ACM Ethics
The Official Site of the Association for Computing Machinery's Committee on Professional Ethics

Case: Malicious Inputs to Content Filters

Using the Code: Malicious Inputs to Content Filters

The U.S. Children's Internet Protection Act (CIPA) mandates that public schools and libraries employ mechanisms to block inappropriate matter on the grounds that it is deemed harmful to minors. Blocker Plus is an automated Internet content filter designed to help these institutions comply with CIPA's requirements. To accomplish this task, Blocker Plus was designed with a centrally controlled blacklist maintained by the software maker. In addition, Blocker Plus provided a user-friendly interface that made it a popular product for home use by parents.

Due to the challenge of continually updating the blacklist, the makers of Blocker Plus have begun to explore machine learning techniques to automate the identification of inappropriate content. During the development of these changes, Blocker Plus combined input from both home and library users to aid in the classification of content. Pleased with their initial results, Blocker Plus deployed these techniques in their production system. Furthermore, Blocker Plus continued to collect input from users to refine their learned models.

During a recent review session, the development team reviewed a number of recent complaints about content being blocked inappropriately. An increasing amount of content regarding gay and lesbian marriage, vaccination, climate change, and other topics not covered by CIPA, had been added to the blacklist. Initial investigations into these incidents suggested that there were a number of activist groups that had exploited Blocker Plus's feedback mechanism to provide input that corrupted the classification model. Determining that there was no easy way to correct the model, Blocker Plus's leadership chose to disable accounts linked to the activist groups, while keeping the existing model intact in the hope that a correction could eventually be made. The legal and business risk, they determined, would be greater by switching to an outdated model that failed to block known bad content.

Analysis

Blocker Plus is a system designed to block content that has been legally designated as harmful to children. While this filtering constitutes a form of censorship, children are considered a protected vulnerable class and experience several limitations on their autonomy. To reduce the impact on adults, CIPA also mandates that these filters must be disabled on request. Given that Blocker Plus is complying with U.S. federal regulations to facilitate socially responsible uses of computers, the system is generally consistent with Principles 1.1 and 2.3

At the same time, Blocker Plus's change to integrate machine learning techniques is morally problematic. Given the complexity and risk involved, Principle 2.5 mandates that extraordinary care is required when deploying machine learning systems. Blocker Plus failed to demonstrate the necessary caution, as the design failed to protect against intentional misuse; this oversight also constitutes a violation of Principle 2.9. At the same time, the activist groups coordinated to provide malicious input to the system. Although the system was publicly accessible,

Principle 2.8 declares that this is insufficient justification for this misuse, as there is no reason to believe that such malicious input would be authorized.

The end result is that Blocker Plus's deployment of machine learning causes harm by suppressing information of legitimate public interest and safety, as well as by discriminating on the basis of sexual orientation. Thus, this change violates Principles 1.2 and 1.4. Blocker Plus's response to the problem exacerbates the moral dilemma, as well. In particular, the makers of Blocker Plus have made no attempt to provide sufficient disclosure to stakeholders and the public regarding the limitations of their system, violating Principles 1.3, 2.4, and 2.7.

Finally, Blocker Plus provides an example of a system integrated into the educational infrastructure of society. Principle 3.7 emphasizes that the developers of such systems have an added responsibility to provide good stewardship. By failing to build adequate protections into the system, as well as failing to disclose the system's limitations publicly, the makers of Blocker Plus have violated the stewardship requirements of Principle 3.7.

These cases studies are designed for educational purposes to illustrate how to apply the Code to analyze complex situations. All names, businesses, places, events, and incidents are fictitious and are not intended to refer to actual entities.
