This is concerned with three parts of the anti-CSAM proposal: detection of "known" material, detection of "unknown" material, and detection of "grooming" or "solicitation". The existing proposal demands doing this under specific detection orders, but the comments below apply to any circumstances whatsoever in which they might be done.

"Known" material

It is clearly possible to detect an already known media file. However--

1. Any infrastructure for doing so would be easily repurposed to detect any other targeted data.

   If the EU forces providers to create the infrastructure to respond to orders to detect "known" CSAM, governments outside, and perhaps within, the EU will find it politically and technically easy to demand that the same infrastructure be used to detect politically disfavored content, content associated with disfavored groups, and content associated with inappropriately criminalized activities. One might, for example, expect states in the United States to demand searches for abortion-related material, should the United States again permit

   It is not technically possible to build a detection system, let alone a complete detection and reporting infrastructure, that can detect known CSAM, but cannot be used, possibly with trivially modifications or extensions, to detect other material.

   The machinery you propose is exactly the machinery of a police state.

2. Targeting can also be expanded surreptitiously beyond CSAM by anyone who can insert non-CSAM material into the database(s) of "known" CSAM. This certainly includes anyone authorized to expand that database (including many national or subnational law enforcement agencies, and in fact various nongovernmental organizations with varying levels of oversight). It will probably include criminal actors who learn to defeat the protections around the database(s).

3. Detection is incompatible with good cryptographic practices. Even if detection is done at the endpoint, any meaningful response to an actual detection event requires disclosing cleartext. This creates opportunities for attackers to try to induce the disclosure machinery to operate when it should not, thus circumventing all cryptographic protection. Complexity and "exceptional cases" are extremely prolific sources of security holes in real software.

4. If the system uses "approximate hashing", as does Microsoft's ContentID, it will be relatively easy for attackers to generating data which the system identifies as targeted material, but are not. This capability can be used to maliciously direct response mechanisms such as account lockouts, additional data gathering, or

   reports to law enforcement, against innocent targets.

5. It is also possible for an attacker to plant actual CSAM in a
   targeted person's device with the intention of triggering these
   mechanisms. The proposed infrastructure creates a powerful tool
   for framing a target with the false appearance of CSAM use or
   distribution. Such an attack can be devastating even if the target
   is eventually exonerated, especially if it is delivered at the
   right time.

Some, but not all, of these can be protected against with various security controls, but such controls are necessarily imperfect, and generally become prohibitively expensive well before they reach even a remotely adequate level of reliability. There can of course never be any technical protection against concern number one, the intentional misuse of the system under government authority.

In short, any proposal for detection of "known" material in otherwise private communication carries vast risks, including but by no means limited to intentional retargeting and serious degradation of cryptography.

These risks entirely outweigh any potential benefit.

In a largely unrelated technical comment, the proposal's emphasis on URLs as primary "indicators" of known content fails to address the way such content is likely to be distributed in modern messaging systems (as complete files rather than as links) and likewise fails to recognize modern detection practices (which are often based on hashes). The mistake is so glaring that it seems unlikely that the EC assigned even minimally competent staff to preparing this proposal.

"Unknown" and "grooming" material

Detection of "unknown" or "grooming" material has all the same risks as detection of "known" material, often in worse forms. It would be even easier to intentionally abuse a system that tried to detect such material than to abuse one that

However, the idea of detecting "unknown" and "grooming" material suffers from additional problems not shared with detection of "known" material.

Unintentional endangerment

The material intentionally targeted by an "unknown/grooming" detecion system would include sexual material generated by children or young adolescents, either spontaneously or in response to actual "grooming".

This means that, even if it somehow worked as intended, such a detection system would create a database of young people perhaps particularly susceptible to sexual solicitation, possibly including sexual images usable for blackmail.

Such a database would be extremely attractive to potential abusers, and would put the listed children at increased risk of further abuse... or indeed often of *initial* abuse. There would also be a large risk of widespread disclosure of their sexual images or other sexual material.

Not only would such a database be open to insider abuse, but it would be a likely target for any outside data breach

1. The database could be sold to potential abusers. The well-developed channels now used to sell stolen passwords and similar material could easily be adapted to sell such material. Once so sold, the database would probably eventually be shared among the abusers themselves, increasing its distribution.

2. The threat of disclosing the database could be used to extort payments from the organization from which it had been stolen, as in any of the other "ransomware attacks" now widely prevalent on the Internet.

Security standards among internet services are not adequate to protect such a high-value target. These services often fail to protect their own and payment password data. Security is not improving very fast, if at all, in spite of huge existing incentives and vigorous efforts.

It would be extremely foolish to demand, or even encourage, the collection of such data. Such collection would directly increase child sexual abuse as well as other crimes.

Technical impossibility

The needed technology for "unknown/grooming" detection simply does not exist, and there are no promising approaches to create it.

All existing approaches, and all future approaches so far proposed in the technical community, inevitably lead to many false positive detections. Because of the base rate effect, the false positives would *dramatically outnumber* the true positives.

Each false positive would represent both a waste of response resources, and a likely disclosure of innocuous private information, unrelated to CSAM, both to provider staff and to law enforcement or quasi-law-enforcement personnel. Many false positives will also generate investigations that directly inconvenience, embarass, or endanger innocent users.

It is almost certain that false positives would disproprtionately burden certain content and certain users. In particular, false positives would probably be concentrated in (1) content concerned nonsexually with children, (2) content concerned with adult sexuality, and (3) communication among children themselves.

Dangerous claims about software capability

Some actors, probably commercially motivated, appear to have made claims that adequate technology exists or can quickly be built.

This is simply false. Neither machine learning nor any other technology is adequate to the task. In fact, it can be difficult for actual humans to detect "grooming" material without extensive context which would usually not be available to the detection system.

The proposal itself feels the need to defend the idea of technological feasibility, by referring to an unnamed Microsoft system and claiming "88 percent accuracy" for that system. There is no explanation of whether the claimed accuracy is against false positives (type 2) or against false negatives (type 1). Conventionally, reporting a single value indicates that the system has been tuned to equalize the two error rates.

88 percent accuracy against false positives would mean 12 percent false negatives. It is obviously absurd to "detect" 12 percent of the material in any real Internet service as potential CSAM. Even the Stasi would not have been able to handle such volume. In fact, 99 percent accuracy would still be totally inadequate, and even 99.99 percent accuracy might not be enough.

It is far easier to get from zero to 88 percent accuracy, than to get from 88 percent to 95 percent. It is harder still to get from 95 percent to 99 percent, and *vastly* harder to get from 99 to 99.9. If a large technology company has put any significant effort into a problem, and has only achieved 88 percent accuracy, that is strong evidence that the problem is totally intractable in practice.

This is especially true if there is some attempt to maintain even substandard cryptographic protection for data, since doing so implies that any detection must be done using the limited local resources of the user's device. Most of the machine learning technologies proposed for this sort of application are demand very large neural networks which most user devices can't apply, and which would indeed be extremely expensive to apply to every message even on the server side.

Furthermore, in practice, because of the attacks described in the section on "Known" material, any detection method would have to be hardened against adversarial examples: non-target material intentionally constructed to trigger the detection. Machine learning that resists adversarial examples is an open research topic, and appears to be especially difficult when the adversary has a copy of the detection network... which would again necessarily be the case for any system that allowed even substandard cryptography.