279.
109. Ibid,10-11.
110. R. Binns, M. Veale, M. Van Kleek & N. Shadbolt, 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation' in *Social Informatics: 9th International Conference, SocInfo 2017*, https://arxiv.org/pdf/1707.01477 accessed 10 March 2025.
111. F. Dakalbab et al. 'Artificial intelligence & crime prediction: A systematic literature review' (2022) 6 *Social Sciences & Humanities Open* 100342; A. Limantė, 'Bias in Facial Recognition Technologies Used by Law Enforcement: Understanding the Causes and Searching for a Way Out' (2023) *Nordic Journal of Human Rights* 1; F. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making* (2018, Council of Europe); O. Lynskey, 'Criminal justice profiling and EU data protection law: precarious protection from predictive policing' (2019) 15 *International Journal of Law in Context*, 162.
112. OSA, s.3(1).
113. OSA, s.7(1) for platforms and s.24(1) for search.
114. OSA, s.10 for platforms and s.27 for search.
115. OSA, s.10(4)(a) and (b) for platforms and s.27(4)(a) and (b) for search.
116. S. Dawood, "Will the Online Safety Act protect us or infringe our freedoms?" *New Statesman* 17 November 2023: https://www.newstatesman.com/spotlight/tech-regulation/online-safety/2023/11/online-safety-act-law-bill-internet-regulation-free-speech-children-safe accessed 10 March 2025.

service that is likely to be accessed by children. This involves further risk assessments and safety duties aimed at keeping children safe online.[117]

Together, these indirect safety duties are designed to ensure that illegal content, terrorist content, CSAM, and so-called priority offences[118] Platforms and search engines should not carry them; where they do, they should be identified and removed immediately. These are all complex sociological and contextual offences. What constitutes fraud depends on the context and the existence of a form of deception, which algorithms are weak at detecting.

Similarly, the context around policing some forms of CSAM is more complex than can be captured in the rather blunt safety duties found here. A considerable proportion of CSAM today is created by children for consumption by other children in the form of consensual (or perhaps coerced) sexting.[119] This is, without doubt, a highly risky activity, and there is no doubt that a proportion of those who send sexts later regret doing so, especially if they are then shared without their consent. However, it is also part of the process of growing up and exploring their sexual and social development. It cannot be sociologically beneficial for children in the upper age bracket of 15-18 to be criminalised via this activity or to have it algorithmically regulated by platforms as part of an illegal content or child safety duty.

This, however, is a likely outcome of the OSA. Popular messaging services such as WhatsApp, Snap, Signal and Telegram are all caught by the definition of a user-to-user service and so have to meet both the illegal content and children's safety duties, which includes a risk assessment and enforcement process, which we remember includes "design of functionalities, algorithms and other features" to prevent an "individual from encountering priority illegal content" including an "indecent photograph of a child". This means that these services will be required to use algorithmic tools to prevent the sharing of sexual images of anyone under eighteen.

This is a laudable aim when discussing adults accessing such content, however, it is more questionable when it regulates sharing between minors, especially when they are of the age of sexual majority, which in the UK is sixteen. Now, you might ask how anyone would ever know if children were sexting each other using messaging services, especially given that all the major services offer end-to-end encryption. The answer here is the controversial nature of s.121 of the Act. Under s.121, the regulator Ofcom may issue a notice that requires a provider to use accredited technology to deal with content found on or in its services, including identifying and taking down CSEA content. There is concern that this may create a backdoor into end-to-end encryption. Little is known about how s.121 will work in practice, as currently no accredited technology exists.[120] However, s.121 creates the potential for platform providers to be required to police private encrypted communications to identify CSAM content and then block and/or remove it. The implications for children's sexual development in a complicated time for adolescents are clear.

CSAM is also directly regulated in Part 4 of the Act. Like the workflow of the CSAM Proposal, platforms and search providers must notify the National Crime Agency of any CSAM content they find on their services.[121] Precise reporting requirements will be set out by the Secretary of State under s.67 and will include timeframes for reporting content, the preservation of evidence and data, and the format of reports. Anyone found to make a false report commits an offence under s.69. This will likely create a substantial duty on regulated services. While a pre-existing industry agreement exists to report CSAM content to the Internet Watch Foundation and the NCA's Child Exploitation and Online Protection Command, this can be done

117. OSA, ss.11-12 for platforms and ss.28-29 for search.
118. Priority Offences are found in Schedule 7 and include harassment, threats, and violent speech, offering to supply drugs or firearms, human trafficking, fraud, and other offences.
119. Research by Bianchi et al in 2017 found that 19.3% of individuals under 18 had sent a "sext," 34.8% had received one, and 14.5% had forwarded a sext" See D. Bianchi, et al, 'Sexting as the mirror on the wall: Body-esteem attribution, media models, and objectified-body consciousness' (2017) 61 *Journal of Adolescence* 164.
120. Currently Ofcom are consulting on that forms accredited technology may take. See https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/consultation-technology-notices/ accessed 10 March 2025.
121. OSA, s.66(1) for platforms, s.66(3) for search.

selectively. Currently, the likelihood of consensual sexting content being reported is low. However, by s.66 of the OSA, it becomes a duty to operate the service using systems and processes which secure (as far as possible) that the provider reports all detected and unreported CSEA content on the service to the NCA. This positive duty, if allied to algorithmic detection tools, is likely to flag large volumes of consensual image sharing, leading to not only the likelihood that the speech of adolescents will be chilled but also that they are opening themselves up to a police investigation by consensual sexting.

Children's sexual development and maturity are a complex psychological and sociological stage of development. In recent years, whether adults like it or not, sexting has become part of that development. The issue of adolescent sexting is further complicated by the fact that at sixteen, adolescents become legally sexually mature in England & Wales and may engage in all sexual acts, including penetrative sex, but under s.7(6) of the Protection of Children Act 1978 the definition of a child for an indecent image of a child is a person under the age of 18. This means that two sixteen-year-olds may engage in sexual intercourse or other sexual acts legally, but if one takes an indecent image of the other, they commit an offence under s.1 of that Act. Further, if they then sext that image (or any other) to their sexual partner, they are likely to be caught by both the illegal content safety duty and the reporting duty of the OSA. Before technology even gets involved, the sociological and legal pitfalls of 16-18-year-olds engaged in sexual activity are already a minefield. Adding technological solutions into the mix by requiring algorithms and other features to prevent an individual from encountering priority illegal content, including an indecent photograph of a child, will undoubtedly cause considerable and unpredictable problems.

The Act's focus on content removal and user-generated content platforms presupposes a technological fix to predominantly social problems. This technodeterministic approach ignores the role of platform design, algorithms, and opaque corporate cultures in perpetuating online harms. Moreover, content removal can stifle legitimate expression and dissent, raising concerns about freedom of speech and potential censorship.

Further, the Act's reliance on algorithmic solutions for content moderation raises concerns about bias, transparency, and accountability. Algorithmic bias can unfairly target marginalised groups, and the lack of transparency in these algorithms makes it difficult to assess their fairness and effectiveness. Furthermore, attributing sole responsibility to platforms for content moderation absolves the state of its role in addressing the root causes of online harms, such as social inequalities and a lack of digital literacy. Ultimately, the OSA risks becoming a technosolutionist fantasy, neglecting the multifaceted nature of online safety. A more nuanced approach that acknowledges the limitations of technology and prioritises social interventions alongside platform accountability is crucial for effectively tackling online harms in the UK.

3.3 The CSAM Proposal: A Technosolutionist Misstep?

While aiming to address a critical societal issue, the CSAM proposal has also sparked concerns about placing undue faith in technological solutions and mandatory scanning of encrypted content. First, the core tenet for compulsory scanning of encrypted content by platforms represents a technodeterministic approach that presumes a technological silver bullet to a deeply entrenched social ill. It overlooks the broader societal factors that contribute to the creation and dissemination of CSAM, including poverty, inadequate education, and limited access to support networks for offenders.[122]

Second, the technical feasibility and effectiveness of mandatory scanning remain questionable.[123] The amount of data involved, the complexities of encryption technologies, and the potential for false outcomes raise significant concerns about the accuracy and efficiency of such a system.[124] Furthermore, the Proposal's potential impact on the broader digital ecosystem cannot be ignored. Mandatory scanning

122. CRIN, *Privacy and Protection: A children's rights approach to encryption*, https://static1.squarespace.com/static/5afadb22e17ba3eddf90c02f/t/6528233c310af102f2d59c6e/1697129284489/Privacy+and+Protection+-+CRIN+defenddigitalme+encryption+report.pdf at 107, accessed 10 March 2025.

123. Abelson et al, n.79 above, 16.

124. European Parliamentary Research Service, above n.97.

could have a detrimental effect on innovation, security, and user trust in online platforms.[125] It could also disproportionately impact smaller platforms and privacy-focused services, hindering their ability to compete and innovate.[126]

It is essential to acknowledge that there are arguments in favour of the Proposal, including the need to protect children from online harm and the potential effectiveness of technological solutions in preventing the spread of CSAM. Online grooming conversations often adopt an informal, everyday tone. Rather than being overtly sexual, grooming typically involves building trust and gradually introducing explicit elements. Crucially, grooming conversations usually include logistical discussions, such as arranging a meeting place or providing travel instructions, that are not inherently sexualised but remain central to the offence. Despite this, the Impact Assessment explicitly acknowledges that speech detection is the most intrusive form of detection for users.[127] This raises significant concerns about the proportionality and necessity of AI-driven monitoring of private communications, particularly given the challenges in distinguishing benign interactions from grooming attempts.

However, these arguments must be carefully weighed against the potential risks and limitations of the proposed approach. Ultimately, the CSAM Proposal, in its current form, risks becoming a technosolutionist misstep. A more nuanced approach that prioritises social interventions, strengthens law enforcement cooperation, and fosters international collaboration alongside responsible technology development is crucial for effectively tackling CSAM online without compromising fundamental rights and freedoms. The Proposal presents a complex challenge that requires careful consideration of its potential benefits and drawbacks.

Having explored the intricate frameworks of the UK's OSA and the EU's CSAM Proposal, which underscore a commitment to SbD principles and proactive measures against the dissemination of CSAM, we stand at a critical juncture. These legislative efforts illuminate the growing consensus on the urgent need for robust regulatory mechanisms tailored to the complexities of the digital age. However, as we examine cyber-regulatory theory, a broader canvas of strategic responses to the scourge of CSAM comes into focus.

