ACT Coding Guideline

The goal of this project is to identify conspiratory content from social media posts. The first step is to create a set of "gold-standard" samples from real posts in order to train machines capable of recognizing such content on a large scale.

Instruction:

1. Create your copy of the coding sample sheet.
2. Read the following working definitions for "*Conspiracy Theory*" (CT)
3. Based on the definition, annotate each of the posts from the coding sample.
    a. Each row represents a sampled Reddit post. Read the content of each post (title and text). You can also read the original Reddit post by clicking the post's URL. Comments on the posts are not included in the annotation consideration. The judgment should mainly be based on the title+text, and you can check out the embedded link (URL) to understand the context of the text.
        i. The attached image and embedded link only provide context, but you shouldn't rely on the image or link to decide the coding.
    b. Following the definitions below, determine your answers about the post. Note that your answer is expected to be yes/no binary. Fill in your confidence level in the decision, the value should be H/M/L to indicate high/medium/low confidence level. The final column is an open response, where you provide notes and justifications as to how/why you reached the final label decision.
    c. Fill your answers in the copied spreadsheet and submit your coding results.

Working definitions:

- **Conspiracy Theory** - *Theoretical Definition*:

"A conspiracy theory is a set of **narratives** designed to accuse an **Agent**(s) (be they individuals, groups, or organizations) of committing a specific **Action**(s), which is believed to be working towards a secretive and malevolent **Objective**(s) (secret plot)"

- **Conspiracy Theory** - *Operational Definition*:

"Following the theoretical definition, an online Conspiracy Theory is a social media post that contains a **main narrative** or claims that (a) represents a *known conspiracy theory* or (b) *suggests a secret plan*, along with (c) *evidence of agreement or support* to some extent for the mentioned conspiracy theory or secret plan"

Possible topics (please review the more comprehensive list of CTs from Wikipedia):

- Holocaust Denial / Anti-Semitic Rhetoric / Christian Identity
- White Genocide / Great Replacement/ Great Reset / White Victimization
- Anti-muslim
- Anti-immigrant
- Anti-LGBT
- Male Supremacy
- Anti-government
- 5G
- COVID

Example 1: "Electoral College. Votes don't matter. The banks own all 52 electoral voters. Your actual voting numbers do not matter. Your vote does not matter, the representative of your state could vote against you. Fraud doesn't matter one bit, the electoral voter already voted. Unless you have shit on them, it doesn't matter. Welcome back to reality. Make sure to vote for a rich guy who used slaves to become rich while your vote doesn't even matter lol. No votes matter."

Code: **CT**
Justification: The author elaborates on a scenario where a group of people, bankers, and rich individuals, control the election results and the process of democracy in the country, robbing the nation of the freedom of election
Agent: Bankers/rich
Action (plot): Control election votes in such a way that individual votes do not matter.
Objective: Accuse the agent of control democracy

Example 2: “Suppose you have it backwards? What if the aim is not to eliminate the vaccinated, with the vaccine, but to eliminate the unvaccinated, as in all the MAGA types? Maybe it's a way to get rid of all the followers of Donald J. Trump? Seems like a lot of unvaccinated getting sick right now."

Code: CT
Justification: The post explains a conspiracy theory where COVID is created to eliminate Trump supporters and MAGA members, a group of people tend to disapprove of the vaccine and mostly refuse to take it
Agent: Government
Action (plot): COVID virus as a bio-weapon
Objective: Accuse agent of using vaccine to eliminate Trump supporters

Additional Guidelines:

- **Popular and known conspiracies:** If the post mentioned existing or known CTs, such as Ukraine Biolab, Great Replacement theory [1], Space Force, 9/11, pizzagate/pedogate, Qanon, 5Ghttps://en.wikipedia.org/wiki/Misinformation_related_to_5G_technology, Communist plot to pollute Americans' "precious bodily fluids" (Dr. Strangelove), etc (https://en.wikipedia.org/wiki/List_of_conspiracy_theories), and the post authors **also express agreement** (**to some extent**) to the theories, the post is considered as a CT.
    - [1] Great Replacement theory: (*Agent*= US gov., liberals; *Action*= replace White people with immigrants; *Objective*= remove the power of White people)

Example 3: “Space Force. And now we are being separated into our homes...Does anyone think these 2 events may be related..."
Original post: www.reddit.com/r/conspiracy/comments/fk5wm3/space_force/

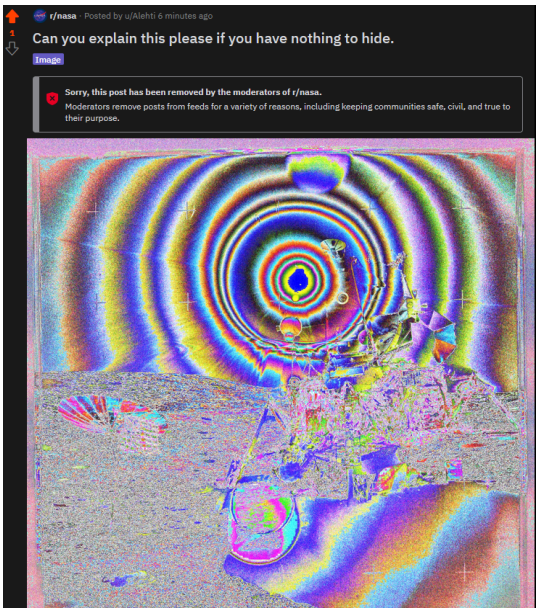
Code: CT

Justification: The post refers to a known CT called “Space Force”. The conspiracy followed a request from Trump to create a new military branch, which some people speculated would handle alien attacks. The affirmation of the first part of the post “separated into our homes” and the rhetoric question after reflect an agreement with the theory.

