This was a great post on the codes of ethics and professional conduct specifically around potentially malicious inputs to content filters. You raised important questions about the ethical, professional, and legal implications of implementing centrally controlled blacklists, particularly in the context of the Children's Internet Protection Act (CIPA). I also found the question of who determines what content is acceptable and/or harmful to be very thought-provoking and relevant, especially in the current era in which specific topics are labelled on either side of the spectrum i.e. either acceptable or harmful.

The issue of professional conduct by computing professionals involved in the implementation of content filtering systems is worth exploring further. You rightly point out that Blocker Plus acted unethically and unprofessionally by failing to disclose the manipulation of their machine learning model and the collection of user data, which may have legal implications. In addition to the excellent references you provided, I would like to include the work of Floridi and Cowls (2019), who discuss propose a framework for professional ethics in the field of computing. They emphasize the importance of ethical behavior, transparency, and accountability among computing professionals to ensure the responsible development and use of technology.

It's also interesting to note that these issues are inherent to machine learning models. A study by Caliskan et al. (2017) found that machine learning models used in natural language processing tend to reflect biases present in the training data, thereby potentially perpetuating social biases. This highlights the need for constant evaluation and scrutiny of the algorithms and datasets used in content filtering systems.

In conclusion, your blog post sheds light on the ethical complexities surrounding content filtering and highlights the need for a more comprehensive and transparent approach to protect children while safeguarding ethical principles. By considering the potential biases in content determination and emphasizing professional conduct, we can strive for a more ethical and inclusive digital environment.

References:

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186. doi: 10.1126/science.aal4230

Floridi, L. and Cowls, J., 2022. A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, pp.535-545.