4. Lessons from Regulatory Theory

4.1 Introduction: The Limits of Techno-Solutionism in CSAM Regulation

CSAM is not a homogenous phenomenon; it presents distinct technological, legal, and investigative challenges that require differentiated regulatory responses. A techno-deterministic approach treats CSAM as a singular problem solvable through uniform detection techniques, failing to account for the evidentiary requirements, enforcement priorities, and rights-based concerns that vary across different forms of abusive material. The CSAM Proposal and the OSA exemplify a growing reliance on techno-solutionism. Both rely heavily on automated detection tools. Arguably, a more nuanced strategy is required—one that draws from past regulatory successes and failures. The history of cyberspace regulation is one of encountering and addressing complex challenges, from combating online fraud to regulating harmful content. These past efforts offer valuable lessons in navigating the tensions between security, privacy, and accountability.

125. Internet Architecture Board, 'IAB Statement on Encryption and Mandatory Client-side Scanning of Content' (IETF, 2023) https://datatracker.ietf.org/doc/statement-iab-statement-on-encryption-and-mandatory-client-side-scanning-of-content/ accessed 10 March 2025.

126. Recent negotiations appear to have resulted in mitigating the proposed detection order. According to the European Parliament, "to avoid mass surveillance or generalised monitoring of the internet, the draft law would allow judicial authorities to authorise [only] time-limited orders, as a last resort, to detect any CSAM and take it down or disable access to it (...). MEPs excluded end-to-end encryption from the scope of the detection orders to guarantee all users' communications are secure and confidential. Providers would be able to choose which technologies to use as long as they comply with the strong safeguards foreseen in the law (...)."; See P. Breyer, Chat Control https://www.patrick-breyer.de/en/posts/chat-control/; European Parliament, 'Child Sexual Abuse Online: Effective Measures, No Mass Surveillance' (Press Release, 10 November 2023) https://www.europarl.europa.eu/news/en/press-room/20231110IPR10118/child-sexual-abuse-online-effective-measures-no-mass-surveillance accessed 10 March 2025.

127. European Parliamentary Research Service, above n.97, 99.

To start, CSAM should be disaggregated into three categories to develop targeted regulatory strategies:

1. Known CSAM.
2. Unknown CSAM and
3. Grooming solicitations.

Known CSAM (previously identified material) is most effectively addressed through hash-matching technologies, which enable precise detection with minimal risk of false positives in law enforcement databases.[128] Unknown CSAM (newly created or AI-generated material) presents more significant challenges as classifiers struggle with age estimation, contextual interpretation, and adversarial evasion techniques.[129] Addressing this category requires a risk-based regulatory model that balances automated detection with robust oversight and human review. Conversely, grooming solicitations heavily depend on context, which poses a challenge to automated systems.[130] This category necessitates networked, communitarian responses, including user education, platform interventions (e.g., nudges, real-time moderation), and enhanced collaboration with law enforcement. By calibrating regulatory responses to the nature of the CSAM, policymakers can develop a more effective and proportionate framework that mitigates harm while upholding privacy, security, and due process.[131]

4.2 Categorising CSAM: A Problem-Specific Regulatory Response

4.2.1 Known CSAM (Previously Identified Material)

Detecting known CSAM is one of the most technologically mature and legally established aspects of online child protection. Platforms and law enforcement agencies rely primarily on hash-matching technology, which compares digital fingerprints (cryptographic hashes) of previously identified CSAM against user-uploaded content. Industry-standard tools such as PhotoDNA, Google's CSAI Match, and the UK's Internet Watch Foundation (IWF) hash database enable near-real-time detection and removal of previously recorded abuse material with high accuracy.[132] Unlike classifiers that rely on probabilistic assessments, hash-matching provides deterministic detection; a flagged file is either an exact match or not.

Despite its efficacy, hash-matching remains dependent on access to verified law enforcement databases, such as those maintained by the National Center for Missing and Exploited Children (NCMEC) in the US or the proposed EU Centre under the CSAM Regulation.[133] These repositories aggregate hashes of identified CSAM, facilitating cross-platform enforcement. Platforms, such as Meta, Google, and Microsoft, are legally required under US law (through NCMEC's CyberTipline) to report identified CSAM, and similar obligations are emerging under the CSAM Proposal.[134] and the OSA.

This ecosystem creates a triangular enforcement model, where platforms detect, law enforcement verifies, and regulators oversee compliance. It demonstrates Lessig's argument that technological design can

128. M. Steinebach, 'An Analysis of PhotoDNA' ARES '23: Proceedings of the 18th International Conference on Availability, Reliability and Security https://doi.org/10.1145/3600160.360504 accessed 10 March 2025.
129. European Parliamentary Research Service, above n.98; Security Researchers and Academics, 'Open Letter: Hundreds of Scientists Warn Against EU's Proposed CSA Regulation' (European Digital Rights, 12 July 2023) https://edri.org/our-work/open-letter-hundreds-of-scientists-warn-against-eus-proposed-csa-regulation/ accessed 10 March 2025.
130. N. Mylonas et al., 'Online Child Grooming Detection: Challenges and Future Directions' in Gkotsis et al. (eds) *Paradigms on Technology Development for Security Practitioners* (Springer, 2025).
131. A. van der Spuy, S. Witting, P. Burton, E. Day, S. Livingstone, & K. R. Sylwander, Guiding Principles for Addressing Technology-Facilitated Child Sexual Exploitation and Abuse (LSE 2024) https://eprints.lse.ac.uk/126219/ accessed 24 March 2025.
132. European Commission, 'Impact Assessment Accompanying the Proposal for a Regulation on Preventing and Combating Child Sexual Abuse' (SWD(2022) 209 final); Electronic Frontier Foundation (EFF), 'Why Automated Systems Fail at Detecting Grooming' (2023); See also Berkman Klein Center for Internet & Society at Harvard University, 'The Limits of Automated Detection of Online Grooming' (2022); Council of Europe, 'Challenges in Detecting Online Grooming: A Human Rights Perspective' (2022).
133. National Center for Missing and Exploited Children (NCMEC), '2022 Annual Report' (2023); European Parliamentary Research Service (EPRS), 'The EU's Proposed CSAM Regulation: Key Issues and Challenges' (2023).
134. Article 10, CSAM Proposal, Online Safety Act 2023 (UK), Part 3

implement legal and regulatory norms.[135] This represents a code-based enforcement model, wherein compliance is self-executing and immediate through code-based enforcement delivered by embedding regulatory constraints directly into digital architectures, reducing reliance on retrospective legal remedies.[136] The CSAM Proposal extends this model, hardcoding child protection into platform architectures.[137]

However, this model is not without challenges. Encryption poses a fundamental barrier—with end-to-end encrypted (E2EE) communications on WhatsApp, Signal, and Apple iMessage, hash-matching becomes infeasible unless content is scanned pre-encryption on-device. This proposal raises concerns due to its privacy implications.[138] Proposals for on-device scanning raise significant concerns regarding user privacy, security, and the potential for function creep in surveillance mechanisms.[139] Additionally, while hash-matching excels at detecting known material, it struggles to identify newly generated CSAM or AI-synthesised content.[140] With the rise of generative AI, this limitation has become increasingly relevant as existing detection frameworks struggle to adapt to novel, previously unseen content.[141] Without balance, the regulatory proposals risk under-enforcement (due to encryption barriers) or overreach (through mass surveillance mandates).[142]

4.2.2 Unknown CSAM (Newly Created or AI-Generated Material)

Detecting unknown CSAM—newly created or AI-generated material—presents substantial technological, legal, and enforcement challenges. Unlike known CSAM, which relies on hash-matching for detection, unknown material requires probabilistic classifiers and AI-based models that operate with varying degrees of accuracy.[143] These models attempt to assess image features (such as nudity and contextual markers) or language patterns (for textual grooming analysis), but often struggle with context, intent, and age classification—three critical dimensions in distinguishing lawful from unlawful content.[144] As a result, automated detection becomes highly prone to error.[145], with a significant risk of false outcomes.[146]

AI-generated CSAM, created via deepfake technology, complicates legal definitions, as jurisdictions vary on criminalising synthetic exploitative content.[147] These synthetic materials complicate the legal landscape:

---

135. This foundational work introduces the concept that code functions as a form of regulation, shaping behaviour in digital spaces. It is particularly relevant to understanding how platforms enforce CSAM laws through technological design; European Parliamentary Research Service (EPRS), 'The Role of Technology in Enforcing Online Safety Regulations' (2023); Internet Watch Foundation (IWF), 'Automated Detection of CSAM: Challenges and Opportunities' (2022).

136. T. Gillespie, 'Platforms Are Not Intermediaries' (2018) 2 *Georgetown Law Technology Review* 198 https://georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Gilespie-pp-198-216.pdf accessed 10 March 2025.

137. Articles 3–4, CSAM Proposal.

138. Ludvigsen, Nagaraja, & Daly, n.79, above; See also Abelson et al, n.79, above.

139. European Parliamentary Research Service (EPRS), *Detecting Child Sexual Abuse Material Online: The Impact of AI and Encryption* (Study, European Parliament 2023) 59 https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf accessed 2 May 2025; citing Fact Sheet: Client-Side Scanning, Internet Society, September 2022, accessed 30 March 2023; On function creep, see B-J. Koops, 'The concept of function creep', Law, Innovation and Technology, Vol 13(1), Routlegde, pp.2956, 2021

140. ActiveFence, 'Detecting Novel CSAM – Why Image Hash Matching Isn't Enough Anymore' (ActiveFence, 2023) https://www.activefence.com/#:~:text=One%20major%20challenge%20is%20its,bypass%20hash%2Dmatching%20detection%20systems. accessed 10 March 2025.	

141. L. Struppek et al., 'Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash' *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022 https://arxiv.org/abs/2111.06628 accessed 10 March 2025.

142. European Parliamentary Research Service, above n.97.

143. B. Westlake & E. Guerra, 'Using file and folder naming and structuring to improve automated detection of child sexual abuse images on the Dark Web' *Forensic Science International: Digital Investigation* (2023). https://doi.org/10.1016/j.fsidi.2023.301620; European Commission, Joint Research Centre, *Technological Solutions to Detect Child Sexual Abuse in End-to-End Encrypted Communications* https://ec.europa.eu/jrc/en/publication/technological-solutions-detect-child-sexual-abuse-end-end-encrypted-communications both accessed 10 March 2025.

