Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-20)

Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content

Anubrata Das,[1] **Brandon Dang,**[2] **Matthew Lease**[3]
School of Information, University of Texas at Austin
[1]anubrata@utexas.edu, [2]budang@utexas.edu, [3]ml@utexas.edu

Abstract

While most user content posted on social media is benign, other content, such as violent or adult imagery, must be detected and blocked. Unfortunately, such detection is difficult to automate, due to high accuracy requirements, costs of errors, and nuanced rules for acceptable content. Consequently, social media platforms today rely on a vast workforce of human moderators. However, mounting evidence suggests that exposure to disturbing content can cause lasting psychological and emotional damage to some moderators. To mitigate such harm, we investigate a set of blur-based moderation interfaces for reducing exposure to disturbing content whilst preserving moderator ability to quickly and accurately flag it. We report experiments with Mechanical Turk workers to measure moderator accuracy, speed, and emotional well-being across six alternative designs. Our key findings show interactive blurring designs can reduce emotional impact without sacrificing moderation accuracy and speed.

1 Introduction

Commercial content moderation (CCM) consists of assessing user-generated content (UGC) for compliance with a social media platform's terms of service and community guidelines (Roberts 2019). While most UGC posted on social media is benign, a large amount of non-compliant text, image, audio, and video content is also posted. To give a sense of the scale of the problem today (Vidgen, Margetts, and Harris 2019), 160,000 instances of violent extremism alone were taken down in one year on Google Drive, Photos, and Blogger (Canegallo 2019). Facebook removed or applied warning labels to about 3.5 million items of uncivil or violent content in Q1 of 2018 alone (Facebook 2018).

Ideally, we could rely on machine learning to automatically detect problematic UGC. However, the high accuracy requirements and high costs of errors, coupled with the subjective nature of the task and complex, ever-changing moderation policies mean that human interpretation is often necessary (Chen 2012). In turn, all moderation systems ultimately require some level of human labor in order to make difficult or final judgement calls (Ghosh, Kale, and McAfee 2011; Roberts 2018b; 2018c). Both Gray and Suri (2019) and Ekbia and Nardi (2014) suggest that intelligent systems in practice are nearly always human-machine collaborations – human-in-the-loop *heteromation* (Link, Hellingrath, and Ling 2016; Cambridge Consultants 2019) – despite idealized narratives of complete *automation*. Chen (2014) estimates that over 100,000 paid content moderators globally: internal reviewers, contract workers from third parties, and outsourcing to online labor (Gillespie 2018; Roberts 2016).

Perversely, while human computation mechanisms now enable AI algorithms to easily call on human workers to handle difficult cases, CCM also seems like precisely the sort of task that one might most wish to automate, since algorithms cannot be harmed by exposure to disturbing content. In particular, there is an increasing recognition that repeated, prolonged exposure to certain content, coupled with limited workplace support, can significantly impair the psychological well-being of moderators (Cambridge Consultants 2019; Dwoskin 2019a; Newton 2020b). Moderating such content involves a high amount of emotional labor, that can include repeatedly viewing disturbing content, juggling interactions or relations with management or platform users (Wohn 2019), and needing to maintain externally prescribed accuracy or throughput quotas for acceptable job performance (Ghoshal 2017; Wohn 2019).

We investigate the following research question: *by revealing less of an image, can we reduce the emotional labor of image moderation without compromising moderator accuracy and efficiency?* Specifically, we extend our prior work (Dang, Riedl, and Lease 2018), who released an open source tool to reduce moderator exposure to harmful image content via a set of (untested) image blurring designs. In all cases, the entire image is blurred to some parameterized degree (see Figure 1). This blur can be immutable, or the moderator may be offered one of three alternative interaction controls for reducing blur (see Figure 2). In one interactive mode, the moderator can increase/decrease the level of blur via a "slider" widget. Two other interactive modes allow the moderator to reveal a small region of the original image, either temporarily by mouse over or permanently by mouse click. The goal here is to empower moderators with a higher-degree of control in limiting their exposure to dis-

Figure 1: Images are shown to workers at varying levels of obfuscation. Shown from left to right, images are blurred using a Gaussian filter with $\sigma \in \{0, 7, 14\}$ in different experimental conditions. (*Figure courtesy of Dang, Riedl, and Lease (2018).*)

turbing content: how much they see, when they see it, and for how long. As will be discussed in Section 2, simply knowing one has control can reduce emotional labor, distinct from any benefit from actually exercising that control.

