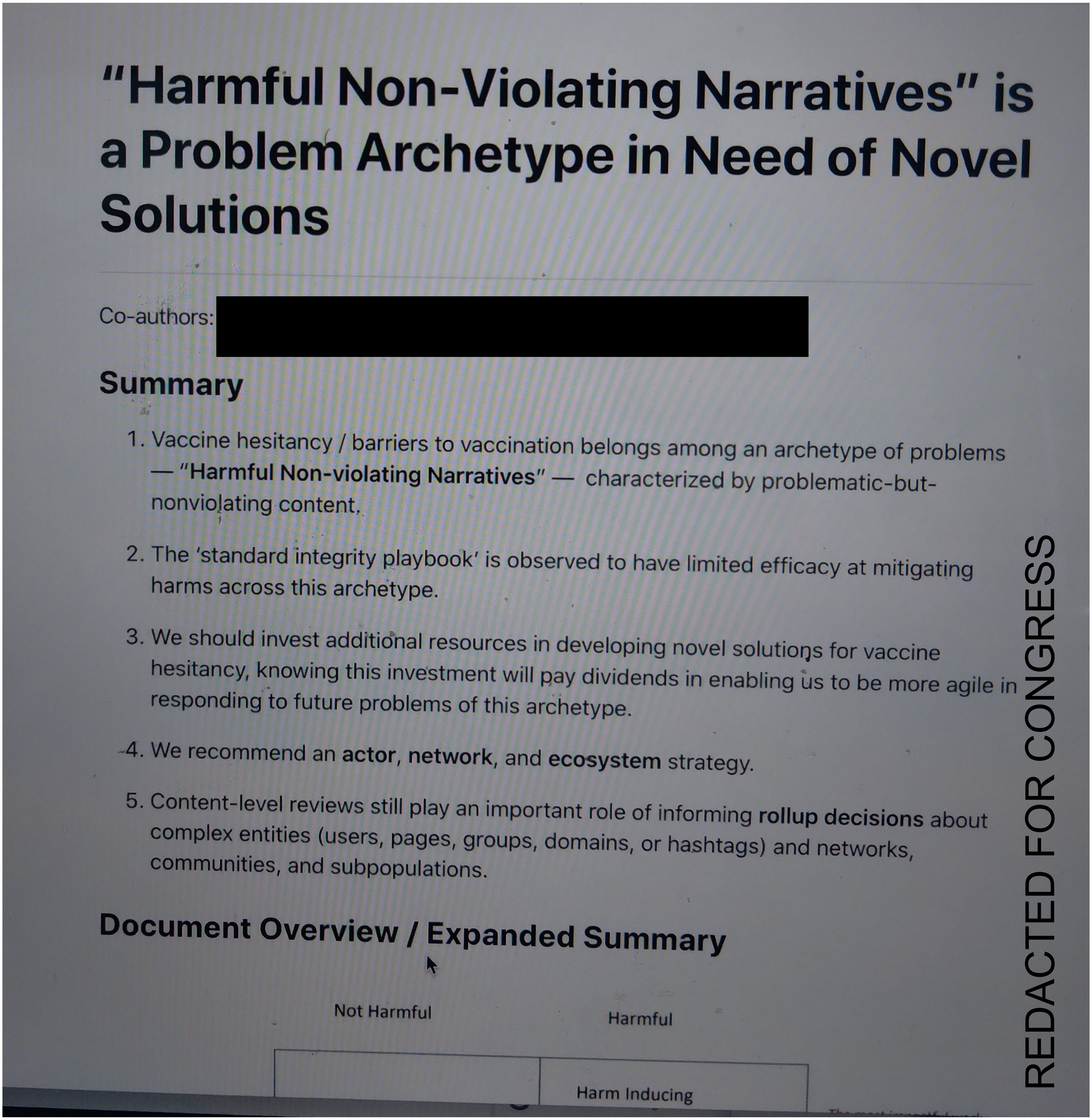


“Harmful Non-Violating Narratives” is a Problem Archetype in Need of Novel Solutions

Co-authors: [redacted]