Example 4: “New Forensics Tool to Detect NASA Fakes https://medium.com/p/434c0a85affa"

Original post:
www.reddit.com/r/conspiracyundone/comments/utm2rn/new_forensics_tool_to_detect_nasa_fakes/

Preview Image (within post body):



Code: **Not CT**
Justification: The text may indicate a potential relation to a CT. However, the attitude of the author towards the event (the existence of a new forensic tool) is not clear. Furthermore, images and links are not used as signals to label online CT content.

- **Content sharing:** If the author presents a second-hand comment to a cited video or post, and the linked video/post may be a CT, we will NOT consider the post itself is a CT.

Example 5: *"Elon Musk Neuralink Snuff device. I am into Snuff show for Elon Musk fun -link to neuralink CT- https://ww[Neuralink Snuff device] (https://www.humorousmathematics.com/post/mk-ultra-victim-exposes-international-money-laundering-system-that-funds-terrorism-trafficking) "*

Original post: www.reddit.com/r/conspiracy_commons/comments/mwc2uf/elon_musk_neuralink_snuff_device/

Code: Not CT
Justification: The content of the link contains elements of CT. However, the post itself shares the link without commenting on its content. Since the labeling criteria does not consider links' content as a signal to indicate CT, this post is not CT.

- **Rhetorical question vs. genuine inquiry:** Reddit as a social media platform presents a ground for discourse among people who are interested in a specific topic share opinions and ask questions. CT forums are no different. Some users join these forums to gain knowledge regarding a topic from a perspective different than mainstream media. To this extent, not all posts that mention a known CT are meant to promote the mentioned CT, some are genuine inquiries. Part of your task is to distinguish between legitimate questions and rhetorical ones that ask about a CT while enclosing its elements hinting at a clear answer. Check the examples below.

Example 6: "Who's skeptical of the $1200? What are the odds that they will force you to get the vaccine? Feels like a trap"

Code: CT
Justification: The post asks for information regarding the stimulus checks in the United States (event). The second part of the question elaborates on a potential answer or a perspective for an answer to the question portraying a potential hidden agenda or secret intention behind the event. Furthermore, the last part suggests that the author believes in the

existence of such an agenda.

Example 7: “Are there any live streams from Afghanistan that are not from a news source? Like people filming right now? Can’t find anything on YouTube.”

Code: **Not CT**
Justification: The post asks for a source of information regarding the war in Afghanistan. The topic may have surrounding CTs, but the content of the post does not promote any CTs, which makes it a simple inquiry for information.

- **Support/promotion vs. criticism/frustration/debunking:** Some topics related to CTs tend to provoke strong opinions and criticism. For example, the content of mainstream media attracts many CTs while being a topic of interest to many people who express their opinions without the intention of spreading conspiracies. Presenting critical viewpoints and negative sentiments towards controversial subjects in a post does not necessarily qualify it as a CT post unless it expresses endorsement or support for a conspiracy belief.

Example 8: “Oregon has made reading, math, and writing racist which I never thought we could be racist just for breathing! We should all embrace this and bring peace and global health!”

Code: **Not CT**
Justification: The post represents an observation with criticism. Despite the controversial nature surrounding the topic, this post does not contain a hidden or malicious agenda. Furthermore, the last part of the post promotes a peaceful and positive message reducing the chance of sharing the content with harmful intentions.

- **Borderline cases:** The intention of the post’s author is a critical element in choosing the correct label. However, this is a difficult task for some posts, especially the short ones that do not provide enough context to make the decision. If the intention is not clear,

label the sample as “non-CT”. If you are not sure, feel free to choose the label “Borderline”, we can address them during consensus meetings.

- Note: After the consensus meeting with all annotators, there were only two examples in which we never reached a consensus. Here are the examples:

Example 9: “Has anyone actually watched mainstream news lately? Ss: Traumatized everyone for a year and then subject them to this repetitive mind numbing terror frequency. Holy heck. I now see how people are like fucking zombies."

Code: **Borderline**
Justification: The team split on the label decision for this sample. One group thinks this is a CT related to media control, and how it is being used as a tool to control the minds of the population turning them into zombies to achieve hidden agendas. The other group thinks this post represents criticism and frustration toward a controversial topic, mainstream media, and how it numbs people’s minds without a clear agent and hidden agenda.

Example 10: “The LEFT exists to lure in the youth and radically change our culture/politics. The RIGHT exists to pacify patriots/old people by pretending to “oppose” the left. Truly take a step back and think about it. What exactly have the conservatives *actually conserved*? I mean really, they’ve quite **literally** conserved nothing. Nothing at all. For 60 years. All they do is pacify the elderly and patriot-types until the leftist media has normalized whatever bullshit they’re trying to push. Then they move onto the next thing and the “conservatives” move on as well, pretending to be outraged again. And so on and so fourth." ”

Code: **Borderline**
Justification: The team split on the label decision for this sample. One group thinks this is part of a big conspiracy that is designed to divide people into two camps, Left and Right, to keep them busy and unaware of bigger and more important hidden agendas. The other group thinks this is a frustration regarding the current political scene and how the two major political camps have never achieved what they promised but instead are busy with minor and less important issues, without promotion of any conspiracy theories.