Peer Response to Jasim Alzaabi

Thank you, Jasim, for your enlightening inquiry into the “Malicious Inputs to Content Filters” case. Your post clearly shows how systems designed to benefit the public can be exploited if there is no foresight in the development processes.

To avoid such ethical and legal shortcomings, one primary measure could have been using proactive moderation strategies— like the combination of automated flagging and human review. The unrestrained soliciting of user-generated feedback in sensitive environments such as libraries and the impact it has on their machine learning systems is dangerous. Without some form of feedback verification, the risk is heightened (Boyarskaya et al., 2020).

Another essential measure is the introduction of so-called bias checkers and fairness audits at all stages of system development. Raji et al. 2020 showed that these tools enable developers to identify and mitigate negative impacts on marginalized groups before system deployment, reinforcing both ACM and BCS codes.

Accountability and GDPR's proactive transparency principles alongside strengthening public trust demand unreserved enforcement through revealing algorithms utilized by Blocker Plus’s developers and offering trusted ways to appeal blocked content. It is not enough to only foster public trust; lawmakers demand fundamental principles revealed through Veale & Edwards (2018).

As your post accurately highlights, ethical foresight is not a choice to make—it is necessary for AI systems that engage with the public. This example highlights the need for AI governance to include legal structure, design constraints, and holistic safeguards for the public good.

References:

Boyarskaya, E., Mark, G. and Voida, A., 2020. Ethical Considerations in AI-Based Content Moderation. *Proceedings of the 2020 ACM Conference on Computer Supported Cooperative Work*, pp.1–15. Available at: https://doi.org/10.1145/3313831.3376399 [Accessed 18 May 2025].

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P., 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT)*, pp.33–44. https://doi.org/10.1145/3351095.3372873 [Accessed 18 May 2025].

Veale, M. and Edwards, L., 2018. Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, 34(2), pp.398–404. https://doi.org/10.1016/j.clsr.2017.12.002 [Accessed 18 May 2025].

Peer Response to Craig Norris

Thanks, Craig, for the precise and well-reasoned explanation of the Malware Disruption case. Your post sorely highlights Rogue Services' inaction as a breach of professional ethical boundaries and the resulting sociotechnical damage.

Services like Rogue must adopt a structured governance approach that features automatic malware and suspicious traffic detection and takedown systems. As pointed out by Li et al. (2019), automated systems that employ behavioural and AI analysis greatly mitigate the hosting and proliferation of malicious content.

In addition, segmentation-addition of cybersecurity compliance requirements as part of the company's internal policy manual is crucial too. Compliance frameworks like NIST's Cybersecurity Framework or ISO/IEC 27001 should be followed by host service providers for compliant legal and ethical governance (NIST, 2018). It is one thing to avoid liability, but even more important is to protect users and safeguard public trust towards digital infrastructures.

Lastly, ethical anticipatory governance should be formalized and integrated through routine corporate reporting and staff training sessions. Floridi and Cowls (2019) make the case that embedding ethics within the technical life cycle—design to deployment—aligns decisions with public interest.

Your analysis clearly portrays that the negligence from a service provider has implications that extend far beyond the provider's operational boundaries.

Every responsible digital service provider will always adhere to ethical responsibility, maintain technological surveillance, obey legal standards, and comply with regulations.

References:

Floridi, L. and Cowls, J., 2019. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). Available at: https://doi.org/10.1162/99608f92.8cd550d1 [Accessed 18 May 2025].

Li, Y., Sun, Y. and Wang, Z., 2019. Detection and prevention of malware in cloud computing environments. *Future Generation Computer Systems*, 95, pp.647–655. https://doi.org/10.1016/j.future.2018.12.050 [Accessed 18 May 2025].

NIST (National Institute of Standards and Technology), 2018. *Framework for Improving Critical Infrastructure Cybersecurity*, Version 1.1. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf [Accessed 18 May 2025].

Peer Response to Martyna Antas

Thank you, Martyna, for your thoughtful and well-referenced discussion of the Corazón case. You've rightly highlighted the company's strong ethical stance, especially in prioritising privacy, regulatory compliance, and social inclusion. It's refreshing to analyse a case that presents a *positive* example of ethical computing practice—while still acknowledging areas for improvement.

The lack of security you mention also addresses a critical point: compliance does not equal safety. While it is commendable that Corazon worked together with the researcher, greater passive risk mitigation measures could have made it so that the vulnerability was not able to be deployed in the first place. Ethical design, as Sedenberg and Hoffmann (2016) point out, "anticipatory governance" conduct scrutiny of potential exploitation weaknesses prior to systems being put into operation, especially in life-dependent systems.

Furthermore, Corazon could routinely employ external pen testers and implement IoT security frameworks like OWASP to better manage risk from sensitive health data they possess. This also supports Article 32 of the GDPR alongside ENISA (2021) that speaks of adequate technical and organizational indications of safeguards).

Finally, your argument regarding policy gaps is equally important. One does not need to breach a policy to be held responsible and liable in emerging technologies spaces. A responsible and ethical approach should have consideration for risky situations without legislation being prepared. This attitude must be adopted in other processes from all health techs in development.

Outstanding breakdown of a complicated yet significant subject.

References:

ENISA (European Union Agency for Cybersecurity), 2021. *Guidelines for Securing the Internet of Things*. Available at: https://www.enisa.europa.eu/publications/guidelines-for-securing-the-internet-of-things [Accessed 18 May 2025].

Sedenberg, E. and Hoffmann, A.L., 2016. Recovering the history of informed consent for data science and internet industry research ethics. *IEEE Security & Privacy*, 14(4), pp.73–79. Available at: https://doi.org/10.1109/MSP.2016.75 [Accessed 18 May 2025].