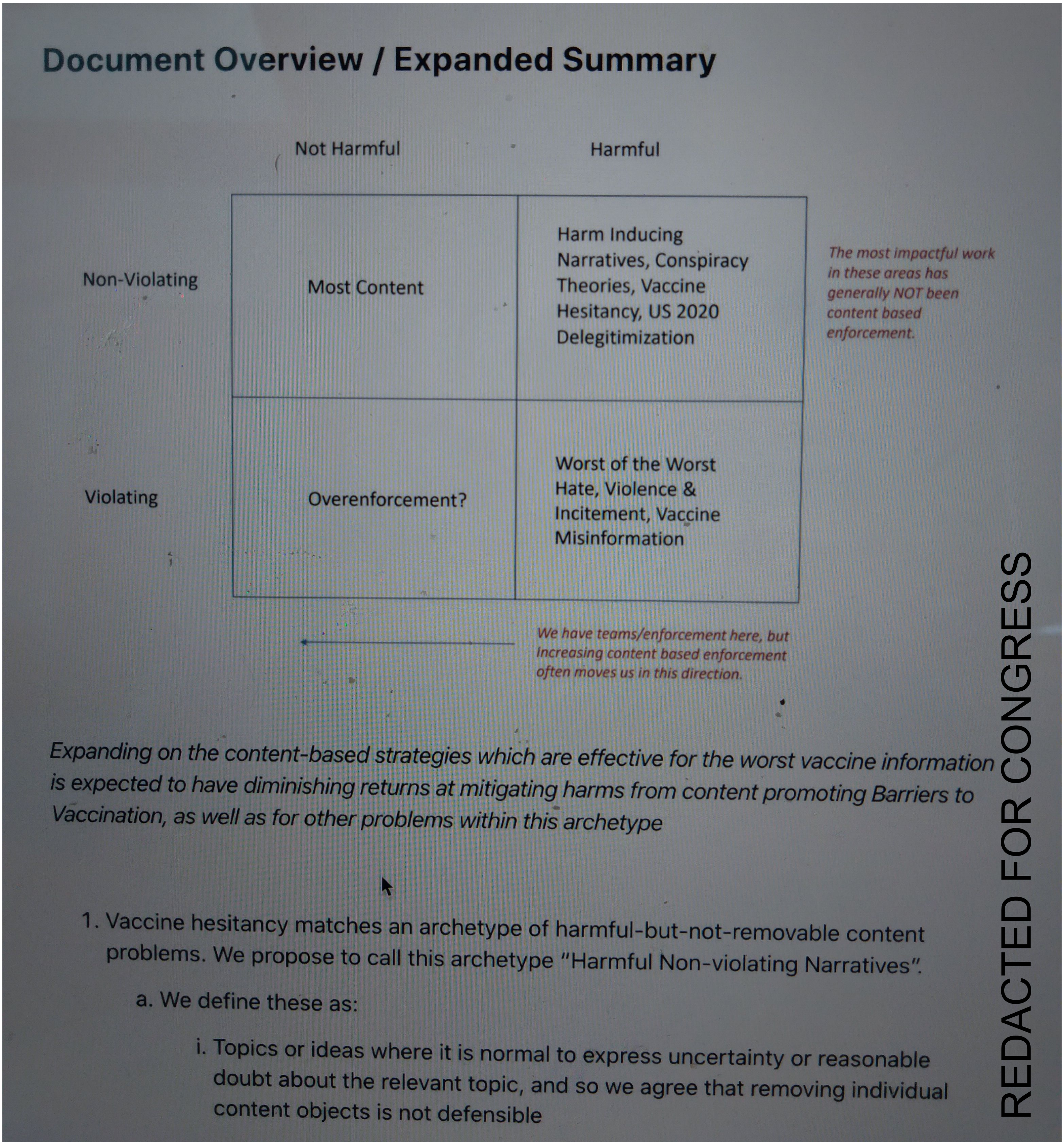
Summary

1. Vaccine hesitancy / barriers to vaccination belongs among an archetype of problems — “**Harmful Non-violating Narratives**” — characterized by problematic-but-nonviolating content.
2. The ‘standard integrity playbook’ is observed to have limited efficacy at mitigating harms across this archetype.
3. We should invest additional resources in developing novel solutions for vaccine hesitancy, knowing this investment will pay dividends in enabling us to be more agile in responding to future problems of this archetype.
4. We recommend an **actor**, **network**, and **ecosystem** strategy.
5. Content-level reviews still play an important role of informing **rollup decisions** about complex entities (users, pages, groups, domains, or hashtags) and networks, communities, and subpopulations.