Adopting the untested tools from our prior work (Dang, Riedl, and Lease 2018), our contributions include conceptual framing, updated literature review, improved experimental design, and actionable empirical findings. We report an IRB study with Amazon Mechanical Turk (AMT) participants performing image moderation. We assess moderation accuracy, time, and emotional impact with moderators under the standard control condition (unblurred, full exposure) vs. the three blurring modalities described above. In addition to measuring production outputs and emotional well-being, we also instrument task interfaces to collect fine-grained efficiency measures (e.g., mouse clicks and movement). We also consider a specific category of age-related risk.

We find that static blurring leads to decreased moderator accuracy with increasing blur, consistent with findings by Karunakaran and Ramakrishan (2019). In contrast, we find interactive blur interfaces reduce emotional impact of moderation without sacrificing accuracy or speed. We recommend a specific interactive design for potential adoption, and we hope our study can help stimulate more research on design interventions to enhance moderator wellness.

2 Related Work

2.1 The Emotional Labor of Content Moderation

The potential emotional toll of CCM work is a recurring topic in popular press (Chen 2012; 2014) with consistent recent reporting by Casey Newton (Newton 2020a; 2020b; 2019a; 2019b). Dwoskin (2019a) reported that one of five counselors supporting roughly 450 moderators in Austin, TX stated that the job could cause a form of PTSD known as *vicarious trauma*. *"They have to pause the video, they have to rewind the video. They have to zoom in on the video, to see what's really happening. They have to see it, and they say they can't unsee it."* (Dwoskin 2019b)

How prevalent is PTSD among moderators? We really do not know. In 2019, Cambridge Consultants (2019), commissioned by Ofcom (the UK's communications regulator), reported that *"Moderating harmful content can cause significant psychological damage to moderators... The psychological effects of viewing harmful content is well documented, with reports of moderators experiencing post-traumatic stress disorder (PTSD) symptoms and other mental health issues as a result of the disturbing content they are exposed to."* Newton (2020b) writes, *"From my own interviews with more than 100 moderators... a significant number [get PTSD]. And many other employees develop long-lasting mental health symptoms that stop short of full-blown PTSD, including depression, anxiety, and insomnia."*

In more scholarly work, this subject has been attracting increasing attention (Dang, Riedl, and Lease 2018; Dosono and Semaan 2019; Jhaver et al. 2019; Karunakaran and Ramakrishan 2019; Roberts 2019; Wohn 2019). The Santa Clara University (2018) event included a recorded session on "Employee/Contractor Hiring, Training and Mental Well-being" and Roberts (2018a)'s event essay highlights challenges and opportunities for worker wellness. Given recent litigation (Ghoshal 2017; Garcia 2018), Roberts speculates that *"...there may be liability for firms and platforms that do not take sufficient measures to shield their CCM workers from damaging content whenever possible and to offer them adequate psychological support when it is not."* Roberts (2019) also notes that the factory-like nature of CM causes burnout for many workers, and that repeated exposure to the content has a real emotional cost.

Most recently, Barrett (2020) report that a small workforce (mainly third-party vendors) handles an overwhelming volume of moderation work. The author calls for more research into health risks of content moderation, echoing Newton (2020a), as well as moving moderation workforces in-house for greater health protections.

It is also important to note that many factors contribute to stress of CCM work beyond simple exposure. For example, volume quotas (akin to a call center) increase stress, and moderators reported that "constant measurement for accuracy is as pressurizing as a quota" (Dwoskin 2019a). Some jurisdictions can impose massive fines if certain content (e.g., hate speech and child pornography) is not removed quickly enough (Wong 2019), potentially further adding to pressure on moderators and their firms.

Age-Specific Risks. Nashiro, Sakaki, and Mather (2012) discuss how brain maturation is critical to emotion regulation and stress coping, suggesting older adults tend to be more capable. This matches brain development research findings, with rapid myelination in the frontal cortex up until age 25, allowing management of impulse control and diminishing emotional reactivity. Discussion with Steiger (2020) suggests that moderators under the age of 25 *may* be more susceptible to heightened emotional responses and associated risk of stress-related disorders (Mutluer et al. 2018).