144. D. Cook et al., 'Can We Automate the Analysis of Online Child Sexual Exploitation Discourse?' (2022) arXiv https://arxiv.org/abs/2209.12320 accessed 10 March 2025.

145. K. Parti & J. Szabó, 'The Legal Challenges of Realistic and AI-Driven Child Sexual Abuse Material: Regulatory and Enforcement Perspectives in Europe' (2024) 13 *Laws* 67 https://doi.org/10.3390/laws13060067 accessed 10 March 2025.

146. R. Anderson, 'Chat Control or Child Protection?' (2022) arXiv:2210.08958 [cs.CY] https://arxiv.org/abs/2210.08958 accessed 10 March 2025.

147. Internet Watch Foundation, 'How AI is Being Abused to Create Child Sexual Abuse Imagery' (IWF, 2023) https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/ accessed 10 March 2025.

some jurisdictions criminalise Computer-Generated Child Sexual Abuse Material (CG-CSAM) outright, while others define content as illegal only when it involves a real child.[148] Already stretched law enforcement resources face an escalating volume of flagged material, requiring additional layers of forensic analysis to differentiate authentic abuse imagery from synthetic content.[149] Algorithmic bias exacerbates these risks. Facial recognition models may misclassify teenagers as adults or vice versa, leading to wrongful criminal suspicion or under-detection of genuine CSAM.[150] These challenges underscore the need for human oversight in AI-assisted moderation, rather than relying solely on fully automated detection.

Rather than mandatory blanket surveillance, a risk-based regulatory approach is essential to mitigating these risks. Unlike the CSAM Proposal, which mandates platform-wide scanning of private communications, a targeted, risk-tiered system that focuses detection efforts where harm is likely to occur, offers a more proportionate and rights-preserving alternative. As discussed in the following section, Laidlaw's gatekeeper theory provides a critical lens through which to assess the role of platforms as regulatory intermediaries. As digital gatekeepers, platforms do not merely host content; they actively shape information flows, risk governance, and regulatory compliance. Their role carries a dual responsibility: ensuring robust mechanisms for detecting and preventing CSAM while upholding fundamental rights, including privacy, due process, and freedom of expression. A pre-emptive, indiscriminate scanning model, such as that envisioned in the CSAM Proposal, risks collapsing these competing obligations into a techno-deterministic enforcement paradigm, where algorithmic detection becomes the *de facto* regulatory instrument, regardless of its efficacy or proportionality.

A normatively sound gatekeeping model must embed principles of procedural justice, legal accountability, and institutional oversight within platform governance structures. Achieving this shift requires transitioning from blanket surveillance to a more differentiated, context-sensitive approach, where enforcement mechanisms are based on demonstrable risk rather than default technological mandates. Platforms must function as nodes within a distributed regulatory ecosystem, where risk mitigation strategies are co-produced through engagement with independent oversight bodies, civil society actors, and law enforcement agencies rather than dictated solely by legislative fiat or technical capabilities.[151]

From a structural regulatory perspective, addressing these challenges requires a departure from automated determinism and a recalibration toward a multi-tiered enforcement framework, where interventions are proportional to risk exposure, evidence-based, and subject to independent adjudication. This approach aligns with a principled regulatory approach that balances Lessig's architectural constraints with Murray's network communitarianism,[152] leveraging platform design, user-driven governance, and *ex-post* enforcement mechanisms to regulate harmful content without entrenching mass surveillance or undermining encrypted communications. The challenge, therefore, is not merely technological but institutional and normative: how to construct a governance architecture that is neither techno-utopian in its reliance on AI-based detection nor techno-pessimistic in its dismissal of regulatory intervention, but instead adaptive, responsive, and rights-preserving within a complex digital landscape.

4.2.3 Grooming & Solicitation

Grooming is inherently dynamic, adaptive, and context-dependent, making it one of the most challenging forms of online child exploitation to regulate. Grooming involves progressive behavioural manipulation that unfolds over time. Perpetrators exploit linguistic ambiguity, social engineering, and deception, often migrating conversations across platforms or into encrypted spaces to evade detection. Automated detection

148. Parti & Szabó, n.146 above; Connecticut General Assembly, 'Criminalization of Computer-Generated Child Sexual Abuse Material' (OLR Research Report, 2024-R-0167, 31 January 2024) https://www.cga.ct.gov/2024/rpt/pdf/2024-R-0167.pdf accessed 10 March 2025.
149. Wilson Center, 'Combatting AI-Generated CSAM' (Wilson Center, 2024) https://www.wilsoncenter.org/article/combatting-ai-generated-csam accessed 10 March 2025.
150. T. Ganel, C. Sofer & M.A. Goodale, 'Biases in Human Perception of Facial Age Are Present and More Exaggerated in Current AI Technology' (2022) 12 *Scientific Reports* 22519 https://doi.org/10.1038/s41598-022-27009-w accessed 10 March 2025.
151. E.B. Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights, and Corporate Responsibility* (CUP, 2020).
152. A. Murray, *The Regulation of Cyberspace: Control in the Online Environment* (Routledge-Cavendish 2006).

remains unreliable, as AI classifiers struggle to interpret the nuanced context in text-based interactions.[153] While natural language processing models can flag specific keywords or phrases associated with predatory behaviour, they are highly prone to false outcomes.[154] Lorenzo-Dus et al. highlight that online groomers frequently manipulate language to obscure their abusive intentions, particularly during the trust-building phase. The linguistic strategies employed can vary significantly among perpetrators, further complicating detection efforts. This underscores the challenges of relying solely on automated or AI-driven language recognition tools for identifying grooming behaviour.[155] A teenager discussing relationships may inadvertently trigger the same alerts as a predator, while sophisticated offenders who deliberately avoid explicit language may evade detection.[156] The fundamental difficulty of codifying intent, power dynamics, and coercion means that blanket algorithmic monitoring is ineffective and presents significant risks to privacy and free expression.[157]

The complexities of regulating CSAM—whether in the form of known, unknown, or grooming-related exploitation—highlight the fundamental limitations of enforcement models that rely on automated detection, platform mandates, or reactive content moderation. Accordingly, the following section argues that a more effective and sustainable regulatory framework must draw from established regulatory theories that account for the interplay between technological enforcement, legal oversight, and institutional accountability. Code-based regulatory approaches provide insights into the strengths and weaknesses of embedding enforcement within digital infrastructures. At the same time, network communitarianism emphasises the need for multi-stakeholder governance models that strike a balance between platform responsibilities and societal norms. Gatekeeper theory offers a framework for understanding platforms not merely as passive hosts but as active regulatory intermediaries with obligations to structure content dissemination in ways that mitigate harm. In contrast to deterministic enforcement models, risk regulation enables more proportionate interventions by targeting regulatory scrutiny where it is most needed rather than imposing indiscriminate surveillance mechanisms. Applying these historical lessons from cyber-regulatory theory allows for a recalibrated approach to CSAM regulation that moves beyond techno-solutionism to integrate a more adaptive, accountable, and rights-preserving enforcement framework.

5. Application of Regulatory Theory

5.1 Code as Law – The Over-Reliance on Safety-by-Design

Integrating SbD principles into digital regulation reflects a growing reliance on code as a means of enforcement. Lessig's seminal argument that "Code is Law" illustrates how technological architectures can function as de facto regulatory tools, dictating permissible actions and structuring online environments to render non-compliance technically infeasible rather than legally impermissible.[158] In the context of CSAM regulation, safety-by-design manifests through pre-emptive content moderation systems, such as hash-matching technologies (PhotoDNA), automated classifiers, and real-time scanning mechanisms, which allow platforms to detect, filter, and remove illicit material before it reaches users.

However, the over-reliance on code-based solutions raises significant concerns about proportionality, transparency, and due process.[159] First, these systems inherently operate *ex-ante*, pre-emptively blocking content without judicial oversight, shifting enforcement from *ex-post* legal adjudication to automated

153. Mylonas et al., above n.131; Social and Economic Council (SER), 'SER Points to the Risks of the Use of Emotional AI, Including Language Recognition AI' (SER, 2024) https://www.ser.nl/nl/actueel/Nieuws/emotion-ai accessed 10 March 2025.

154. See n.99.

155. N. Lorenzo-Dus, C. Evans, and Re. Mullineux-Morgan, *Online Child Sexual Grooming Discourse* (CUP, 2023) 70–71.

156. S.K Witting, 'Regulating Bodies: Child Sexuality in the Digital Era' (2019) *Zeitschrift für Internationale Strafrechtsdogmatik* 1, 5 https://www.nomos-elibrary.de/10.5771/2193-7869-2019-1-5.pdf accessed 10 March 2025.

157. Z. Lan & A. Turchin, 'Impact of possible errors in natural language processing-derived data on downstream epidemiologic analysis' 6(4) *JAMIA Open* December 2023 https://academic.oup.com/jamiaopen/article/6/4/ooad111/7502595 accessed 10 March 2025.

158. L. Lessig, *Code v2* (Basic Books 2006) 123–127, 133–136.

159. A. de Streel et al, 'Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform' (Study requested by the IMCO Committee, European Parliament, June 2020) 12 https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf accessed 10 March 2025.

decision-making. Pre-emptive blocking risks chilling legitimate content (e.g., sexual health education) and expanding encryption backdoors undermines privacy.[160] Second, integrating encrypted messaging backdoors undermines core privacy protections by embedding state-mandated surveillance into digital infrastructures. Such measures are fundamentally at odds with democratic principles, effectively dismantling encryption as a tool for secure communications, whistleblower protections, and journalistic confidentiality.[161]

A more proportionate and rights-respecting approach requires recalibrating code as a regulatory tool, ensuring that SbD measures are context-aware, narrowly tailored, and subject to independent oversight.[162] Rather than mandating blanket surveillance, regulatory frameworks should adopt a tiered enforcement model, where interventions scale with demonstrated risk rather than imposing indiscriminate scrutiny. Implementing this approach requires human-in-the-loop oversight, where human moderation teams supplement algorithmic decisions. Ensuring responsiveness to context and alignment with due process protections depends on integrating human oversight into enforcement mechanisms. By embedding principles of necessity, proportionality, and accountability into SbD systems, it is possible to mitigate the risks of regulatory overreach while maintaining a robust child protection framework that leverages technological interventions without subverting fundamental rights.

