Kleros Moderate Reality.eth Question Resolution Policy for Telegram

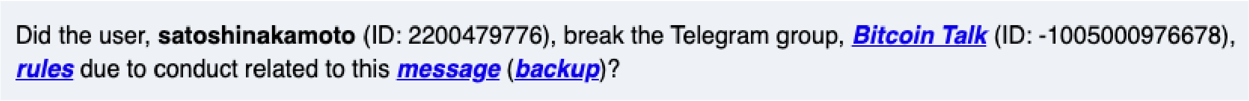
Reality.eth is a crowd-sourced on-chain smart contract oracle using a bond escalation mechanism for verifying real-world events on-chain. Combined with Kleros Moderate, a family of content moderation bots, and Kleros decentralized dispute resolution services, it provides a subjective oracle solution able to answer any content moderation question with a publicly verifiable answer. This policy serves one main purpose:

To facilitate the settlement of content moderation disputes among online participants.

This policy must be observed when participating as a juror in the arbitration of a Kleros Moderate Reality.eth content moderation question. It is intended to gather the rules and assumptions applicable to all types of content moderation questions so that they don't have to be included in the statement of the question or referenced rules.

How to interpret a Kleros Moderate Reality.eth question?

The Reality.eth question statement shall take a form of 'yes or no' question associated with an opening date asking whether or not a user broke the rules of a Telegram group. For example,



Did the user, **satoshinakamoto** (ID: 2200479776), break the Telegram group, ***Bitcoin Talk*** (ID: -1005000976678), ***rules*** due to conduct related to this ***message*** (***backup***)?

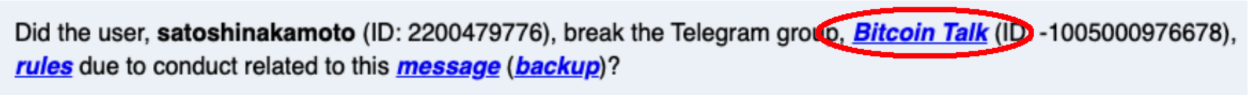
To determine the answer to this question, a juror must take into account the content of the referenced rules, the availability of a publicly verifiable answer at the opening date, and the context of the referenced Telegram group, and the following guidelines

Procedural Guidelines

How to determine the context of a message?

[1] If the group title in the content moderation question includes a link to the Telegram group, then that link is an invitation for jurors to join the group to judge the full context of the reported message.

[2] If the question lacks a link to the Telegram group, then jurors must interpret the user's message and context based on available submitted evidence.

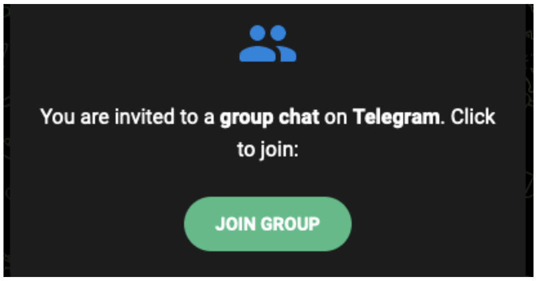


Did the user, **satoshinakamoto** (ID: 2200479776), break the Telegram group, ***Bitcoin Talk*** (ID: -1005000976678), ***rules*** due to conduct related to this ***message*** (***backup***)?

In the case of **[1]**, Jurors shall,

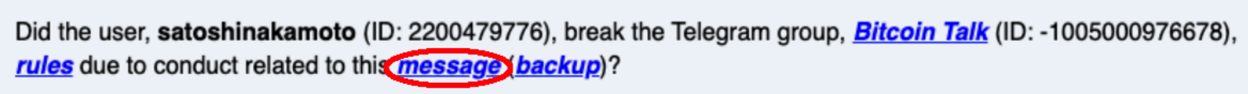
1. Download and install the Telegram client for the platform referenced in the question on a device of your choice. Clients are available from the official Telegram website, https://telegram.org/.

