WHITE PAPER

Beyond Graphic Content:
The Need for Wellbeing in Non-Egregious Moderation

ZEVO HEALTH

Dr. Michelle Teo
Clinical Director, Zevo Health

TABLE OF CONTENTS

INTRODUCTION

Much of what is written about content moderation and Trust & Safety is around egregious content and how to support people who are reviewing this content every day. But for many Content Moderators, they support non-egregious content workflows, which includes content like SPAM, ads moderation, hacked accounts and identity verification, and copyright infringement. The need to protect these Content Moderators is as important as their colleagues in the egregious queues. Let's look at why this is so important.

EGREGIOUS CONTENT CAN SHOW UP ANYWHERE, ANYTIME

One of the major reasons for ensuring Content Moderators have access to psychological health and wellbeing services is because egregious content may accidentally appear in non-egregious workflows. This could be due to various reasons such as bad actors manipulating content and exploiting loopholes in policy, glitches in automated content filtering, or human error in user reporting or content moderation itself.

> **"Harmful content comes in all shapes and sizes, all content formats – from video, audio, text, and images, in a wide range of languages – predominantly those not covered by automated detection technologies. Moreover, content shared or produced by bad actors may be disguised as something inculpable – using creative linguistic nuances, hiding text in images, l33t speak, and more – allows bad actors to hide content from automated detection methods."**
>
> https://www.activefence.com/blog/content-moderators-cost-of-burnout/

> **"Most trust-and-safety policy work is about establishing specific cases where a particular action is warranted. In 2013, platforms were still mostly caught off guard by nonconsensual pornography; now, every major platform has a formal workflow for reporting that content. "I'm in this photo and I don't like it" is well-known enough to be a meme. Similar workflows have been developed for inauthentic accounts and manipulated media. The maddening challenge of moderation is always to draw clear lines between what's allowed and what isn't — but in these specific cases, we've been able to draw those lines."**
>
> https://restofworld.org/2023/exporter-gaza-social-media-moderation/

> **"Many of the posts have been seeded by Hamas to terrorize civilians and take advantage of the lack of content moderation on some social media sites — particularly X and Telegram — according to a Hamas official and social media experts interviewed by The New York Times. The strategy mirrors efforts by extremist groups like the Islamic State and Al Qaeda, which took advantage of the lack of guardrails at social media companies years ago to upload graphic footage to the internet. Social media companies reacted then by removing and banning accounts tied to those groups."**
>
> https://www.nytimes.com/2023/10/10/technology/hamas-violent-videos-online.html

AI IS ABOUT TO MAKE CONTENT MODERATION MORE DIFFICULT, ESPECIALLY NON-EGREGIOUS

One AI is about to create more content, and content which is harder to identify as obviously fake. The improvements in AI mean that images, videos and sound files have all become difficult to determine whether something is real or fake. This type of content will likely be part of non-egregious queues such as fraud investigations and identity verification.

Recently a bank employee paid scammers who had orchestrated a video call with multiple members of the bank's financial team, including the CFO, who were all deepfakes. Believing everyone else on the call was real, the worker agreed to remit a total of $200 million Hong Kong dollars – about $25.6 million.

A recent deepfake phone message was sent to multiple New Hampshire households purportedly from Joe Biden asking people not to vote in the forthcoming primaries and to "keep your vote for November." According to Time Magazine, many have warned that new artificial intelligence-powered video and image generators will be used this year for political gain, while representation for nearly half of the world is on the line in polls. But it's audio deepfakes that have experts worried now. They're easy to edit, cheap to produce and particularly difficult to trace. Combine a convincing phone message with a voter registration database, and a bad actor has a powerful weapon that even the most advanced election systems are ill-equipped to handle, researchers say.

BOREDOM TIED TO MEANING AND PURPOSE IN THE ROLE

One of the most important aspects of content moderation is that there is consistency in how content is treated. Working on non-egregious content can be boring, and therefore there is a danger of employees lacking a sense of meaning and purpose in their work. Productivity is an important metric that can impact bonuses and salary increases for moderators, but quality is also important. And maintaining the quality of moderation can be difficult when the content being reviewed is non-egregious, and potentially boring.