5.2 Network Communitarianism – The Need for Social Solutions to Social Problems

Digital platform regulation extends beyond the simplistic dichotomy of state versus platform self-governance. A dynamic interplay of law, social norms, and technical architectures shapes it. Network communitarianism offers a compelling alternative to rigid statutory controls and unaccountable self-governance, recognising that social problems necessitate social solutions and providing a viable framework for policy design in CSAM regulation. Network communitarianism is a socio-technological paradigm that reconfigures traditional communitarian values within digital ecosystems, emphasising the interplay between decentralised participation, algorithmic mediation, and shared governance. It posits that communities no longer form solely through geographic or cultural proximity. Instead, they emerge from dynamic, technology-enabled interdependencies, where norms, trust, and collective agency develop through iterative digital interactions. Unlike classical communitarianism, which roots itself in the moral and social fabric of close-knit societies, network communitarianism acknowledges the fluidity of digital affiliations, the influence of platform architectures on communal bonds, and the contested spaces where algorithmic governance both enables and constrains collective identity formation. It critically examines how power, agency, and solidarity shape interactions in networked environments, where the balance between participatory autonomy and systemic control remains in a state of constant flux.

Community-based moderation provides a valuable lens through which to explore the potential of decentralised regulation. Platforms like Reddit demonstrate effective community moderation, though challenges include inconsistent enforcement and moderator bias.[163] Therefore, effective policy design must ensure that community-driven moderation efforts are not exploited as a substitute for systemic safeguards but integrated into a structured regulatory approach. Platforms must provide clear guidelines, training, and institutional support to moderators while ensuring that decisions remain subject to external oversight and due process.

Network communitarianism underscores the potential of community-driven moderation as a complementary mechanism to platform-led and regulatory interventions. Decentralised moderation models illustrate how

160. J.F. Gomez et al, 'Algorithmic Arbitrariness in Content Moderation' (arXiv, 26 February 2024) 1–2, 5–6, 14 https://arxiv.org/abs/2402.16979 accessed 10 March 2025.

161. Center for Democracy & Technology, 'Chilling Effects on Content Moderation Threaten Freedom of Expression for Everyone' (CDT, 2023) https://cdt.org/insights/chilling-effects-on-content-moderation-threaten-freedom-of-expression-for-everyone/ accessed 10 March 2025.

162. Brookings Institution, 'Using 'safety by design' to address online harms' (Brookings, 2022) https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/ accessed 10 March 2025.

163. A. Fang, W. Yang, & H. Zhu, 'Shaping Online Dialogue: Examining How Community Rules Affect Discussion Structures on Reddit' (2023) arXiv https://arxiv.org/abs/2308.01257; H.M. Wang, B. Bulat, S. Fujimoto, & S. Frey, 'Governing for Free: Rule Process Effects on Reddit Moderator Motivations' (2022) arXiv https://arxiv.org/abs/2206.05629 both accessed 10 March 2025.

engaged user communities can identify and remove harmful content. The agility of such frameworks, where distributed actors can respond in real time, offers a level of contextual intelligence that large-scale algorithmic moderation lacks.

For network communitarianism to be a viable component of CSAM policy design, policymakers must integrate it into a broader co-regulatory framework that balances community-driven initiatives with statutory mandates and enforcement mechanisms. Regulatory models such as the DSA and Australia's eSafety Commissioner framework demonstrate how platforms can operate under clear legal obligations while retaining flexibility in enforcement.[164] These frameworks provide structured oversight mechanisms, including independent audits, transparency reporting, and accountability benchmarks, ensuring that participatory governance models do not become de facto strategies for deregulation.[165] Moreover, a structured collaboration between platforms, civil society organisations, and regulatory authorities is essential to maintaining an adaptive and resilient governance architecture. Policies should facilitate information-sharing mechanisms, enabling regulators to access meaningful transparency data while respecting fundamental rights, including privacy and due process. Ensuring that community-driven safety initiatives operate within defined regulatory frameworks enhances their legitimacy while mitigating the risks of arbitrary or inconsistent enforcement.

A fundamental principle of this approach is integrating user agency within structured governance frameworks. User empowerment through reporting mechanisms, safety tools, and digital literacy initiatives is crucial, but it must be part of a broader system of platform accountability and regulatory oversight. Effective governance structures must ensure that platforms do not shift the burden of child protection onto individuals without corresponding systemic safeguards. Instead, user participation in content moderation should be supplementary rather than substitutive, reinforcing platform responsibility rather than diluting it. Network communitarianism offers a viable alternative to the limitations of rigid state control and the shortcomings of platform self-regulation. Embedding user participation within structured regulatory obligations creates a governance model that is both scalable and adaptive. This approach ensures that community-driven safety mechanisms operate within legal frameworks, striking a balance between effectiveness and the protection of fundamental rights.

5.3 The Gatekeeping Role of Platforms

Gatekeepe theory articulates platforms' heightened responsibilities as intermediaries, ensuring that content governance aligns with legal mandates, fundamental rights, and evolving societal norms.[166] Regulatory frameworks governing online content moderation impose heightened obligations on digital platforms, mandating more expansive detection, reporting, and removal mechanisms to combat illicit material. Recent legislative developments reflect a shift toward proactive enforcement, requiring platforms to strike a balance between legal compliance and considerations of fundamental rights.[167] However, these obligations often introduce structural tensions between child protection imperatives and fundamental rights safeguards, specifically regarding privacy, due process, and freedom of expression. This regulatory dilemma reflects broader challenges in striking a balance between security-driven interventions and the preservation of digital rights.[168] The proliferation of automated enforcement mechanisms, including AI-driven content classification, real-time scanning, and cross-platform information-sharing mandates, risks entrenching a regime of ambient surveillance in which *ex ante content control supplants traditional ex post* judicial oversight.[169]

164. Global Online Safety Regulators Network, Position Statement: Regulatory coherence and coordination: the role of the Global Online Safety Regulators Network (April 2024) https://www.esafety.gov.au/sites/default/files/2024-05/GOSRN-Position-Statement-on-Regulatory-Coherence.pdf accessed 10 March 2025.

165. Rasha Abdul Rahim, 'Cutting Through the Jargon: Independent Audits in the Digital Services Act' (Mozilla Foundation Blog, 6 March 2024) https://foundation.mozilla.org/en/blog/cutting-through-the-jargon-independent-audits-in-the-digital-services-act/ accessed 24 March 2025; J. Owono & B. Ricks, 'Using Safety by Design to Address Online Harms' (Stanford Cyber Policy Center, 15 February 2024) https://cyber.fsi.stanford.edu/news/using-safety-design-address-online-harms accessed 24 March 2025.

166. Laidlaw, n.151 above.

167. G. De Gregorio, *Digital Constitutionalism: The Role of Internet Platforms* (CUP, 2022).

168. J. Slupska, 'Child Protection and Digital Rights: Conflicting Legal Frameworks in Online Content Regulation' (2021) 17 *International Journal of Law and Information Technology* 89.

169. M. Hildebrandt & B-J. Koops, 'The Challenges of Ambient Law and Legal Protection in the Profiling Era' (2010) 73 *Modern Law Review* 428.

The challenge is not merely technological feasibility but regulatory proportionality. The indiscriminate application of automated moderation regimes risks collapsing nuanced legal distinctions, treating lawful but sensitive expression, investigative journalism, and digital human rights advocacy as collateral damage in broad-spectrum enforcement efforts.[170] A principled regulatory approach demands a recalibration towards risk-sensitive regulation, where intervention measures are tiered, proportionate, and subject to independent adjudication.[171] For example, the OSA requires certain service providers to assess reasonably foreseeable risks to individuals in the UK.[172] Platforms must align enforcement mechanisms with legal norms to ensure moderation strategies uphold democratic commitments.[173]

Effective CSAM regulation requires robust transparency and accountability mechanisms beyond perfunctory compliance disclosures. These mechanisms ensure institutional legitimacy and uphold fundamental legal and ethical standards. Given the high stakes in combating CSAM, policymakers must treat systematic transparency reporting as a regulatory requirement. This reporting should include detailed information on detection methodologies, automated enforcement error rates, false outcomes, due process safeguards, and the proportionality of interventions. Without scrutiny, enforcement structures risk operating in a state of legal and ethical ambiguity, which increases the likelihood of overreach, rights violations, and ineffective intervention.

Transparency reporting alone cannot mitigate the risks associated with opaque or overly expansive surveillance-based enforcement models. The complexities of CSAM detection, primarily utilising probabilistic AI classifiers, hash-matching technologies, and metadata analysis, necessitate external verification beyond self-reported compliance data. Independent third-party audits must become an integral part of regulatory oversight. Regulatory bodies, digital rights organisations, technical experts, and child protection advocates should conduct these audits. Their role is to rigorously evaluate the accuracy, proportionality, and impact on fundamental rights of automated and human moderation processes. Oversight ensures that interventions remain narrowly tailored, rights-respecting, and effective in addressing exploitation without undermining privacy or encryption.

Auditability is a fundamental requirement for legitimacy in CSAM enforcement. Without mechanisms for verifiability, procedural fairness, and independent oversight, transparency commitments risk becoming performative exercises in corporate self-regulation. A regulatory framework that balances child protection imperatives with rights-preserving enforcement must be grounded in evidence-based policymaking. Achieving this balance requires a multi-stakeholder approach that prioritises transparency, auditability, and safeguards against both under- and over-enforcement. Embedding these principles into CSAM regulation is essential to developing a strategy that is both effective in protecting children and resilient against the unintended consequences of unchecked surveillance.