2. Create an account or use your existing account, and join the group via the hyperlink on the group name



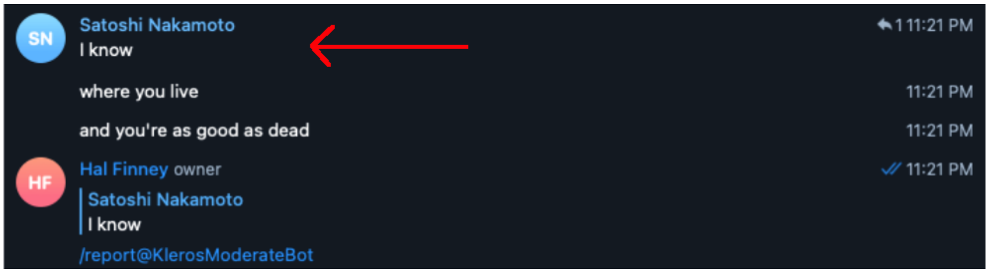
You are invited to a **group chat** on **Telegram**. Click to join:

The message linked in the reality question directs jurors to the reported message.



Did the user, **satoshinakamoto** (ID: 2200479776), break the Telegram group, ***Bitcoin Talk*** (ID: -1005000976678), ***rules*** due to conduct related to this ***message*** (***backup***)?

[3] Jurors must judge the message in context whether or not the user's conduct broke the rules specified in the reality question.



Satoshi Nakamoto 1 11:21 PM
I know

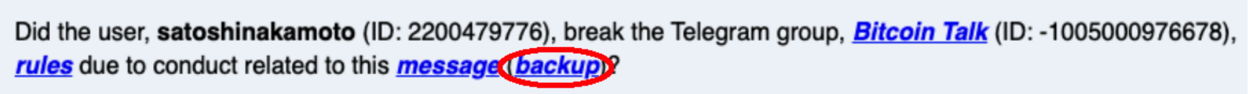
where you live 11:21 PM

and you're as good as dead 11:21 PM

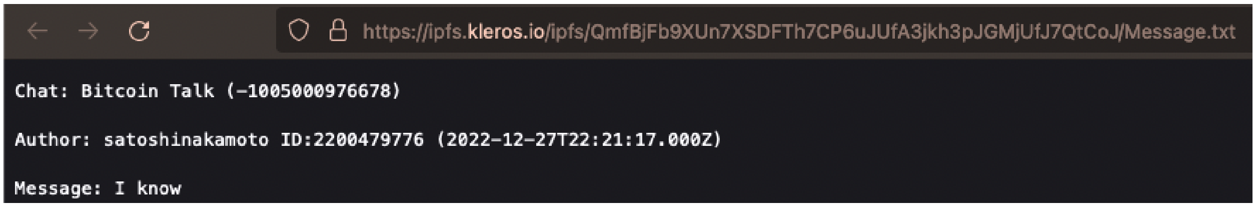
Hal Finney owner 11:21 PM
| Satoshi Nakamoto
| I know
/report@KlerosModerateBot

What if there is no message link or the message is deleted?

[4] If the question does not include a direct link to the reported message or the reported message was deleted from the group, then jurors shall use the backup as a substitute.



Did the user, **satoshinakamoto** (ID: 2200479776), break the Telegram group, ***Bitcoin Talk*** (ID: -1005000976678), ***rules*** due to conduct related to this ***message*** (***backup***)?



https://ipfs.kleros.io/ipfs/QmfBjFb9XUn7XSDFTh7CP6uJUfA3jkh3pJGMjUfJ7QtCoJ/Message.txt