Document Overview / Expanded Summary

Not Harmful	Harmful
	Harm Inducing

REDACTED FOR CONGRESS



Document Overview / Expanded Summary

	Not Harmful	Harmful
Non-Violating	Most Content	Harm Inducing Narratives, Conspiracy Theories, Vaccine Hesitancy, US 2020 Delegitimization
Violating	Overenforcement?	Worst of the Worst Hate, Violence & Incitement, Vaccine Misinformation

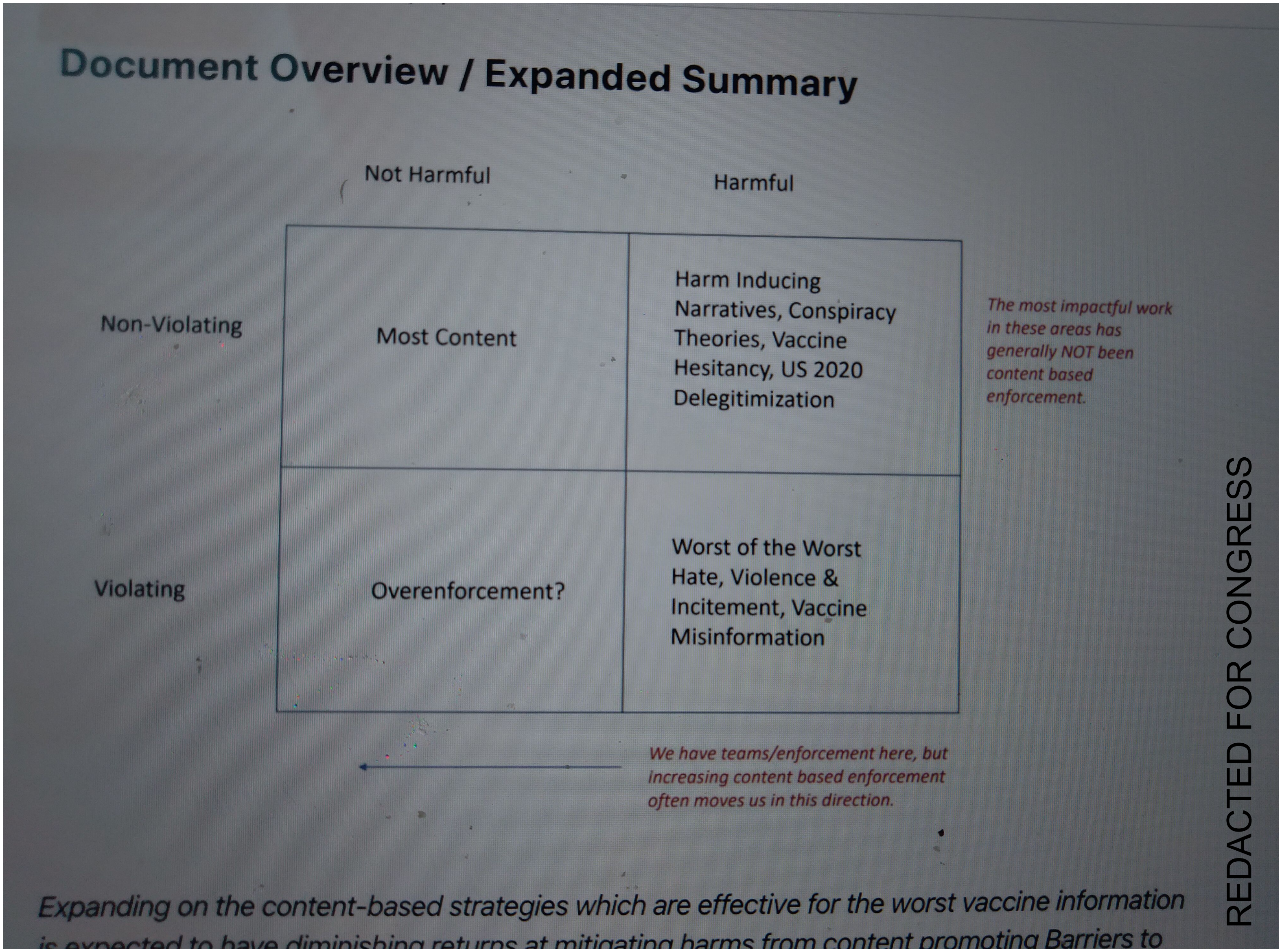
The most impactful work in these areas has generally NOT been content based enforcement.

We have teams/enforcement here, but increasing content based enforcement often moves us in this direction.