The absence of accountability measures enables platforms to function as unconstrained arbiters of digital expression, effectively operating as private regulatory entities with unilateral decision-making authority over online speech and safety frameworks. Without reciprocal institutional checks, this concentration of power poses a structural threat to democratic governance, shifting content moderation decisions from legal institutions to private corporate actors whose incentives may not align with human rights obligations, due process norms, or public interest considerations. Regulatory frameworks must move toward multi-stakeholder governance models, where co-regulation mechanisms ensure platforms remain accountable to legal, civic, and technical oversight bodies, preventing power asymmetry. Establishing independent content adjudication structures and enforceable due process safeguards is imperative to curtail discretionary

170. E. Douek, 'Content Moderation as Systems Thinking' (2021) 136 *Harvard Law Review* 526.
171. U. Gasser & W. Schulz, 'Governance by Transparency: The Emerging Third Wave of Regulation for the Digital Age' (2020) 21 *German Law Journal* 1390.
172. OSA, s.9.
173. Council of Europe, Regulating content moderation on social media to safeguard freedom of expression (2023) https://rm.coe.int/as-cult-regulating-content-moderation-on-social-media-to-safeguard-fre/1680b2b162 accessed 10 March 2025; E. Douek, 'Content Moderation as Systems Thinking' (2022) 136 *Harvard Law Review* 526.

overreach, mitigate algorithmic opacity, and preserve the fundamental balance between platform integrity and user rights in the digital ecosystem.

However, the foundational question remains, even within a co-regulatory framework: how should risk be conceptualised and operationalised in platform governance? The increasing reliance on pre-emptive surveillance measures, such as mandatory AI-driven detection and real-time content scanning, reflects a broader regulatory trend towards securitised risk management—one that presumes that harm can be neutralised through ex-ante intervention rather than reactive enforcement. This assumption, however, overlooks the well-documented shortcomings of pre-emptive regulatory strategies, including overreach, misallocation of enforcement resources, and infringement of fundamental rights.

A risk-based regulatory model offers a more proportionate, targeted, and adaptive alternative to blanket detection mandates, aligning enforcement mechanisms with demonstrable threats rather than speculative harms. Drawing from Beck's risk society theory, which critiques the governance of uncertainty through securitisation, digital child protection frameworks need a shift from mass surveillance imperatives to structured, evidence-based risk mitigation strategies. Risk-based approaches have already demonstrated efficacy in cybersecurity and financial regulation domains, where tiered interventions, adaptive compliance mechanisms, and probabilistic threat modelling have successfully mitigated harm without compromising fundamental legal principles. Applying this logic to CSAM detection necessitates a transition towards risk-tiered detection orders, where intervention scales with the severity of risk, the nature of the platform, and the operational feasibility of enforcement measures. Rather than mandating indiscriminate scanning across all digital communications, a tiered enforcement structure, underpinned by independent oversight and industry-wide best practices, offers a more legitimate and functionally effective regulatory model that ensures child protection and the preservation of fundamental rights.

5.4 Risk Regulation – Why a Risk-Based Approach Outperforms Blanket Surveillance

The CSAM Proposal and the OSA exemplify a deeply flawed regulatory instinct: the securitisation of risk through pre-emptive surveillance mandates. By compelling platforms to deploy AI-driven detection tools across all communications, these regimes assume that mass data interception and algorithmic enforcement will eliminate risk. However, as Beck's risk society theory illustrates, such pre-emptive regulatory strategies often generate more systemic vulnerabilities than they resolve.[174] Beck argues that modern governance increasingly shifts from reacting to harm to attempting to eliminate hypothetical risks before they materialise. In doing so, regulators create unintended consequences, redistribute risk unevenly, and erode fundamental rights under the guise of security enhancement.[175]

The limits of pre-emptive regulation become apparent in the context of CSAM detection. Blanket scanning mandates, such as those proposed under the CSAM Proposal and s.121 of the OSA, operate on the fallacy that risk is static, quantifiable, and best mitigated through universal intervention. Pre-emptive mass surveillance risks generating high false-positive rates, overwhelming law enforcement with non-actionable reports, and disproportionately infringing upon encrypted communications—a cornerstone of digital privacy and security.[176] Rather than ensuring child protection, these measures entrench an algorithmic overreach system that compromises the rights they purport to safeguard.

A risk-based regulatory model offers a more proportionate and effective alternative. Risk-tiered cybersecurity and financial regulation frameworks have efficiently allocated enforcement resources while mitigating systemic vulnerabilities.[177] Financial crime prevention, for instance, employs risk-based due diligence

174. Beck, n.46 above.

175. Ibid, 21–23, 36–40, 50–55.

176. European Parliament, 'Resolution on Artificial Intelligence in Criminal Law and Its Use by the Police and Judicial Authorities in Criminal Matters' (6 October 2021) P9_TA(2021)0405 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52021IP0405 accessed 10 March 2025.

177. Articles 4, 5, 6, and 16 Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014, and (EU) 2016/1011 [2022] OJ L333/1. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2554 accessed 10 March 2025.

protocols. In these protocols, institutions tailor their monitoring intensity based on demonstrable risk factors rather than imposing universal scrutiny on all transactions.[178] Similarly, cybersecurity regulation prioritises critical infrastructure and high-risk threat vectors rather than indiscriminately surveilling all online activity.[179] By calibrating enforcement measures to the specific risks associated with different platforms, regulators can ensure the effective and legally justifiable use of detection technologies. Judicial oversight would play a critical role in determining when and how scanning technologies should be authorised, ensuring their use is both necessary and proportionate to the identified risk.

A risk-tiered model would distinguish between platforms based on their operational characteristics, technological capabilities, and the likelihood of misuse for the distribution of CSAM. Platforms that facilitate anonymity enable rapid content dissemination or support interoperable messaging across multiple services, presenting heightened risks and potentially requiring stricter regulatory obligations. These platforms may be subject to enhanced scrutiny, with a responsibility to implement more robust safeguards while remaining accountable for the necessity and proportionality of their interventions. In contrast, services with demonstrably lower risk profiles could implement targeted, proportionate safeguards without being compelled to adopt indiscriminate content scanning.

This differentiated approach would provide an alternative to blanket surveillance mandates that impose identical obligations across diverse online services. A coherent, risk-based framework would replace the undifferentiated enforcement paradigm with a regulatory model integrating child protection objectives with digital security and privacy considerations. The failure to adopt a calibrated approach risks entrenching an indiscriminate surveillance model operating under the pretext of safety without adequately considering the broader legal, ethical, and societal implications. A more sophisticated regulatory design, grounded in risk assessment and judicial oversight, is necessary to ensure that CSAM detection orders serve their intended purpose without eroding fundamental rights or undermining the structural integrity of digital communications.

6. Towards a More Balanced Approach

The CSAM Proposal and the OSA embody a technocratic overreach, relying on blanket surveillance, indiscriminate AI-driven detection, and regulatory absolutism to combat CSAM. These approaches fail to account for the complexity of the problem, the limitations of technological enforcement, and the necessity of regulatory proportionality. A more balanced, effective, and rights-preserving model must integrate three distinct but complementary regulatory pillars.

6.1 Code-Based Detection for Known CSAM

Platforms should maintain and enhance their deployment of code-based detection systems to identify and remove previously verified (Known) CSAM. Hash-matching technologies, such as PhotoDNA, Google's CSAI Match, and the hash databases managed by the EU Center, provide an essential technical safeguard, enabling the rapid and precise detection of known illegal content without exposing human moderators to the associated harms of direct review. These systems allow cross-platform collaboration while minimising privacy intrusions. This form of architectural enforcement aligns with Lessig's classic conception of "Code as Law", wherein the technical design of digital spaces establishes *de facto* regulatory mechanisms, shaping user behaviour and enforcing legal constraints. When implemented within statutory frameworks, such mechanisms can serve as narrowly tailored regulatory interventions, achieving legal certainty while safeguarding fundamental rights.

178. Articles 8, 13, 15, 18 Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 amending Directive (EU) 2015/849 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing and amending Directives 2009/138/EC and 2013/36/EU [2018] OJ L156/43.

179. Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS2 Directive) [2022] OJ L333/80.

However, effective online harm prevention requires more than reactive detection of illegal content. Beyond user-focused initiatives, embedding proactive friction-by-design interventions within platform architecture is vital for pre-emptively disrupting grooming pathways. Drawing from behavioural science, friction-based design principles intentionally introduce 'sludge' to harmful interactions, requiring additional steps or creating natural pauses that disrupt exploitative behavioural flows. Real-time prompts—drawing on empirical evidence from interventions in online fraud prevention—serve as effective 'cognitive speed bumps' to deter predatory conduct.[180] For instance, platforms like Instagram have introduced warnings when adults attempt to message minors without prior connection, reducing unsolicited adult-to-minor contact by measurable margins.[181] Second, age verification mechanisms, while imperfect, provide an additional line of defence when responsibly deployed, narrowing adult access to spaces designed for minors.[182] Additionally, platform features such as delayed image transmission, two-sided consent for private messaging[183], and enforced 'cooling-off' periods between connection requests materially constrain grooming strategies without resorting to mass surveillance. These design choices embed friction, transparency, and user empowerment into the platform ecosystem, offering a credible, rights-respecting alternative to invasive surveillance while systematically discouraging exploitation at the infrastructural level.

Nevertheless, these technological interventions should not be over-relied upon. Structural reforms must resist technosolutionism's asymmetries and fail to build community-centred digital environments. Proactive design interventions and automated detection systems alike should be framed as complements to, rather than substitutes for, sustained human oversight, robust accountability structures, and legislative frameworks aligned with international human rights standards. At the same time, caution is essential when expanding code-based enforcement beyond its current bounds. While hash-matching identifies previously verified CSAM, expanding scanning mechanisms to target new material fundamentally alters the regulatory landscape, from detecting illegal content to making probabilistic determinations about as-yet-unverified material. The risks of false positives, misclassification, and undue reliance on law enforcement by algorithmic outputs raise significant concerns about due process, proportionality, and the potential criminalisation of lawful content.[184] The history of AI-driven content moderation failures—particularly in detecting nudity, artistic expression, and lawful but sensitive material—illustrates the danger of treating code-based interventions as a panacea for online safety challenges.[185] Any expansion of automated scanning beyond previously verified content must be subject to rigorous legal, technical, and ethical scrutiny, ensuring that enforcement mechanisms remain proportionate, rights-compliant, and aligned with democratic oversight. Failure to establish these boundaries risks embedding flawed algorithmic decision-

180. National Australia Bank, 'Why Helpful Friction is Crucial in the Battle Against Scammers' (NAB News, 5 March 2024)https://news.nab.com.au/news/why-helpful-friction-is-crucial-in-the-battle-against-scammers, accessed 28 April 2025; PYMNTS, 'Behavioral Analytics Inject "Smart Friction" into Battle Against Scams and Payments Fraud' (PYMNTS, 2 February 2025) https://www.pymnts.com/fraud-prevention/2025/behavioral-analytics-inject-smart-friction-into-battle-against-scams-and-payments-fraud accessed 28 April 2025; R. W. Lee et al, 'The Relationship Between Cognitive Abilities and Fraud Detection in Older Adults' (2024) 42(1) *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition* 72 https://pmc.ncbi.nlm.nih.gov/articles/PMC10738753 accessed 28 April 2025.