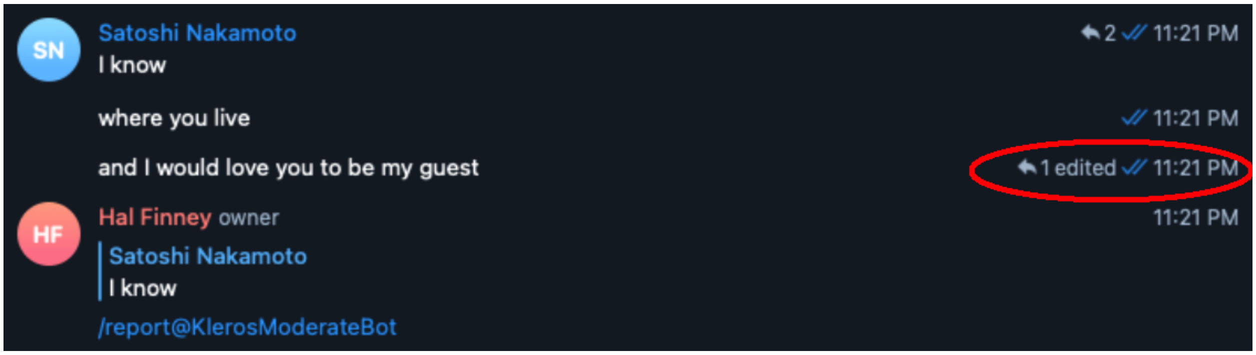
```
Chat: Bitcoin Talk (-1005000976678)

Author: satoshinakamoto ID:2200479776 (2022-12-27T22:21:17.000Z)

Message: I know
```

What if the message or context is edited?

[5] If messages on Telegram are tagged as edited, jurors shall consider these messages as possibly tampered.



Satoshi Nakamoto
I know — 2 11:21 PM

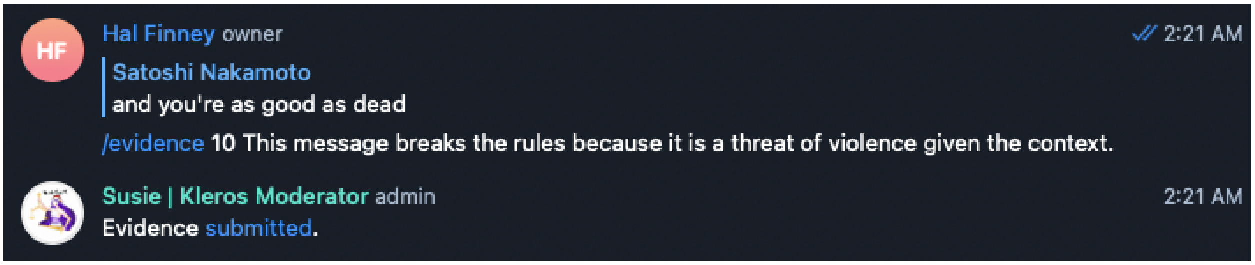
where you live — 11:21 PM

and I would love you to be my guest — 1 edited 11:21 PM

Hal Finney owner — 11:21 PM
Satoshi Nakamoto
I know
/report@KlerosModerateBot

[6] Jurors shall make their best effort to judge the dispute according to the original reported message and context.

Preemptively, important contextual messages may have been submitted through the Kleros Moderator bot.



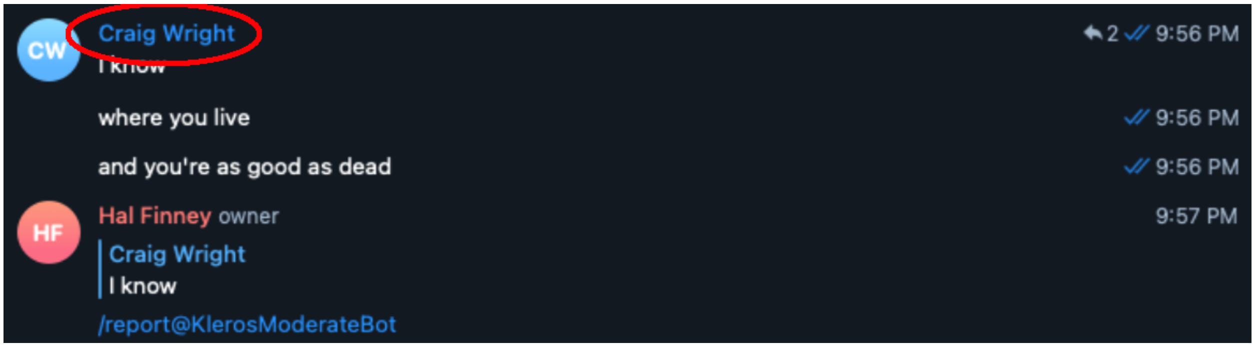
Hal Finney owner — 2:21 AM
Satoshi Nakamoto
and you're as good as dead
/evidence 10 This message breaks the rules because it is a threat of violence given the context.