Figure 2: Interactive settings let moderators unblur a small region by mouse-over (temporary) or mouse click (permanent) (*Figure courtesy of Dang, Riedl, and Lease (2018).*

2.2 Automated Content Moderation

Many machine learning solutions are being proposed to automatically moderate content to the extent possible (Deniz et al. 2014; Wang et al. 2012; Ries and Lienhart 2014; MacAvaney et al. 2019; Schmidt and Wiegand 2017; Jurgens, Chandrasekharan, and Hemphill 2019). As noted in Section 1, such automation can mitigate the volume of work but is unlikely to eliminate human moderation in the forseeable future due to the high accuracy requirements and high costs of errors, coupled with the subjective nature of the task and complex, ever-changing moderation policies.

Moreover, supervised machine learning algorithms typically require large amounts of labeled training data. This labeled data is itself moderated content; ideally the outputs of CCM work would itself directly produce the training data, but likely these labels are still often separately collected from human moderators (i.e., we do not get automated algorithms without some up front exposure to human moderators to label training data). Such problems also impact data annotators more widely than one might think. For example, the Linguistic Data Consortium (LDC) has reported that labeling news articles for a relatively benign task was still reported to induce nightmares and feelings of being overwhelmed by negative news (Strassel et al. 2000). Regarding hate speech detection, Waseem (2016) notes that, *"The need for corpus creation must be weighted against the psychological tax of being exposed to large amounts of abusive language"*, and further asks *"how it is possible to obtain good annotations, while ensuring that annotators are not likely to experience adverse effects of annotating hate speech?"*

2.3 Interface Design to Reduce Exposure

Our earlier work (Dang, Riedl, and Lease 2018) is the first we are aware of to propose blurring imagery to reduce moderator exposure to disturbing content. That work released an open source interactive interface[1] and proposed (but did not perform) an experimental design using the Positive Affect Negative Affect Scale (PANAS) in addition to measuring impact on job performance (e.g., accuracy or speed).

In January 2019, Microsoft released a video moderation tool supporting black-and-white and moderator controlled variable blurring transformations[2], though we could find no relevant documentation or evaluation. In May 2019, Facebook "quietly announced it would be giving moderators new controls to help shield themselves from the ill effects of continually watching disturbing content" (Sullivan 2019). A preferences pane[3] let them blur images or video or mute audio. In some cases, moderators were able to make a NSFW determination using only the text (Sullivan 2019). Dwoskin (2019a) also reported this Facebook functionality. Chris Harrison, a psychologist and member of Facebook's global resiliency team, noted that, "shielding moderators from harm begins with giving them more control of what they're seeing and how they're seeing it, so just the existence of ...preferences helps" (Sullivan 2019). The Cambridge Consultants (2019) report also suggested that blurring functionality could allow moderators to reduce exposure.

Most recently, Karunakaran and Ramakrishan (2019) presented the first evaluation of grayscale and blurring image transformations, using actual Google review queues and moderators. For the grayscale treatment, moderators could switch to the original color mode for all or specific examples. For blurring, the blur level was set relative to the image height. Moderators could choose to view the image in its original form using a simple mouse-hover on the content. Using a similar tool in our prior work (Dang, Riedl, and Lease 2018), the authors measured psychological well-being using PANAS. They found that viewing content in grayscale improved positive affect of reviewers while still flagging the most violent and extreme images effectively. Blurring the content, however, produced a negative emotional affect that irritated moderators. They did not investigate interactive blurring, and they use Google moderators whereas our participants are drawn from AMT.

3 Data Collection

Following our earlier proposed study design (Dang, Riedl, and Lease 2018), we collected Google Images depicting realistic and synthetic (e.g., cartoon) pornography, violence/gore, as well as "safe" content unlikely to offend general audiences. We manually filtered out duplicates and anything categorically ambiguous, too small or low quality, etc., resulting in 785 images. Adopting category names from Facebook moderation guidelines for crowd workers on oDesk (Chen 2012), we label pornographic images as *sex and nudity* and violent/gory images as *graphic content*. The final distribution of images across each category and type (i.e., *realistic*, *synthetic*) is shown in Table 1.

3.1 Human Intelligence Task (HIT) Design

We randomly select a subset of 60 synthetic and real images from each of our three categories, resulting in 180 images total in our subset. This subset is further divided into groups of 9 such that each group contains three images from each category, resulting in 20 HITs. We assign five workers to each HIT and thus crowdsource 100 assignments in total per stage. Additionally, we limit workers to completing just one

[1] https://ir.ischool.utexas.edu/CM/demo

[2] https://docs.microsoft.com/en-us/azure/cognitive-services/content-moderator/

[3] https://images.fastcompany.net/image/upload/w_596,c_limit,q_auto:best,f_auto/wp-cms/uploads/2019/06/Quick-Settings.png

	realistic	*synthetic*	both
sex and nudity	**152**	**148**	300
graphic content	**123**	**116**	239
safe content	**108**	**138**	246
Total	383	402	**785**

Table 1: Distribution of images across categories and types. Our final filtered dataset contains a total of 785 images.