Expanding on the content-based strategies which are effective for the worst vaccine information is expected to have diminishing returns at mitigating harms from content promoting Barriers to Vaccination, as well as for other problems within this archetype

1. Vaccine hesitancy matches an archetype of harmful-but-not-removable content problems. We propose to call this archetype "Harmful Non-violating Narratives".
   a. We define these as:
      i. Topics or ideas where it is normal to express uncertainty or reasonable doubt about the relevant topic, and so we agree that removing individual content objects is not defensible

REDACTED FOR CONGRESS



Document Overview / Expanded Summary

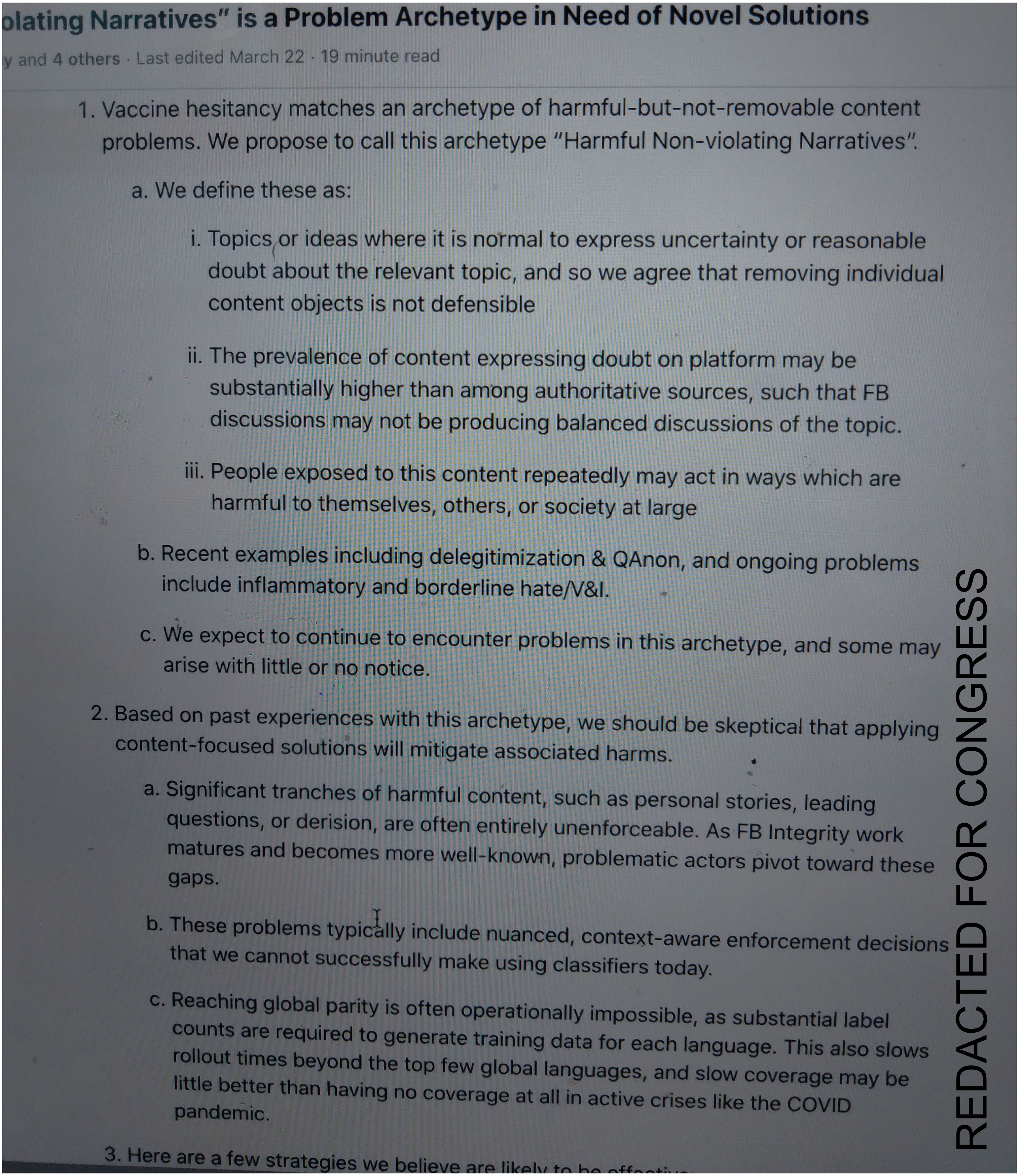
	Not Harmful	Harmful
Non-Violating	Most Content	Harm Inducing Narratives, Conspiracy Theories, Vaccine Hesitancy, US 2020 Delegitimization
Violating	Overenforcement?	Worst of the Worst Hate, Violence & Incitement, Vaccine Misinformation

The most impactful work in these areas has generally NOT been content based enforcement.

We have teams/enforcement here, but increasing content based enforcement often moves us in this direction.