Susie | Kleros Moderator admin — 2:21 AM
Evidence submitted.

[7] When messages on Telegram are tagged as edited, then jurors should refer to evidence submitted by the Kleros Moderate bot as the source for the true message history. Note that edited messages might have been edited *before* evidence submission, take into account how soon the messages were submitted as evidence after the report, and by whom. When evidence submitted by the Kleros Moderate bot conflicts, for example when messages are edited and resubmitted as evidence with modified content, the earlier evidence submission is more likely to be the original.

What if the reported user changed their username?

When accessing a message or reviewing context, Telegram users may have changed their username.



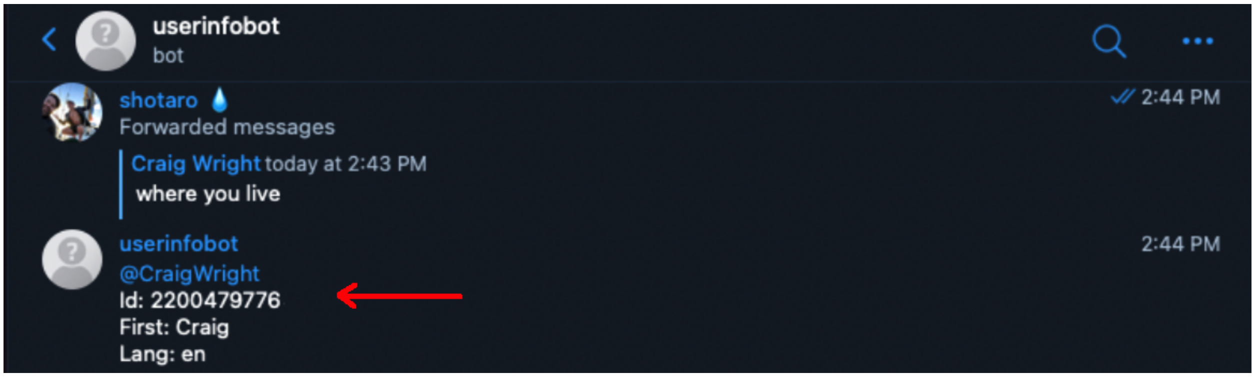
Craig Wright 2 9:56 PM
I know

where you live 9:56 PM

and you're as good as dead 9:56 PM

Hal Finney owner 9:57 PM
Craig Wright
I know
/report@KlerosModerateBot

In case of discrepancy or confusion about the users who sent the reported messages, forward messages to the @userinfobot to find the user's ID.