181. Instagram, 'Continuing to Make Instagram Safer for the Youngest Members of Our Community' (Instagram Blog, 16 March 2021) https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community accessed 28 April 2025.

182. UK Information Commissioner's Office, *Age Appropriate Design Code* (2021) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/ accessed 28 April 2025.

183. See Article 28 and Recital 71, DSA: "Providers of online platforms used by minors should take appropriate and proportionate measures to protect minors, for example by designing their online interfaces or parts thereof with the highest level of privacy, safety and security for minors by default where appropriate or adopting standards for protection of minors, or participating in codes of conduct for protecting minors".

184. European Commission, *EU Strategy for a More Effective Fight Against Child Sexual Abuse* COM(2020) 607 final.

185. European Union Agency for Fundamental Rights, *Artificial Intelligence: Implications for Fundamental Rights* (FRA 2020) 9–11 https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights accessed 8 May 2025; D. Kaye, *Speech Police: The Global Struggle to Govern the Internet* (Columbia Global Reports 2019); D. K. Citron & F. Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89 *Washington Law Review* 1; The Law Society, Algorithms in the Criminal Justice System (The Law Society 2019) https://prdsitecore93.azureedge.net/-/media/files/topics/research/algorithms-in-criminal-justice-system-report-2019.pdf?rev=c4f153dd6d9e4d528e10f9abe1ed3e55&hash=60D4B31873F801C579C2E1A04D43F1EE accessed 8 May 2025.

making into law enforcement infrastructure, with profound implications for privacy, freedom of expression, and legal accountability.

6.2 Network Communitarian Responses to Grooming

Grooming is fundamentally a behavioural process rather than a static, content-based offence, making it resistant to algorithmic detection.[186] Unlike the identification of previously verified CSAM through hash-matching, the detection of grooming requires an understanding of progressive, context-dependent patterns of exploitation.[187] Relying on AI-driven surveillance to identify grooming behaviours risks overreaching and underreaching: false positives may misidentify benign interactions as predatory. At the same time, sophisticated offenders can circumvent detection by adapting their strategies to automated enforcement. Therefore, a more effective regulatory model must move beyond techno-solutionist approaches, instead embedding network communitarian principles that integrate user empowerment, real-time intervention mechanisms, and platform accountability.

A network communitarian response to grooming emphasises harm reduction over AI-driven moderation, recognising that collective responsibility fosters better safety than centralised algorithmic policing.[188] Digital literacy initiatives must equip young users, parents, and educators with the knowledge to identify risks and threats.[189] Parental education programs should focus on risk awareness and open, non-punitive communication strategies that encourage children to report suspicious interactions without fear of reprimand. At the platform level, friction-based design interventions—such as nudges, delayed messaging features, and consent-driven interactions—can introduce subtle yet effective barriers that disrupt predatory engagement.[190] These interventions operate pre-emptively, reducing the likelihood of grooming attempts without infringing on legitimate user interactions.[191]

Enforcement strategies in the online child protection ecosystem must prioritise intelligence-led, multi-stakeholder collaboration rather than rely on indiscriminate, dataset-driven methodologies that risk conflating predictive analysis with effective intervention. Recent attempts to deploy risk-scoring algorithms and behavioural profiling have demonstrated considerable limitations: such systems often fail to incorporate the contextual nuance necessary to distinguish between credible threats and innocuous behaviour, leading to Type I (false-positive) and Type II (false-negative)errors at an unacceptable scale. As the FRA notes, automated risk-scoring and behavioural profiling tools "can misclassify individuals, leading to false positives where innocent individuals are wrongly identified as potential threats, and false negatives where genuine threats go undetected," thereby raising "serious concerns for the presumption of innocence, proportionality, and non-discrimination."[192]

A proportionate enforcement model demands verified behavioural analysis and structured intelligence-sharing frameworks among platforms, public authorities, and civil society organisations. Community moderation initiatives embedded within a formal intelligence network have demonstrated greater efficacy

186. National Child Advocacy and Investigation Association, 'Addressing the Unique Challenges of Online Child Abuse Investigations: Leveraging Technology and AI' (NCACIA, 2023) https://www.ncacia.org/post/addressing-the-unique-challenges-of-online-child-abuse-investigations-leveraging-technology-and-ai accessed 10 March 2025.

187. J. Street, I.K. Ihianle, F. Olajide, & A. Lotfi, 'Enhanced Online Grooming Detection Employing Context Determination and Message-Level Analysis' (2024) arXiv https://arxiv.org/pdf/2409.07958 accessed 10 March 2025.

188. E. Calvete, I. Orue & M. Gámez-Guadi, 'A Preventive Intervention to Reduce Risk of Online Grooming Among Adolescents' (2022) 31 *Psychosocial Intervention* 177 https://pmc.ncbi.nlm.nih.gov/articles/PMC10268540/ accessed 10 March 2025.

189. For example, Google and Parent Zone, *Be Internet Legends: Impact Report 2023* (2023) https://beinternetlegends.withgoogle.com/en_uk accessed 28 April 2025.

190. L. Jahn, et al, 'Friction Interventions to Curb the Spread of Misinformation on Social Media' (2023) arXiv https://arxiv.org/abs/2307.11498 accessed 10 March 2025.

191. N. Meurens, E. Notté, A. Wanat, & L. Mariano, Child Safety by Design That Works Against Online Sexual Exploitation of Children (Down to Zero Alliance, 2022) 71 https://www.datocms-assets.com/22233/1652864615-child-safety-by-design-report-final-1.pdf accessed 10 March 2025.

192. European Union Agency for Fundamental Rights, Getting the Future Right – Artificial Intelligence and Fundamental Rights (FRA 2020) 30–33 https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights accessed 8 May 2025; See also K. Yeung, 'Algorithmic Regulation: A Critical Interrogation' (2018) 12 *Regulation & Governance* 505, 513–515.

than automated detection systems. For instance, the Australian eSafety Commissioner's "Safety by Design" framework has pioneered an ecosystem-wide model that integrates proactive user empowerment, platform accountability, and government facilitation.[193] Similarly, Discord's Trust and Safety team has operationalised network communitarianism by combining automated detection with manual review by trained community moderators, achieving significant reductions in grooming incidents.[194] Research indicates that community cohesiveness and attachment factors significantly influence active participation within virtual communities, including posting, commenting, and sharing.[195] Significantly, passive participation (such as simple visitation) correlates only with attachment, not cohesiveness, suggesting that embedding users within a cohesive community structure bolsters active safeguarding behaviours.[196]

Equipping users, particularly minors and guardians, with risk-mitigation skills constitutes an essential pillar of a robust prevention strategy.[197] Structured, accessible reporting tools must supplement these initiatives, ensuring that human oversight, not algorithmic automation, remains the primary gatekeeper for critical interventions. However, reporting systems must strike a balance between accessibility and the risk of over-reporting, which can dilute the investigatory focus. Embedding participatory safeguards within platforms, informed by the principles of community attachment and cohesiveness, can strengthen the resilience of reporting ecosystems without overwhelming enforcement capacities.

Structured collaboration between platforms and law enforcement enhances the precision and effectiveness of intervention efforts. Automated classifiers should not be the primary mechanism for identifying high-risk interactions, as their probabilistic nature introduces significant risks of misclassification. Platforms should establish expert moderation teams trained in behavioural risk assessment to evaluate flagged reports, ensuring that law enforcement engagement focuses on identified threats rather than algorithmically generated reports. Intelligence-sharing between digital platforms and investigative agencies enables a more targeted response to emerging grooming tactics while preserving the integrity of due process protections. Any deployment of intrusive investigative measures must remain subject to judicial oversight to prevent misuse, ensuring that child protection efforts do not become a pretext for unchecked surveillance. A regulatory model that fosters structured cooperation between private platforms and public authorities mitigates the risks of indiscriminate enforcement while reinforcing the precision and accountability of intervention strategies.

This approach fundamentally diverges from the regulatory trajectory outlined in the EU's CSAM Proposal and the UK's OSA, which prioritises automated detection and proactive content moderation as enforcement mechanisms. While embedding safety measures within digital infrastructure has clear advantages, its indiscriminate application risks establishing an enforcement model that prioritises algorithmic control at the expense of legal proportionality, institutional oversight, and user agency.

Over-reliance on code-based enforcement mechanisms generates three critical risks. First, it embeds architectural interventions such as hash-matching databases, content classifiers, and encrypted messaging backdoors with insufficient procedural safeguards, creating a regulatory model that lacks transparency and due process protections. Second, it assumes that algorithmic governance is equivalent to regulatory effectiveness, despite extensive evidence that AI-based content moderation systems struggle with bias, misclassification, and the inability to assess the contextual complexities of grooming interactions accurately. Third, it shifts adjudicative power from legal institutions to private platforms, enabling them to function as quasi-judicial entities that determine the boundaries of permissible online behaviour with limited external oversight. This fosters a regulatory imbalance in which platforms, rather than courts or legislatures, dictate

193. Australian eSafety Commissioner, *Safety by Design: Principles and Implementation Guidance* (2021) https://www.esafety.gov.au/safetybydesign accessed 28 April 2025.

194. Discord Trust and Safety Team, *Transparency Report H2 2023* (2024) https://discord.com/safety-transparency accessed 28 April 2025.

195. J. Seering, 'Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation' (2020) 4 Proc ACM Hum-Comput Interact CSCW2, Article 107 https://doi.org/10.1145/3415178 accessed 28 April 2025.

196. *Ibid.*

197. Google and Parent Zone, above n.189.

enforcement priorities, raising serious concerns about accountability, proportionality, and the protection of fundamental rights.

A more nuanced regulatory approach necessitates shifting from deterministic architectural interventions to a governance framework incorporating multiple regulatory modalities. Reconsidering the role of platforms as networked communities with shared responsibility for safety aligns with Murray's theory of network communitarianism.[198] It also demands a reassessment of intermediary liability frameworks, incorporating gatekeeper theory to ensure platform responsibilities are calibrated based on risk exposure rather than imposed as blanket obligations. Risk-based regulation provides an alternative to mass surveillance mandates by tailoring regulatory interventions to platform-specific risk profiles, ensuring that enforcement remains proportionate and targeted. Engaging with cyber-regulatory theory enables a critical evaluation of the strengths and limitations of prevailing enforcement mechanisms while creating pathways for a more effective and rights-preserving digital child protection model.