Expanding on the content-based strategies which are effective for the worst vaccine information is expected to have diminishing returns at mitigating harms from content promoting Barriers to

REDACTED FOR CONGRESS

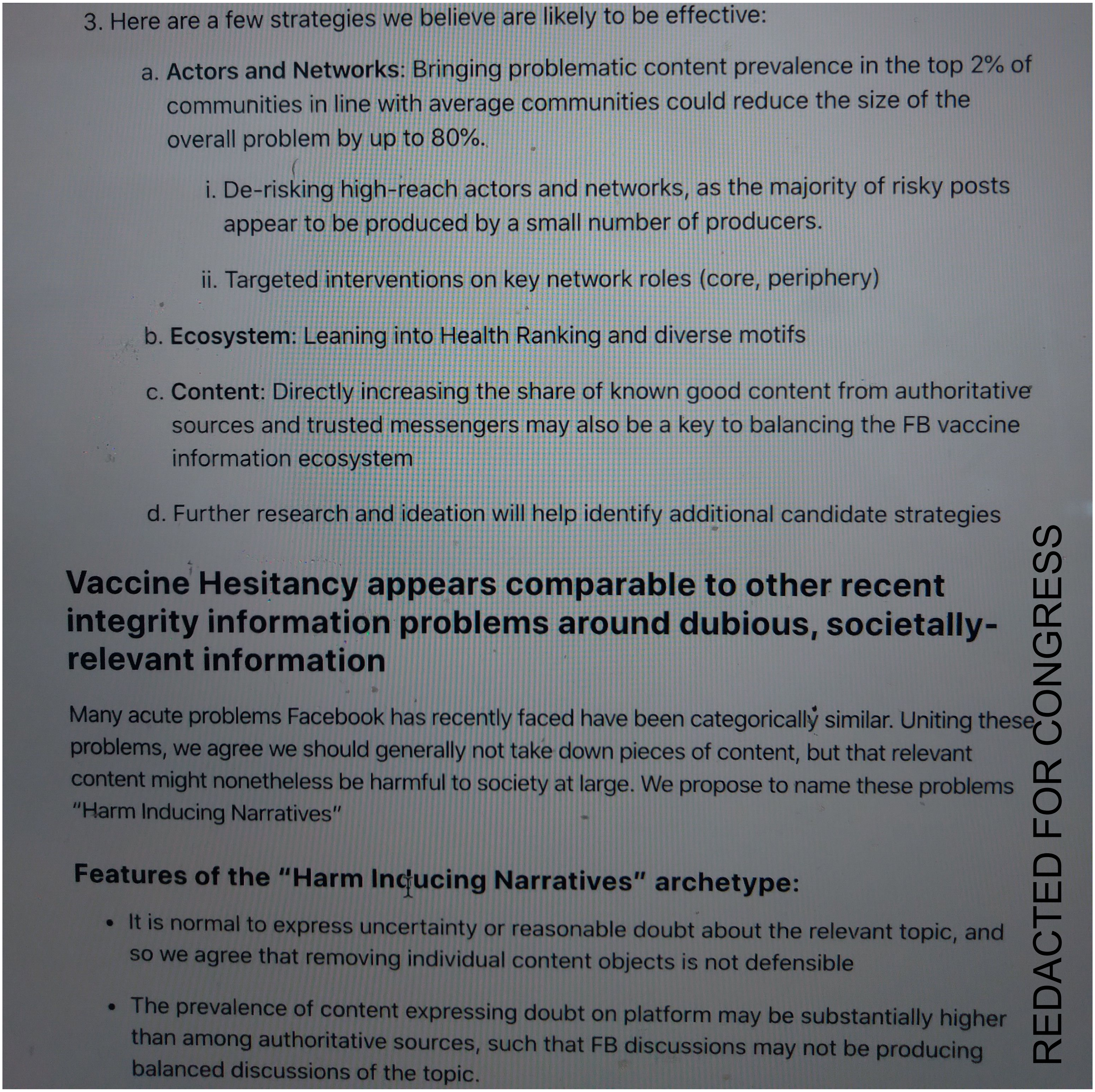


olating Narratives" is a Problem Archetype in Need of Novel Solutions

y and 4 others · Last edited March 22 · 19 minute read

1. Vaccine hesitancy matches an archetype of harmful-but-not-removable content problems. We propose to call this archetype "Harmful Non-violating Narratives".
    a. We define these as:
        i. Topics or ideas where it is normal to express uncertainty or reasonable doubt about the relevant topic, and so we agree that removing individual content objects is not defensible
        ii. The prevalence of content expressing doubt on platform may be substantially higher than among authoritative sources, such that FB discussions may not be producing balanced discussions of the topic.
        iii. People exposed to this content repeatedly may act in ways which are harmful to themselves, others, or society at large
    b. Recent examples including delegitimization & QAnon, and ongoing problems include inflammatory and borderline hate/V&I.
    c. We expect to continue to encounter problems in this archetype, and some may arise with little or no notice.
2. Based on past experiences with this archetype, we should be skeptical that applying content-focused solutions will mitigate associated harms.
    a. Significant tranches of harmful content, such as personal stories, leading questions, or derision, are often entirely unenforceable. As FB Integrity work matures and becomes more well-known, problematic actors pivot toward these gaps.