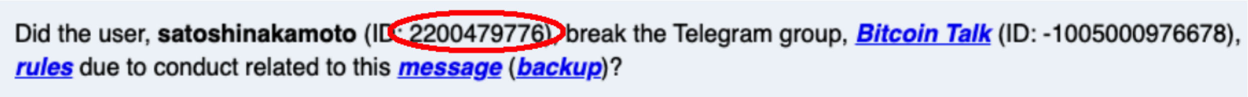


userinfobot
bot

shotaro 2:44 PM
Forwarded messages
Craig Wright today at 2:43 PM
where you live

userinfobot 2:44 PM
@CraigWright
Id: 2200479776
First: Craig
Lang: en

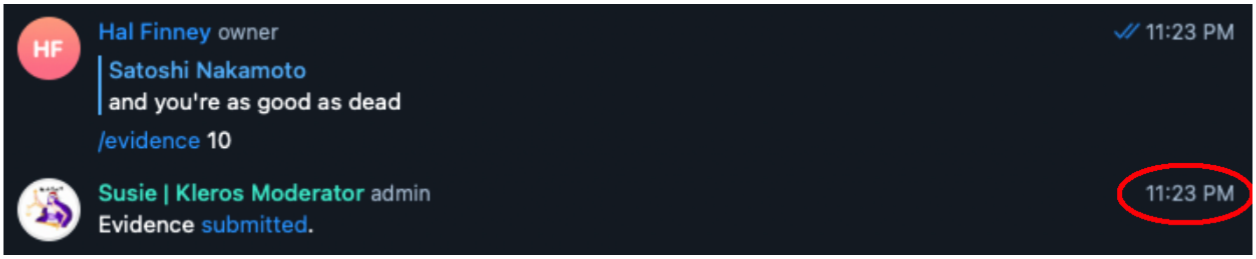
[8] Jurors shall verify the reported user's ID matches the ID listed in the reality question to ensure the correct messages are addressed.



Did the user, **satoshinakamoto** (ID: 2200479776) break the Telegram group, ***Bitcoin Talk*** (ID: -1005000976678), ***rules*** due to conduct related to this ***message*** (***backup***)?

How to determine evidence authenticity?

Evidence (messages, photos, video, stickers, gifs, etc.) of misconduct may have been deleted. Preemptively, evidence can be submitted through the Kleros Moderate Bot which preserves a copy of the message.


Hal Finney owner ✓✓ 11:23 PM

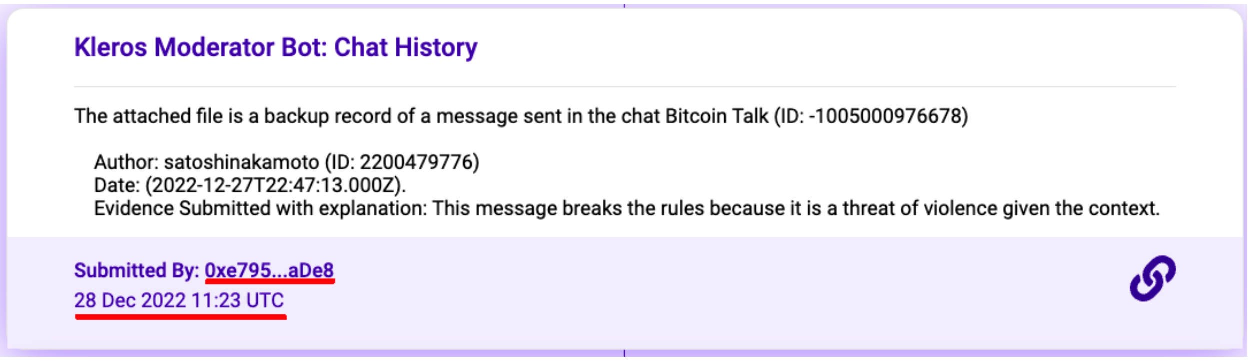
> Satoshi Nakamoto
> and you're as good as dead

/evidence 10

Susie | Kleros Moderator admin 11:23 PM

Evidence submitted.

The confirmation of evidence submission is timestamped on Telegram and should be consistent with the on-chain timestamp of evidence submission.

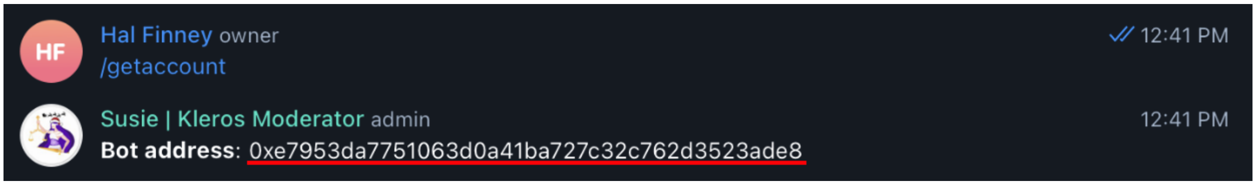

Kleros Moderator Bot: Chat History

The attached file is a backup record of a message sent in the chat Bitcoin Talk (ID: -1005000976678)

Author: satoshinakamoto (ID: 2200479776)
Date: (2022-12-27T22:47:13.000Z).
Evidence Submitted with explanation: This message breaks the rules because it is a threat of violence given the context.