6.3 Risk-Based Regulation for Unknown CSAM

Applying machine-learning classifiers to detect newly created or AI-generated CSAM may lead to a regulatory error by reallocating enforcement resources inefficiently and compromising procedural safeguards due to a high volume of false positives. Hash-based detection, with its high specificity and minimal risk of collateral harm, contrasts with AI-driven classifiers that use probabilistic modelling, introducing uncertainty into enforcement mechanisms. These systems may misclassify lawful content, overreport non-criminal imagery, and divert investigative resources to algorithmically generated alerts rather than focusing on high-risk cases. A risk-based regulatory approach ensures detection obligations are proportionate, evidence-driven, and aligned with fundamental rights protections. Regulatory interventions should be tiered and contextual, with enforcement measures tailored to platform functionality, data flows, and user behaviour. Higher-risk services—platforms with open discoverability features, end-to-end encrypted messaging, or user-generated content functionalities—may require targeted detection orders. Conversely, low-risk platforms, such as closed enterprise systems or private file storage services, should be exempt from indiscriminate scanning mandates. The absence of a risk-based framework could erode privacy rights, chill lawful expression, and entrench disproportionate surveillance mechanisms across digital ecosystems.

Detection orders should be structured and proportionate, with robust procedural safeguards in place to prevent regulatory overreach. Judicial oversight and independent reviews must ensure that enforcement measures address actual risks rather than speculative harm. Orders should be narrowly scoped, requiring platforms to prove technical feasibility, effectiveness, and rights compliance before deploying automated detection tools. Transparent accountability structures—such as reporting obligations, audits, and redress mechanisms for wrongful flagging—must uphold due process protections. Current strategies risk embedding flawed AI-driven enforcement in child protection regimes without careful recalibration, undermining CSAM detection and regulatory legitimacy. An evidence-based policy framework prioritising risk proportionality safeguards against overreach, and enforceability grounded in legal certainty is essential for balancing child protection with fundamental rights.

7. A Comprehensive Regulatory Design for CSAM Regulation

Tackling CSAM demands a sophisticated, multi-layered regulatory architecture grounded in evidence-based interventions, proportionality, and robust institutional oversight. Existing legislative frameworks, such as the United Kingdom's Online Safety Act 2023 and the European Commission's CSAM Proposal, impose extensive obligations, often anchored in AI-driven surveillance systems. Critically, these frameworks usually fail to distinguish between qualitatively distinct forms of harm, including known CSAM, unknown CSAM, and grooming solicitation. Treating these disparate challenges as a monolithic category undermines regulatory

198. For a detailed discussion on network communitarianism, see Chapter 6, 'Networks and Nodes', in C. Reed & A. Murray, *Rethinking the Jurisprudence of Cyberspace* (Edward Elgar, 2018).

precision and risks generating disproportionate interventions that erode fundamental rights. An effective CSAM regulatory framework must be grounded in five foundational principles:

1. **Risk Calibration** — Platforms should assume obligations proportionate to their documented exposure to CSAM-related threats, determined through rigorous, externally auditable risk assessments.
2. **Technical and Procedural Safeguards** — Automation must continuously operate within a framework of human oversight, with intrusive detection measures deployed only under judicial authorisation and subject to regular external auditing.
3. **Multi-Stakeholder Governance** — Effective oversight demands institutionalised reporting, monitoring, and redress mechanisms, incorporating public authorities, civil society, and independent supervisory bodies to safeguard against regulatory capture.
4. **Transparency and Accountability**—Systematic publication of detection protocols, enforcement outcomes, error rates, and rights impacts is essential to reinforcing public trust and regulatory legitimacy.
5. **Targeted, Intelligence-Led Enforcement** — Enforcement strategies must prioritise context-sensitive, intelligence-driven interventions over indiscriminate, dataset-driven surveillance, ensuring that regulatory actions are proportionate and effective.

This foundation supports a proportionate, intelligence-led enforcement model that prioritises multi-stakeholder collaboration over indiscriminate automation, ensuring precision and respect for rights. Recent attempts to deploy risk-scoring algorithms and behavioural profiling have demonstrated considerable limitations: such systems often fail to incorporate the contextual nuance necessary to distinguish between credible threats and innocuous behaviour, leading to errors at an unacceptable scale. These predictive tools risk criminalising lawful conduct, undermining the presumption of innocence, and misallocating enforcement resources away from high-risk offenders.[199] Our proportionate enforcement model demands verified behavioural analysis and structured intelligence-sharing frameworks among platforms, public authorities, and civil society organisations.[200]

Our model's strength lies in its balanced integration of automated detection, behaviourally informed design interventions, and community-centred governance, all operating under rigorous legal oversight.[201] Empirical research and Vranaki's work highlight that regulatory strategy must be tailored to each platform's unique socio-technical ecosystem.[202] Efforts to regulate or safeguard users must avoid top-down, one-size-fits-all prescriptions and instead consider the nuanced assemblage of actors, affordances, and resistance practices unique to each environment.[203] Proactive design interventions should be framed as complements to, rather than substitutes for, sustained human oversight, accountability structures, and legislative frameworks aligned with international human rights standards.[204] An effective regulatory model must distinguish between these phenomena and structure obligations accordingly.

199. V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018).
200. European Commission, *EU Strategy for a More Effective Fight Against Child Sexual Abuse* COM(2020) 607 final.
201. R. Thaler & C. Sunstein, *Sludge: What Stops Us from Getting Things Done and What to Do about It* (MIT Press 2022).
202. T. Heimburg & M. Wiesche, 'Time to Break Up? The Case for Tailor-Made Digital Platform Regulation Based on Platform-Governance Standard Types' (2024) *Electronic Markets* https://link.springer.com/article/10.1007/s12525-024-00747-7 accessed 28 April 2025; Bucher H and others, 'Platforms' Regulatory Disruptiveness and Local Regulatory Outcomes: A Comparative Analysis' (2024) *Internet Policy Review* https://policyreview.info/articles/analysis/platforms-regulatory-disruptiveness accessed 28 April 2025; Eckardt M, 'EU Digital Law and the Digital Platform Economy—An Inquiry into the Co-Evolution of Law and Technology' (2024) *European Journal of Law and Economics* https://link.springer.com/article/10.1007/s43253-024-00135-z accessed 28 April 2025; Asma A I Vranaki, 'Social Networking Site Regulation: Facebook, Online Behavioral Advertising, Power and Data Protection Laws' (2017) 43(2) Rutgers Computer and Technology Law Journal 169 https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3004651_code1151450.pdf?abstractid=2731159&mirid=1 accessed 28 April 2025.
203. Ibid.
204. UN Committee on the Rights of the Child, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment* CRC/C/GC/25; See also Sonia Livingstone, 'Children's Rights in the Digital Age: A Download from Children Around the World' (UNICEF Office of Research, 2014); T C Li, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2020) 98 *Fordham Law Review* 439.

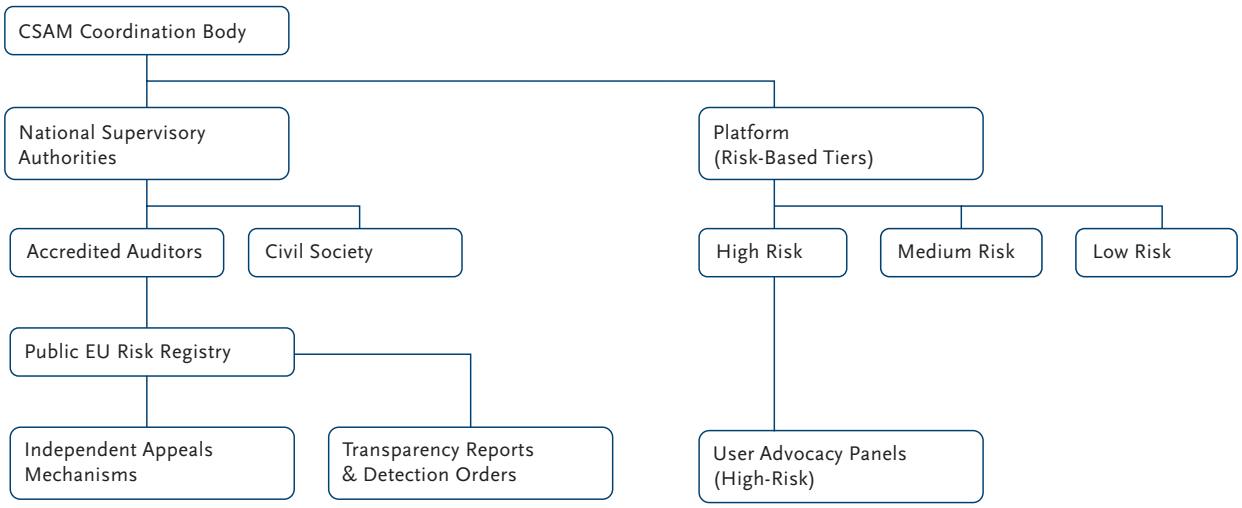
Building on insights from code-based regulation, network communitarian theory, gatekeeper theory, and dynamic risk regulation[205], our proposed model integrates technological enforcement with calibrated legal oversight. Central to this design is a risk-based structure that imposes obligations proportionate to each platform's exposure to CSAM-related threats. Platforms must therefore undertake formal, externally auditable risk assessments, enabling regulatory obligations to be tailored based on documented exposure profiles.[206] Automation must operate under human oversight, with surveillance subject to judicial authorisation and audits.[207] Multi-stakeholder oversight—via supervisory authorities, auditors, and civil society—is vital to prevent regulatory capture.[208] Transparency, accountability, and auditability must be embedded as structural principles, requiring the systematic publication of annual transparency reports detailing detection protocols, enforcement actions, error rates, and the impact on user rights.[209]