    b. These problems typically include nuanced, context-aware enforcement decisions that we cannot successfully make using classifiers today.
    c. Reaching global parity is often operationally impossible, as substantial label counts are required to generate training data for each language. This also slows rollout times beyond the top few global languages, and slow coverage may be little better than having no coverage at all in active crises like the COVID pandemic.
3. Here are a few strategies we believe are likely to be effecti

REDACTED FOR CONGRESS



3. Here are a few strategies we believe are likely to be effective:
   a. **Actors and Networks**: Bringing problematic content prevalence in the top 2% of communities in line with average communities could reduce the size of the overall problem by up to 80%.
      i. De-risking high-reach actors and networks, as the majority of risky posts appear to be produced by a small number of producers.
      ii. Targeted interventions on key network roles (core, periphery)
   b. **Ecosystem**: Leaning into Health Ranking and diverse motifs
   c. **Content**: Directly increasing the share of known good content from authoritative sources and trusted messengers may also be a key to balancing the FB vaccine information ecosystem
   d. Further research and ideation will help identify additional candidate strategies

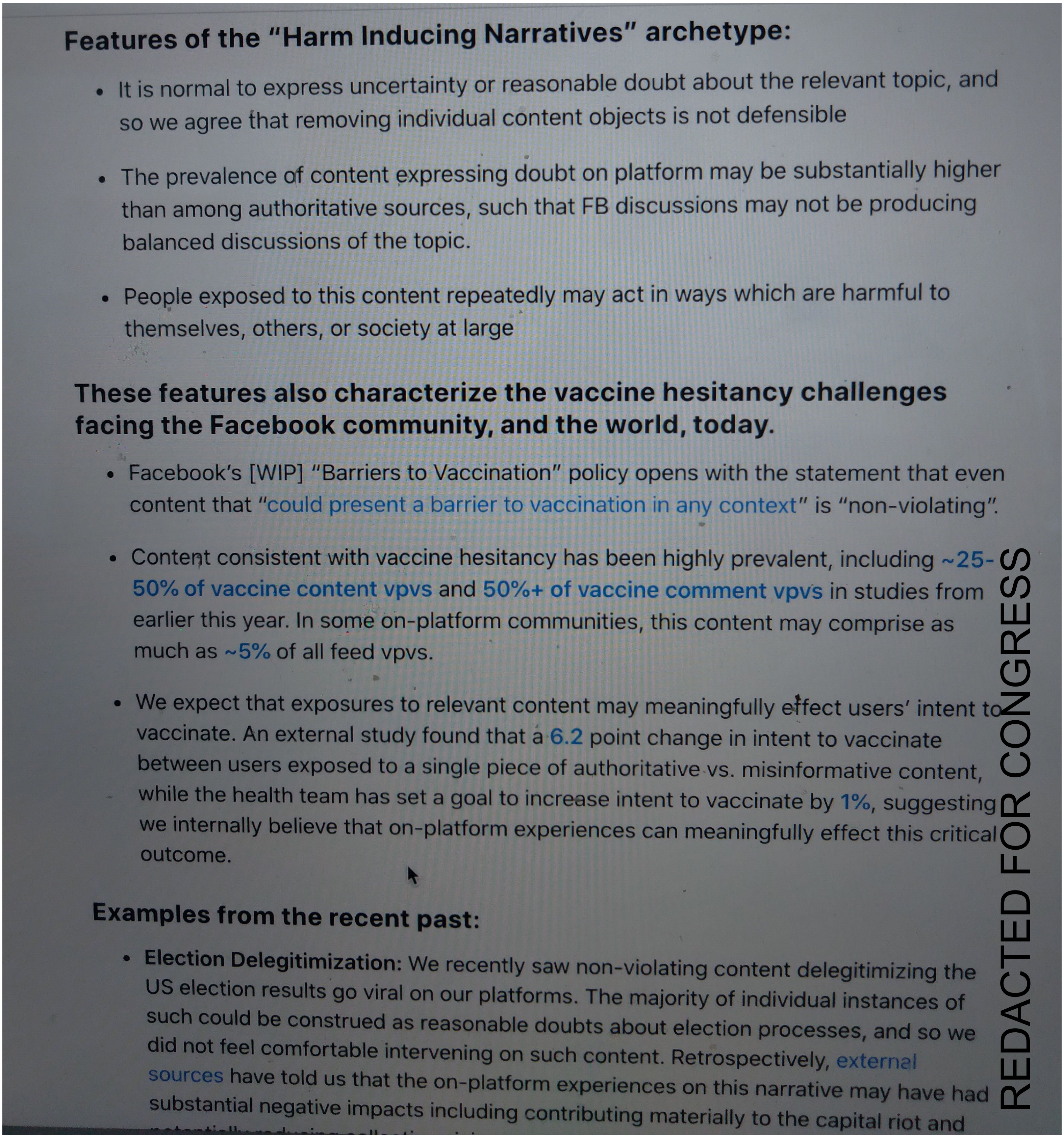
Vaccine Hesitancy appears comparable to other recent integrity information problems around dubious, societally-relevant information

