Design principles for Automoderator

1. Transparency
2. Human-on-appeal
3. Community control
4. Ease of use
5. Timeliness
6. Bounded

Transparency

We value **transparency**, visibility, and accountability so that all stakeholders can easily discover, understand, and audit Automoderator should they want to. Without transparency Admins may misunderstand, and therefore accidentally misuse or refuse to use Automoderator, and users impacted by its decisions may feel censored. These situations would lead to decreased trust in and use of the automated system.

We also want to ensure that Automoderator is not invisible but rather easily visible to the editing community. We also value transparency because it makes intentional abuse or misuse discoverable. AM's actions are visible, auditable and trackable; information about how the Revert Risk models is freely available and comprehensible for AM's users.

We will continuously study data and how the Revert Risk algorithm affects Wiki projects.

Human on appeal

We value a **human-on-appeal system** because it is guaranteed that Automoderator will make mistakes. It is therefore important for there to be systems in place that allow users to provide feedback to Automoderator and appeal its decisions. Automoderator is designed to function with human involvement in its processes. Decisions made by AM should be able to be overturned by humans where necessary.

Community control

We value **community control** because we want Admins who configure Automoderator to have and feel a sense of agency over the tool and their wiki. Automoderator exists to be in service of Admins and their work. *It* works *for* them, and not the other way around. We want the community in general to feel that they have control, not just the one or two users who understand how to configure it.

Ease of use

We value **ease of use** because we do not want Admins to get frustrated when configuring Automoderator or feel alienated by its presence on their wiki. We want to ensure that anyone that interacts with Automoderator as a system is able to engage meaningfully.

Timeliness

We value **timeliness**. Whether it's reverting vandalism or providing feedback when decisions are appealed, the tool should do these actions in a timely fashion.

Bounded

We value an intentionally **bounded** system. Automoderator does not perform actions or moderator functions beyond the limits we have set out for it (i.e. not operating on other forms of actionable behaviour, non-obvious vandalism, non-moderator tasks).