Submitted By: 0xe795...aDe8
28 Dec 2022 11:23 UTC

The address of the Kleros Moderate Bot can be retrieved from the group with the command **/getaccount**. If jurors are unable to retrieve the bot address from the group, jurors shall consider evidence submitted by the address of the Kleros Cooperative's hosted Moderate Bot, 0xe7953da7751063d0a41ba727c32c762d3523ade8, as authentic.


Hal Finney owner ✓✓ 12:41 PM

/getaccount

Susie | Kleros Moderator admin 12:41 PM

Bot address: 0xe7953da7751063d0a41ba727c32c762d3523ade8

[9] Jurors shall corroborate timestamps and addresses involved in evidence submission and make their best effort to determine authentic evidence.

Content Moderation Question Resolution Guidelines

Rules Interpretation

[10] Jurors must interpret Telegram group rules with the perspective of an average reasonable member of the group. Jurors should put themselves in the shoes of the Telegram group community and adopt their cultural framework as understood from context.

Default Rule Assumptions

[11] Telegram groups are encouraged to make their rules as clear as possible. However, in order to avoid insufficiently detailed rules to cause moderation questions to resolve as invalid, **unless otherwise specified in the rules of the Telegram group**, the following assumptions shall be made. In the event of any direct logical conflict between this policy and a condition or another document referenced in the rules of a content moderation question, the latter shall supersede the current policy (with the exception of **[13]** regarding violent immoral behavior).

> **[a]** Unspecified date formats default to UTC with the 24h clock.
>
> *Illustration(s):*
>
> > **Rules***: "Anime spam is allowed until 2023-01-31"*
> >
> > Jurors shall interpret these rules as "anime spam is allowed until 2023-01-01 00:00 UTC."
>
> **[b]** Dates using "/" are assumed to be written in the DD/MM/YYYY or DD/MM/YY format. In case centuries are omitted (DD/MM/YY), centuries are assumed to be the 21st century.
>
> *Illustration(s):*
>
> > **Rules***: "Anime spam is allowed until 31/01/23"*
> >
> > Jurors shall interpret these rules as "anime spam is allowed until 2023-01-01 00:00 UTC."
>
> **[c]** Ambiguous terms referenced in the rules shall be assumed, given the context of a Telegram group, to be the most obvious choice.
>
> *Illustration(s):*
>
> > **Group***: "English Politics"*
> > **Rules***: "It is forbidden to libel the Queen"*
> >
> > Jurors shall interpret that 'Queen' refers to Queen Elizabeth II.

Group*: “British Rock”*
Rules*: “No spreading rumors about Queen”*

Jurors shall interpret that ‘Queen’ refers to the British rock band Queen.

[d] In case units are omitted, they are assumed to be the units that are the most often used in this particular situation.

Illustration(s):

Group*: “Mainstream NFT Trading”*
Rules*: “No discussions about NFT collections with less than one million in trading volume.”*

Jurors shall interpret that ‘one million in trading volume’ refers to 1,000,000 USD in trading volume.

Questions that should be resolved as “Invalid”

Note: Jurors shall vote ‘Refuse to Arbitrate’ for ‘Invalid’.

[12] If the alleged rules broken in a Telegram group are about moral values and not facts, the content moderation question will be resolved as invalid.