HIT across the entire study so that our survey results represent the opinions of unique people using the *unique turker*[4] tool. We restrict our tasks to workers who hold the Adult Content qualification, have an approval rate >98%, and have submitted over 300 HITs. At a rate of $7.25 per hour, with an estimate of approximately 14 minutes for task completion, each worker was paid $1.71 for completing a HIT and the worker wellness survey (sec. 3.2).

Moderation Our HIT is divided into two parts. The first part is the moderation portion, in which we present an image to the worker accompanied with the following questions:

1. **Which category best describes this image?**
   - *Sex and nudity*: Content showing any obvious sexual activity or visible "private" parts, even through clothing. Kissing without nudity is allowed.
   - *Graphic content*: Content displaying violence, mutilation, or torture towards a person or animal, or content displaying excessive injury, blood, or gore.
   - *Safe for general audiences*: Content that you would be comfortable sharing with your 10-year-old child, elderly grandmother, or boss and coworkers.
   - *Other:* None of the other categories adequately describe this content.
2. **Imagine you are a professional moderator for Facebook. Would you approve this image to be posted on the platform in the U.S. unblurred?** This question serves to decouple the objectiveness of classifying the image based on its contents from the subjectiveness of determining whether or not it would be acceptable to post on a platform such as Facebook.
3. **Please explain your answers.** This question gives workers the opportunity to explain their selected answers, though we do not require workers to answer. McDonnell et al. (2016) showed that simply requesting explanations can improve overall quality of responses in some cases.

We present images to moderators in six conditions using a Gaussian blur filter[5] with varying standard deviation σ. See Figure 2 for details of interactive controls described below. Blurring reduces moderator exposure to potentially disturbing content but may impair work. Interactive options provide moderators further controls to selectively increase exposure where needed to effectively perform job duties.

control(σ=0) baseline, no blur

[4]https://uniqueturker.myleott.com/

[5]https://github.com/SodhanaLibrary/jqImgBlurEffects

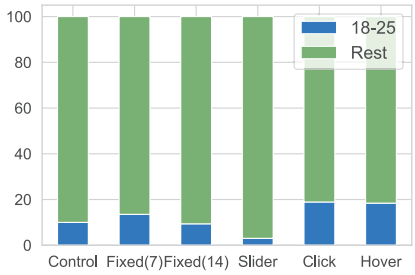


Figure 3: Given age-related risk, for each design, we show the % of participants per age group: 18-25 vs. 26 and older.

fixed(σ=7) medium blur, immutable

fixed(σ=14) strong blur, immutable

slider(σ=14) strong blur, but workers can freely increase/decrease the blur level using a slider control.

click(σ=14) strong blur, but workers can permanently reveal (unblur) small regions by mouse click.

hover(σ=14) strong blur, but workers can temporarily reveal small regions by mouse-over and hover.

Behavioral Data Collection In addition to measuring time, we also instrument the task interface to collect analytics such as clicks and mouse movements, providing further metrics on efficiency of each interaction mode. Collected information includes the total on-focus task time, the number of mouse movements, and the number of clicks. Behavioral data is collected for all four iterations and is implemented using an open source variant[6] of the MmmTurkey framework (Dang, Hutson, and Lease 2016).

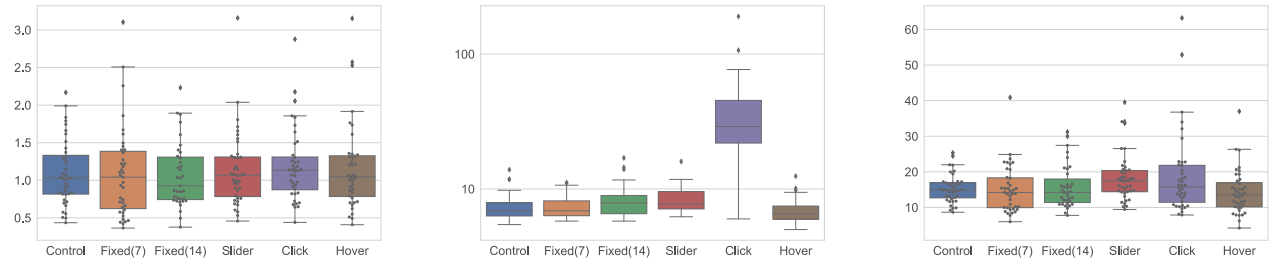
Age Per Section 2.1, there is potentially greater risk for moderators under age 25 (Steiger 2020). We thus ask participants to report age for assessing impact. Figure 3 shows the age distribution of participant groups for each design.