Many acute problems Facebook has recently faced have been categorically similar. Uniting these problems, we agree we should generally not take down pieces of content, but that relevant content might nonetheless be harmful to society at large. We propose to name these problems "Harm Inducing Narratives"

Features of the "Harm Inducing Narratives" archetype:

- It is normal to express uncertainty or reasonable doubt about the relevant topic, and so we agree that removing individual content objects is not defensible
- The prevalence of content expressing doubt on platform may be substantially higher than among authoritative sources, such that FB discussions may not be producing balanced discussions of the topic.

REDACTED FOR CONGRESS



Features of the "Harm Inducing Narratives" archetype:

- It is normal to express uncertainty or reasonable doubt about the relevant topic, and so we agree that removing individual content objects is not defensible
- The prevalence of content expressing doubt on platform may be substantially higher than among authoritative sources, such that FB discussions may not be producing balanced discussions of the topic.
- People exposed to this content repeatedly may act in ways which are harmful to themselves, others, or society at large

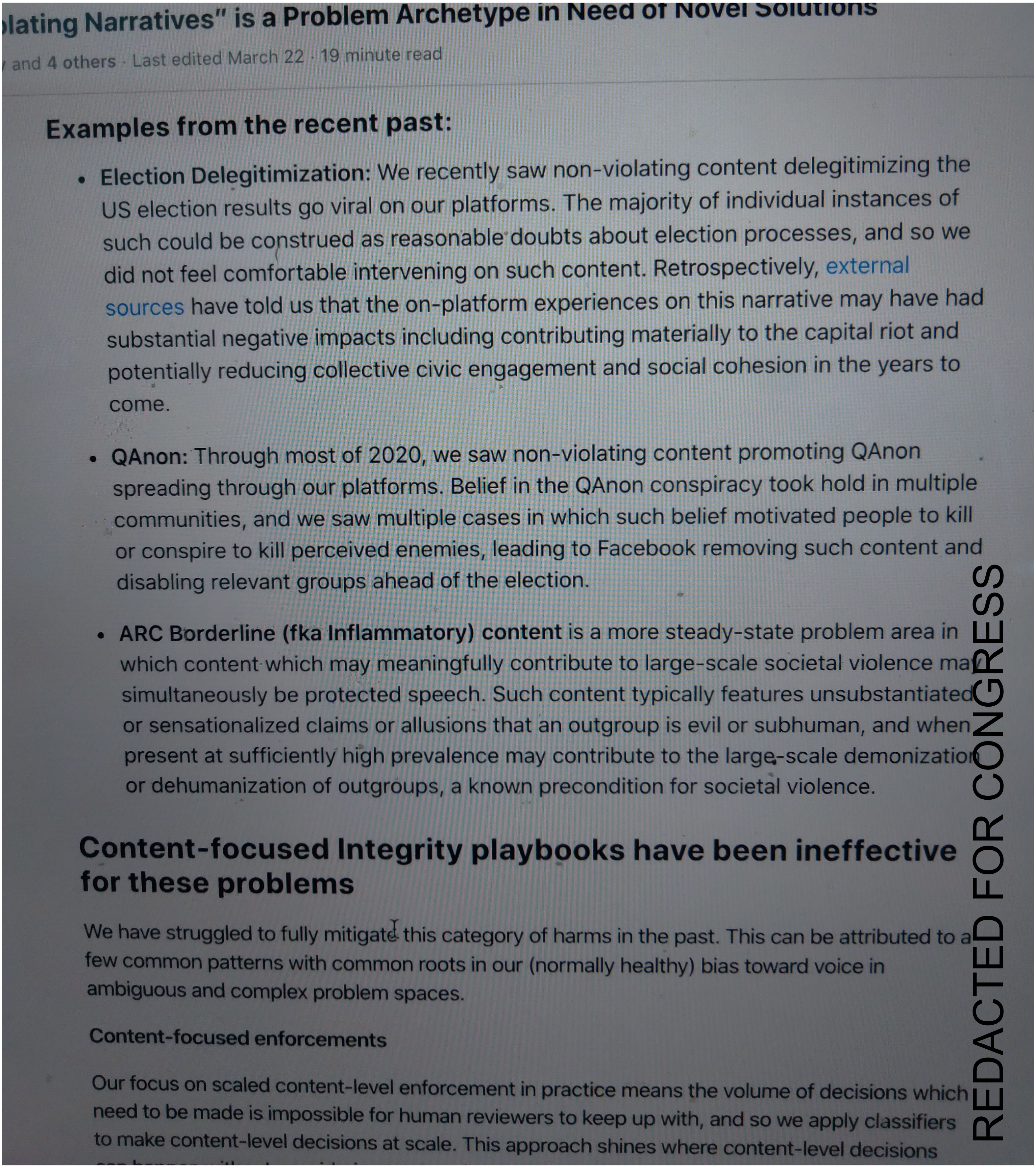
These features also characterize the vaccine hesitancy challenges facing the Facebook community, and the world, today.

