This is the export of a building block from the TMC Threat Library. This project aims to propose mitigations for relevant threats to reusable building blocks of modern tech systems. For more information about the project visit https://threat-modeling.net/library/

Content moderation

Gains

- Split wanted / unwanted content - by TMC DACH
- Highlight good content - by TMC DACH
- Access Control - by TMC DACH 2A
- Delete content - by TMC DACH 2A
- Interaction with the users (thumbs up/down) - by TMC DACH 2A
- Avoid illegal content - by TMC-DACH 2B
- Approval process - by TMC DACH 2A
- Analytics about content - by TMC-DACH 2B
- Legally conform content - by TMC DACH 2A
- Gain trust (from the audience) - by TMC DACH 2A
- Content validation - by TMC DACH 2A
- Enforce house rules/Compliance - by TMC DACH 2A
- Data protection - by TMC DACH 2A

System Model

- Visitor - by TMC DACH 2A
- Content creators - by TMC DACH 2A
- Hoster - by TMC DACH 2A
- Provider - by TMC DACH 2A
- Moderator - by TMC DACH 2A
- Restricted content creator - by TMC DACH 2A
- Reviewer - by TMC DACH 2A
- Reporter - by TMC DACH 2A
- Bots - by TMC DACH 2A
- Robots - by TMC DACH 2A
- Analyst - by TMC DACH 2A
- Algorithm (to show pattern matched content) - by TMC DACH 2A
- Filter - by TMC DACH 2A
- Wanted content - by TMC DACH 2A
- Unwanted content - by TMC DACH 2A
- Unwanted: Malware - by TMC DACH 2A
- Unwanted: off-topic - by TMC DACH 2A

- Unwanted: violence - by TMC DACH 2A
- Unwanted: hate speech - by TMC DACH 2A
- Unwanted: not suitable for work - by TMC DACH 2A
- Unwanted: illegal content - by TMC DACH 2A

More info

- Profit/non-profit focussed - by TMC DACH 2A

Threats and Mitigation Ideas

Confidentiality? Integrity? Availability? / Usability? Privacy? Safety? €$? ...
→ Protect? Detect? Respond? → Follow-up threats?

- Unwanted content was created - by TMC DACH 2A
  - Someone can report the content - detect - by TMC DACH 2A
  - Moderator splits wanted and unwanted content - respond - by TMC DACH 2A
  - Unwanted content is deleted - respond - by TMC DACH 2A
  - Automatic detection - detect - by TMC DACH
  - Mark the content as suspicious as long as it's unchecked - by MC DACH 2A
- Impersonation - by TMC DACH 2A
- Wanted content classified unwanted - by TMC DACH 2A
  - Quality assurance on classifications - by MC DACH 2A
  - Allow creators to appeal (say the classification was wrong) - by MC DACH 2A
  - Back and forth ;-) (wrong, right, wrong, ...) - by MC DACH 2A
- Unwanted content classified wanted - by TMC DACH 2A
  - 4 eye principle - by MC DACH 2A
  - Update classification rules - by MC DACH 2A
  - When in doubt: classify unwanted - by MC DACH 2A
- Denial of Service of the moderation by flooding with unwanted content - by TMC DACH 2A
- Data theft - by TMC DACH 2A
- Trick automatic detection by asteriks - by TMC DACH
- Unwanted content is copied and replicated - by TMC DACH
- Copyright violation - by MC DACH 2A
- Bots use communication for their own communication infastructure, distracting legit users - by MC DACH 2A
- Denial of Service of the moderation by abuse of report function - by MC DACH 2A
  - Report throtteling - by MC DACH 2A
  - Don't accept reports - by MC DACH 2A

- IP-based rate limiting - by MC DACH 2A
- Captchas / proof of work - by MC DACH 2A
- Punish wrong reports - by MC DACH 2A
- Ban false reporters - by MC DACH 2A
- Deduplicate reports - by MC DACH 2A
- Require strong identities for reporters (verified identity) - by MC DACH 2A
- limit report function to validated users only - by MC DACH 2A
- Karma / social reporter score - by MC DACH 2A
- Automatically check reports - by MC DACH 2A
- Make reports easy to check with details - by MC DACH 2A
- Have second factor on reports - by MC DACH 2A

Too high reporting barriers - by MC DACH 2A

- make it easy to report - by MC DACH 2A

Hard to decide / identify if content is malicious - by MC DACH 2A

Policies changed definiting what's wanted and unwanted, content is the same - by MC DACH 2A

Inconsistent judging - by MC DACH 2A

- Detect inconsistencies in judging - by MC DACH 2A
- Allow users to give examples - by MC DACH 2A
- Have clear rules - by MC DACH 2A
- Enforce training - by TMC DACH 2A
- Review by example - by TMC DACH 2A

Content is created, judged ok, then changed to be unwanted - by MC DACH 2A

- Enforce content moderation on each update - by MC DACH 2A
- Rerun detection algorithms - by MC DACH 2A

Link leads to unwanted content - by TMC DACH 2A

Rogue/biased moderation - by TMC DACH 2A

- Audit function - by TMC DACH 2A
- Reputation score for moderators - by TMC DACH 2A
- Sanitized content - by TMC DACH 2A

Trick filter into prompt injection - by TMC DACH 2A

- Apply avoidance patterns - by TMC DACH 2A
- Avoid fancy AI - by TMC DACH

Bypass filters and reponse mechanisms by introducing multiple identities ("Sockenpuppen", Sybil attack) - by TMC DACH

- Advanced fingerprinting - by TMC DACH
- Strong identities - by TMC DACH

Malicious actor learns underlying response mechanisms by testing the report process - by TMC DACH

- Change report process all the time ;-) - by TMC DACH

Too high volume of content makes moderation infeasible - by TMC DACH

- Have enough moderators - by TMC DACH
- Scale moderators with success - by TMC DACH
- Recruit moderators from the community - by TMC DACH
- Benefits and achievements for moderators ("Denuniziant der Woche ;-)") - by TMC DACH

Some parts of contents unmoderated (example: posts are moderated, but not comments) - by TMC DACH

Moderator receive mental breakdown because they have to watch violence all the time - by TMC DACH

- Support mental health of moderators - by TMC DACH
- Limit and rotate - by TMC DACH
- Supervision - by TMC DACH