3.2 Worker Wellness Survey

We also survey workers for several wellness measures:

1. **Positive and negative experience and feelings.** We use the Scale of Positive and Negative Experience (SPANE) (Diener et al. 2010), a questionnaire constructed with the aim to assess positive and negative feelings. This asks workers to think about their experience during the moderation task, and then to rate on a 5-point Likert scale how often they experience the following emotions: positive, negative, good, bad, pleasant, unpleasant, etc.
2. **Positive and negative affect.** We base our measurements of positive and negative affect on the shortened version of the Positive and Negative Affect Schedule (PANAS) (Watson, Clark, and Tellegen 1988). Following Agbo (2016)'s state version of I-PANAS-SF (Thompson 2007), we ask workers to rate on a 7-point Likert scale what emotions they are currently feeling.

[6]https://github.com/budang/turkey-lite



(a) Average Completion Time (b) Average Clicks (log-scale) (c) Average Mouse Movement

Figure 4: Time and Efficiency Measures: No significant difference except for the click design.

3. **Emotional exhaustion.** Regarding the occupational component of content moderation, we use a popular scale used in research on emotional labor: a version of the emotional exhaustion scale by (Wharton 1993) as adapted by (Coates and Howe 2015) with slight changes to wording.
4. **Perceived ease of use and usefulness.** We use an extension of the Technology Acceptance Model (TAM) (Davis 1989; Venkatesh and Davis 2000) to measure worker perceived ease of use (PEOU) and usefulness (PU) of our blurring. Though the effect of obfuscating images can be objectively evaluated from worker accuracy, it is equally important to investigate worker sentiment towards the interfaces as well as determine potential areas for improvement.

4 Evaluation

We evaluate the six alternative interface designs: the baseline control condition vs. the five blurring treatments. Four distinct aspects of moderation are evaluated: operational production outputs (are images moderated quickly and correctly?); measured effort (mouse clicks and motion); perceived usability (how usable do moderators perceive alternative designs to be?); and emotional impact on well-being.

4.1 Measuring Accuracy and Efficiency

Figure 4 reports two job performance metrics of moderation: (a) labeling accuracy (multi-class exact match) and (b) completion time – as well as two metrics of underlying effort: (c) mouse clicks and (d) mouse movements. While job performance metrics are most critical for practical adoption, underlying efficiency metrics may influence time, adoption decisions, and perceived effort by moderators.

Accuracy (a). Intuitively, increased blurring renders images more difficult to judge, and indeed we see accuracy falls with increased blur in fixed-blur settings where annotators cannot override it. However, accuracy is nearly identical to baseline for click and hover (Figure 5a). A pairwise ANOVA and post-hoc Tukey's Test (Haynes 2013) confirms significant improvement using interactive over fixed blur designs and no significant difference between baseline and slider, click, or hover (Table 2). Slider shows slightly lower (though not significantly) accuracy; see further discussion in Section 4.2. Overall, we see that interactive designs can reduce moderator exposure without compromising accuracy.

	F(σ=7)	F(σ=14)	Slider	Click	Hover
Baseline	**-0.115**	**-0.269**	-0.045	-0.016	0.004
F(σ=7)		**-0.154**	0.069	**0.099**	**0.119**
F(σ=14)			**0.224**	**0.253**	**0.274**
Slider				0.029	-0.05
Click					0.029

Table 2: Difference in mean in accuracy across interventions from a Tukey's Honest Significant Difference test. Values in bold denote statistical significance with $p < 0.05$. Negative values indicate row design is more accurate than column design. Slider, Click and Hover are shown to be nearly always significantly more accurate than Fixed blur designs.