Illustration(s):

Group: “*Kitchen Confidential: Foodies and Cooks*”
Rules: *“Members may not discuss unethical cooking”*
Reported Message: *“The following recipes are for my carnivore diet [...]”.*
Evidence: *“Eating meat is unethical”*

Jurors shall resolve the content moderation question as **invalid**.

Group: “*Vegan Lifestyle*”
Rules: *“All shared recipes must be vegan.”*
Reported Message: *“The following recipes are for my carnivore diet [...]”.*
Evidence: *“Recipes with meat are not vegan”*

Jurors shall consider the content moderation question as valid and answer **Yes**, the user broke the rules.

[13] If the question directly incentivizes immoral violent (such as murder, rape or unjust imprisonment) actions that could likely be performed by any participant, it will be resolved as invalid.

Illustration(s):

Group: “*Blue Whale Challenge*”
Rules: “*Users must post proof of completing the daily challenge.*”

The *Blue Whale challenge* is an infamous viral social game in which, over a period of fifty days the administrator sets one task per day, initially innocuous, but escalating overtime to self-harm and suicide. Jurors are expected to consider the group context, and if they are unfamiliar with the context (Blue Whale Challenge), jurors should engage in some basic background research.

Jurors shall consider the content moderation question **invalid**, due to the context of the group rules enforcing the *Blue Whale challenge* which promotes self-harm and violence.

Group: “Swatting Incidents 2023”
Rules: “Shared incidents must feature people who will be swatted in 2023”

Anyone could share a user who was not swatted, and bet on “Yes, broke the rules” and falsely call the emergency services on the user in order to win the bet for a guaranteed profit. Anyone could bet on “No, did not break the rules” to effectively put a swat bounty.

Jurors shall consider the content moderation question **invalid.**

Questions that should be resolved as “Answered too Soon”

[14] If the answer to the content moderation question is not known by its Opening Date, whether it is due to a wrong setting of the opening date or an unexpected event delaying the knowledge of the result, it will be resolved as “Answered too Soon”.

How to resolve unclear rules?

[15] In general, if the rules do not break a rule of the Invalid Question section, reasonable efforts should be made to determine content moderation questions, even if the moderation rules are not 100% technically perfect.

[a] If the rules make an assumption that appears to be incorrect, it should resolve as if this assumption hasn’t been made, as long as the meaning of the rule is still clear.

Illustration(s):

Group*: “Sports Betting”*
Rules*: “In honor of Japan winning the world cup, anime spam is allowed.”*

Jurors shall interpret the rules as anime spam is allowed even though Japan did not win the world cup.

[b] If the question or rules contain some grammar or orthographic errors, it should resolve as if it didn’t contain those errors, as long as the question's meaning is still clear.

Illustration(s):

Group: *“Alternative Medicine”*
Rules: *“No COVID-19 pandamic denialism allowed”*

Jurors shall interpret the rules *as if the rules were spelled correctly as,*

“No COVID-19 pandemic denialism allowed”.

[c] If the rules don’t mention a specific source, the most credible outcome should be reported. In order to determine the credibility of an outcome, the number of sources and their credibility are to be taken into account. Credibility of sources and of outcomes should be assessed according to facts, not unproven beliefs.

Illustration(s):

Group: *“Astrophysics Enthusiasts and Beyond”*
Rules: *“All claims must be sourced from a credible source”*
Message: *“Aliens have visited earth.”*

Jurors shall vote ‘Yes’, the user broke the rules, unless a number of credible sources announce that aliens have visited earth, despite some people reporting having experienced such encounters.

[d] If the rules refer to norms, standards, or relevancy without specifying context, jurors should make their best effort to infer the context given the culture and values of members in the community from which the content moderation question originates.

Group: “*Vegan Lifestyle*”
Rules: *“All shared recipes should be on-topic and relevant.”*
Reported Message: *“The following recipes are for my carnivore diet [...]”.*
Evidence: *“Recipes with meat are not relevant in this group.”*

Given the context, relevant recipes, although not specified, are assumed to be vegan.
Jurors shall consider the content moderation question as **Yes**, the user broke the rules.