- Facebook's [WIP] "Barriers to Vaccination" policy opens with the statement that even content that "could present a barrier to vaccination in any context" is "non-violating".
- Content consistent with vaccine hesitancy has been highly prevalent, including ~25-50% of vaccine content vpvs and 50%+ of vaccine comment vpvs in studies from earlier this year. In some on-platform communities, this content may comprise as much as ~5% of all feed vpvs.
- We expect that exposures to relevant content may meaningfully effect users' intent to vaccinate. An external study found that a 6.2 point change in intent to vaccinate between users exposed to a single piece of authoritative vs. misinformative content, while the health team has set a goal to increase intent to vaccinate by 1%, suggesting we internally believe that on-platform experiences can meaningfully effect this critical outcome.

Examples from the recent past:

- **Election Delegitimization:** We recently saw non-violating content delegitimizing the US election results go viral on our platforms. The majority of individual instances of such could be construed as reasonable doubts about election processes, and so we did not feel comfortable intervening on such content. Retrospectively, external sources have told us that the on-platform experiences on this narrative may have had substantial negative impacts including contributing materially to the capital riot and

REDACTED FOR CONGRESS



lating Narratives" is a Problem Archetype in Need of Novel Solutions

and 4 others · Last edited March 22 · 19 minute read

Examples from the recent past:

- **Election Delegitimization:** We recently saw non-violating content delegitimizing the US election results go viral on our platforms. The majority of individual instances of such could be construed as reasonable doubts about election processes, and so we did not feel comfortable intervening on such content. Retrospectively, external sources have told us that the on-platform experiences on this narrative may have had substantial negative impacts including contributing materially to the capital riot and potentially reducing collective civic engagement and social cohesion in the years to come.
- **QAnon:** Through most of 2020, we saw non-violating content promoting QAnon spreading through our platforms. Belief in the QAnon conspiracy took hold in multiple communities, and we saw multiple cases in which such belief motivated people to kill or conspire to kill perceived enemies, leading to Facebook removing such content and disabling relevant groups ahead of the election.
- **ARC Borderline (fka Inflammatory) content** is a more steady-state problem area in which content which may meaningfully contribute to large-scale societal violence may simultaneously be protected speech. Such content typically features unsubstantiated or sensationalized claims or allusions that an outgroup is evil or subhuman, and when present at sufficiently high prevalence may contribute to the large-scale demonization or dehumanization of outgroups, a known precondition for societal violence.

Content-focused Integrity playbooks have been ineffective for these problems

We have struggled to fully mitigate this category of harms in the past. This can be attributed to a few common patterns with common roots in our (normally healthy) bias toward voice in ambiguous and complex problem spaces.

Content-focused enforcements

Our focus on scaled content-level enforcement in practice means the volume of decisions which need to be made is impossible for human reviewers to keep up with, and so we apply classifiers to make content-level decisions at scale. This approach shines where content-level decisions

REDACTED FOR CONGRESS