Research evidence demonstrates that a sense of meaning and purpose in various areas of our lives – including in our occupations – helps drive better

mental health outcomes. This is the over-arching theme of therapies like Acceptance and Commitment Therapy. People working on the egregious queues know that they are 'fighting the good fight' and taking damaging content off the platform. For non-egregious content, the validation to the employee is less clear. It can be difficult for moderators working on non-egregious workflows to connect with the importance of their work, compared to someone working on the CSAM queue, for example. There is therefore, a different type of risk to the psychological health of Content Moderators who are experiencing boredom and lacking meaning and purpose in their roles.

HIERARCHY WITHIN CONTENT MODERATORS

Within organizations there are certain issues that come up between egregious and non-egregious teams. It is very clear why egregious teams are valuable to the business, and why they would need access to a mental health professional.

Between and within teams, there can be unspoken tensions stemming from perceived hierarchies within the organization. CSAM, TVEC and SSI content are seen as the 'most important' content moderation areas, therefore moderators supporting these workflows may be perceived as more valuable. However, there is often not a clear distinction in work practices relating to this perceived value and tensions may arise in various scenarios. This can lead to reduced loyalty and sense of belonging in the organization and can impact team dynamics and organizational cohesion. For example:

- Jose supports non-egregious workflows but is paid the same as Sanaa who works on highly egregious workflows. Sanaa feels she should be paid more than Jose because she is exposed to more graphic imagery that can impact her psychological health.

- Jose is faced with a lower volume of content daily while Sanaa must work non-stop and still experiences backlogs of content. Sanaa is frustrated that she cannot clear her backlogs and that Jose never has to worry about backlogs.

- Jose can easily reach his quality targets because he reviews less content and has more time for each case, so he receives his full bonus monthly whereas

Sanaa has lower quality scores because her cases are more complex, and her bonus is docked each month. Sanaa thinks the bonus structure is unfair.