Platform Risk Profile	Obligations	Safeguards
High-risk (e.g., encrypted messaging, P2P)	Targeted detection orders, human-in-the-loop, judicial authorisation, mandatory audits, transparency reports	Strict external audits, annual rights impact reports
Medium-risk (e.g., mainstream social media)	Behavioural monitoring (e.g., unsolicited adult-to-minor communication alerts), privacy-by-default settings	Minimise surveillance scope, judicial oversight where needed
Low-risk (e.g., niche apps, low-traffic services)	No mandatory scanning unless a credible risk is documented	Risk-based reassessment only on credible threats

This co-regulatory framework requires independent national supervisory authorities to oversee compliance, authorise detection orders, and audit technologies, extending the Digital Services Act's assessor mechanism.[210].[211] Platforms operating above specific risk thresholds must also be required to establish user advocacy panels, independent of corporate control, that adjudicate disputes relating to erroneous CSAM takedown decisions, modelled on the Facebook Oversight Board.[212] A public risk registry should catalogue platforms subject to detection orders, detailing the risk justification, procedural safeguards, and compliance requirements imposed upon them.[213]

Within this model, detection orders must be graduated according to a platform's risk profile. Encrypted messaging services and peer-to-peer file-sharing networks should be subject to targeted detection obligations, incorporating human-in-the-loop processes, mandatory independent audits, and transparency requirements, contingent on judicial authorisation.[214] Meanwhile, mainstream social media services would be required to deploy behavioural monitoring triggers, such as alerts following unsolicited adult-to-minor communication, while maintaining privacy-by-default settings to minimise rights intrusions. Low-risk services, by contrast, would remain free from mandatory content scanning obligations unless credible risk assessments indicate otherwise. Such a tiered regulatory design prevents disproportionate burdens on low-risk operators while concentrating enforcement where needed.

205. C. Reed, *Making Laws for Cyberspace* (OUP 2012) chs 3–5.
206. K. Yeung, A Study of Risk Assessment Frameworks in Algorithmic Governance (2020) 11 *European Journal of Risk Regulation* 145.
207. European Data Protection Supervisor (EDPS), 'Opinion 9/2022 on the Proposal for a Regulation to Combat Child Sexual Abuse' (2022).
208. European Union Agency for Fundamental Rights (FRA), *Children's Rights and Safeguards in the Digital Environment* (2021).
209. EDPS, 'Opinion 9/2022' above, n.214.
210. Digital Services Act (Regulation (EU) 2022/2065), Art.42.
211. Digital Services Act, Arts. 37–38.
212. Oversight Board, 'Annual Report 2022' (Meta Oversight Board 2023) https://oversightboard.com accessed 28 April 2025.
213. Digital Services Act, Art.45.
214. European Data Protection Board (EDPB), 'Statement on the Proposal for a Regulation to Combat Child Sexual Abuse' (2022).



Differentiation among the various manifestations of CSAM is essential to ensuring proportionate and effective regulation. Known CSAM is best addressed through existing hash-matching technologies, which offer mature and rights-respecting detection methods when deployed under strict procedural safeguards.[215] However, detecting unknown CSAM presents greater difficulty. Automated classifiers remain prone to significant error rates. Therefore, they must undergo stringent external validation, mandatory transparency on accuracy metrics, and judicial pre-approval before being deployed at scale.[216] Grooming solicitation demands a distinct approach, relying on minimal, context-specific behavioural indicators rather than expansive data surveillance, combined with staged human review processes, as exemplified by community-based moderation models.

Fundamental to the legitimacy of this regulatory structure is the existence of accessible and independent appeal mechanisms. Users whose content is removed or accounts suspended must have the right to challenge such decisions through external adjudicatory bodies, free from platform influence.[217] In line with the ethos of Article 37 DSA, annual third-party audits of platform compliance and algorithmic detection technologies should be mandated, with public summaries made available to ensure accountability without jeopardising the operational integrity of child protection efforts. At the supranational level, a CSAM Coordination Body should synthesise national reporting data and publish annual aggregate reports that detail enforcement trends, error margins, rights impacts, and systemic challenges.[218]

The cross-border nature of CSAM dissemination necessitates robust international cooperation. A centralised European mechanism, linked to Europol's European Cybercrime Centre (EC3), should coordinate cross-jurisdictional investigations governed by harmonised evidence-sharing standards that comply with the European Union's Charter of Fundamental Rights.[219] Bilateral and multilateral cooperation agreements must expressly preclude the admissibility of evidence obtained through unlawful surveillance, reinforcing the primacy of procedural safeguards and fundamental rights protections.

While the societal harm posed by CSAM warrants decisive regulatory intervention, it is imperative to avoid disproportionate measures that compromise broader civil liberties. The experience of the Internet Watch Foundation's Keyword List, which by 2020 contained over 5,000 flagged terms[220], illustrates the dangers of

215. NCMEC, 'PhotoDNA Overview' (National Center for Missing & Exploited Children, 2023) https://www.missingkids.org/photodna accessed 28 April 2025.
216. AlgorithmWatch, 'The Challenges of Classifying New CSAM' (2023) https://algorithmwatch.org/en/csam-ai accessed 28 April 2025.
217. Digital Services Act, Art.20.
218. European Commission, COM(2022) 209 final, above n.2.
219. Charter of Fundamental Rights of the European Union [2012] OJ C326/391.
220. Internet Watch Foundation, 'Keyword List' (2020) (details remain confidential for operational reasons) https://annualreport2020.iwf.org.uk/tech/keyservices/keywords accessed 12 May 2025.

well-intentioned but overly broad surveillance, which can inadvertently criminalise lawful communication and chill legitimate expression, particularly in the context of adolescent sexual development.[221] The risk is that technological responses aimed at suppressing visible traces of CSAM may conceal rather than resolve the underlying societal issues, while simultaneously infringing upon the rights to privacy and freedom of expression enshrined in Articles 7 and 11 of the Charter.[222] Proportionality demands a nuanced, risk-calibrated regulatory response that privileges human oversight, community engagement, and cross-border cooperation over mass surveillance and technosolutionist shortcuts.

Thus, the proposed regulatory architecture for CSAM dissemination is best understood as a multi-dimensional, risk-calibrated governance structure that combines differentiated platform obligations with institutionalised safeguards and oversight.[223] Platforms are stratified into high-, medium-, and low-risk tiers based on their exposure to CSAM-related threats, with each tier subject to proportionate duties. High-risk services—such as encrypted messaging platforms and peer-to-peer networks—face the most intensive obligations, including targeted detection orders, mandatory human-in-the-loop processes, independent audits, and judicial authorisation. Medium-risk platforms, encompassing mainstream social media, must implement behavioural monitoring triggers (such as alerts following unsolicited adult-to-minor communication) while adhering to privacy-by-default configurations to mitigate rights intrusions. Low-risk entities remain exempt from mandatory scanning unless credible, externally validated risk assessments justify intervention.

Technological enforcement measures must operate within a robust framework of procedural and structural safeguards. These include formal, externally auditable risk assessments; accredited third-party auditors; and independent supervisory authorities responsible for monitoring compliance, authorising detection deployments, and ensuring adherence to fundamental rights standards. Automation may only be deployed under conditions of verified human oversight, and all intrusive detection technologies must obtain prior judicial approval. To prevent regulatory capture and ensure legitimacy, oversight mechanisms must encompass multi-stakeholder input, with civil society watchdogs and public interest experts empowered to scrutinise compliance activities.[224] Detection approaches are likewise tailored to the specific manifestations of CSAM.

Operational accountability is underpinned by mandatory transparency and auditability obligations. Platforms must publish annual transparency reports that include detection protocols, enforcement metrics, false positive and negative rates, and documented impacts on user rights. Appeals mechanisms must be independent, publicly accessible, and free from platform influence, enabling users to contest erroneous takedown decisions or account suspensions. These mechanisms should be complemented by user advocacy panels, institutionally separated from platform governance structures, and modelled on bodies such as the Facebook Oversight Board.[225]

Annual third-party audits of platform compliance and algorithmic tools must be standardised, and public-facing summaries must be issued to promote accountability without compromising child protection operations. At the supranational level, a centralised CSAM Coordination Body should aggregate national enforcement data and publish harmonised reports on trends, error margins, rights implications, and systemic barriers. Cross-border cooperation remains essential. Ideally linked to Europol's EC3, for example, a European coordination mechanism should facilitate transnational investigations and evidence-sharing

221. S. Livingstone, J. Byrne & K. Third, *Children's Rights in the Digital Age: A Download from Children Around the World* (UNICEF 2014).
222. Charter of Fundamental Rights of the EU, Arts 7 & 11.
223. European Commission, 'Questions and Answers on the Digital Services Act' (European Commission, 15 December 2020) https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348 accessed 30 April 2025.
224. Internet Watch Foundation, *Keywords List* (Internet Watch Foundation, undated) https://www.iwf.org.uk/our-technology/our-services/keywords-list/ accessed 30 April 2025.
225. Oversight Board, *Improving how Meta treats people and communities worldwide* (Oversight Board, undated) https://www.oversightboard.com/ accessed 30 April 2025.

protocols that align with the Charter of Fundamental Rights.[226] Cooperation agreements must explicitly prohibit the use of unlawfully obtained surveillance evidence, reinforce procedural guarantees, and maintain the integrity of investigations. Ultimately, the legitimacy and effectiveness of this regulatory framework hinge on its ability to reconcile urgent child protection imperatives with the preservation of civil liberties. Overly broad surveillance, such as keyword blacklists or speculative profiling, risks chilling lawful speech and criminalising normal developmental behaviours, particularly among adolescents.[227] Therefore, a proportionate response must prioritise calibrated enforcement, community engagement, and transparent, human-centred review over technosolutionist expedience. This model aims to safeguard children from exploitation while upholding the democratic principles, privacy rights, and expressive freedoms that form the bedrock of a rights-respecting digital society.

226. Europol, *European Cybercrime Centre – EC3* (Europol, undated) https://www.europol.europa.eu/about-europol/european-cybercrime-centre-ec3 accessed 30 April 2025.

227. Center for Democracy & Technology, 'CDT Original Research Examines Privacy Implications of School-Issued Devices and Student Activity Monitoring Software' (Center for Democracy & Technology, 29 April 2024) https://cdt.org/insights/cdt-original-research-examines-privacy-implications-of-school-issued-devices-and-student-activity-monitoring-software/ accessed 2 May 2025; R. Ó Fathaigh, "Article 10 and the Chilling Effect Principle," (2013) 18 *European Human Rights Law Review* 304.