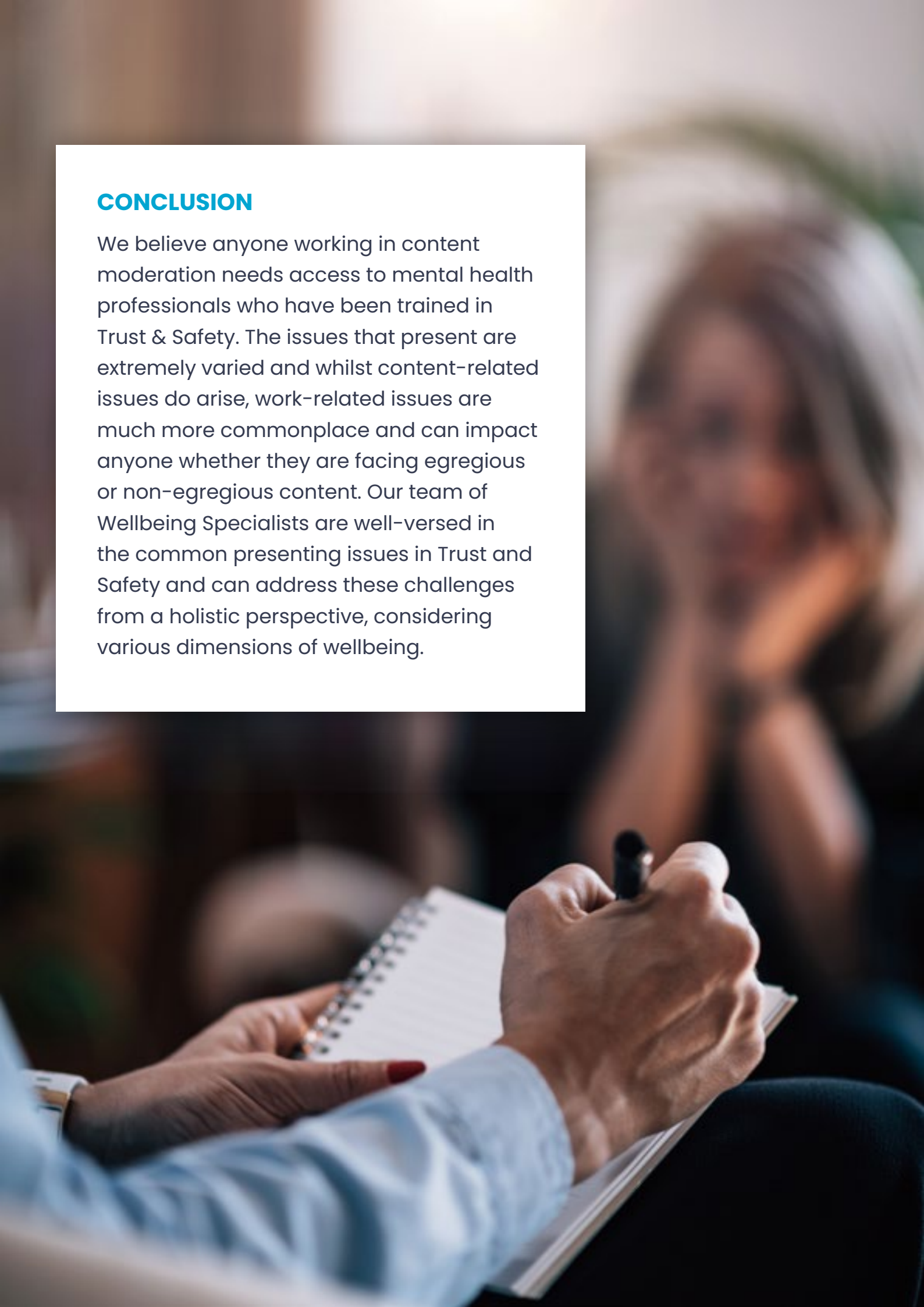
- Jose has recently experienced a bereavement but is hesitant to seek support from a Wellbeing Specialist because his issue is not related to graphic content, and he worries that he is "taking time away" from moderators like Sanaa who might "need it more".

- Jose wants to apply for a position as a Policy SME but was redeployed to non-egregious workflows 6 months ago. He is now unfamiliar with most of the policies relating to egregious workflows. He thinks it is unfair that his chances of being successful are slimmer than his peers who have supported egregious workflows when he was not given a choice in his redeployment.

WORK RELATED ISSUES

Our data demonstrates that between 35% and 45% of issues raised in counselling sessions and peer support sessions with our Wellbeing Specialists are related to work issues and have little to do with the content being viewed, even in egregious content queues. Content Moderators are working on tight deadlines, with high productivity metrics, therefore everyday stress and strain from working in a team will be an important part of their wellbeing. Being able to access support with an accredited mental health professional is important for psychological wellbeing, but many Content Moderators supporting non-egregious workflows may be left feeling that their work isn't worthy of needing support.

Aside from moderators themselves, other Trust & Safety professionals such as management, subject matter experts in policy, learning and development, and even HR functions may be indirectly exposed to egregious and non-egregious materials. We cannot forget the importance of offering support to these adjacent functions to ensure the overall wellbeing of an organization.



CONCLUSION

We believe anyone working in content moderation needs access to mental health professionals who have been trained in Trust & Safety. The issues that present are extremely varied and whilst content-related issues do arise, work-related issues are much more commonplace and can impact anyone whether they are facing egregious or non-egregious content. Our team of Wellbeing Specialists are well-versed in the common presenting issues in Trust and Safety and can address these challenges from a holistic perspective, considering various dimensions of wellbeing.

ZEVO HEALTH

Empowering the psychological health and wellbeing of content moderation teams

www.zevohealth.com

hello@zevohealth.com