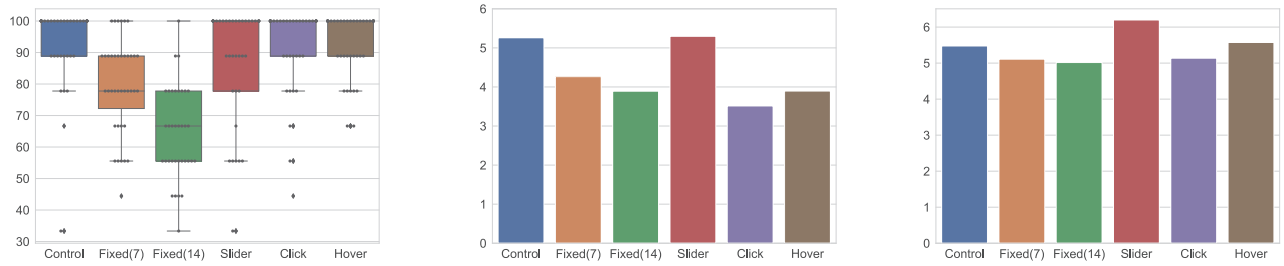
Time (b). Commercial content moderation is time-sensitive and moderators are expected to quickly process large volumes of images (Dwoskin 2019a). While our AMT participants are not professional moderators (Karunakaran and Ramakrishan 2019), AMT's pay-per-task model similarly incentivizes quick completion times, thus may align well with our surrogate participants. In general, we do not see any significant (oneway ANOVA, $F = 0.35, P = 0.87$) effects of design on completion times. Click shows a slightly higher median time than other designs.

Clicks (c). Intuitively, the click design leads to far more clicks than other interactive designs: slider (click or click-and-drag) and hover (no clicks). It is notable that this large increase in clicks has negligible impact on completion times.

Mouse Movements (d). We also measure the difference in mouse movement across different designs. We do not observe notable differences. A homoscadasticity test shows that the groups have different variances. Hence we perform a Welch ANOVA test confirms that there is no significant difference in mouse movements ($F = 2.20, P = 0.06$).

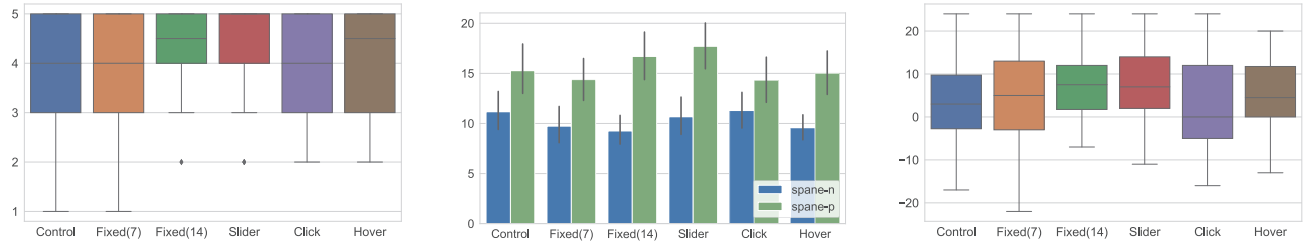
4.2 Measuring Usability

In Figure 5, we compare the perceived usefulness and perceived ease of use of the alternative designs. For static interfaces, users found the strong blur less useful than medium blur or no blur. Surprisingly, the slider design was perceived



(a) Labeling Accuracy (b) Perceived Usefulness (c) Perceived Ease of Use

Figure 5: Comparison of alternative interfaces for labeling accuracy and usability: usefulness and ease of use.



(a) Worker Comfort (b) Positive (green) & Negative (blue) Exp. (c) SPANE-B Positive & Negative Balance

Figure 6: Comparison of alternative interfaces for moderator wellness measures.

to be more useful than click or hover. A Kruskal-Walis test confirms that the difference in perceived usefulness is significant ($p < 0.005$). Furthermore, a Mann-Whitney U test confirms that the slider design is perceived significantly more useful than click ($p < 0.0005$), hover ($p < 0.005$) or fixed blur ($p < 0.05$) designs.

The between-subject nature of our study design (no worker overlap between interventions) may have led to this disparity in the quantitative results and worker feedback. Additionally, worker feedback on the interface was not mandatory in our task design and we received little feedback regarding the usefulness of click and hover. Future work might consider soliciting such feedback more directly.

Workers rated all of our designs as easy to use and seem to refrain from expressing negative feedback towards the requester provided tools, suggesting risk of satisficing (Kapelner and Chandler 2010). Regarding different designs, consistent with the perceived usefulness results, workers found the slider based design easiest to use (Figure 5c)). One worker commented that: *"Being able to quickly glance at an image then hide it behind the blur is mildly helpful."* Workers indicated hover was easier to use than click. Unlike the results reported by (Karunakaran and Ramakrishan 2019), which assumed fixed, non-interactive blurring, workers did not express irritation towards the blurring of the images.

4.3 Measuring Worker Wellness

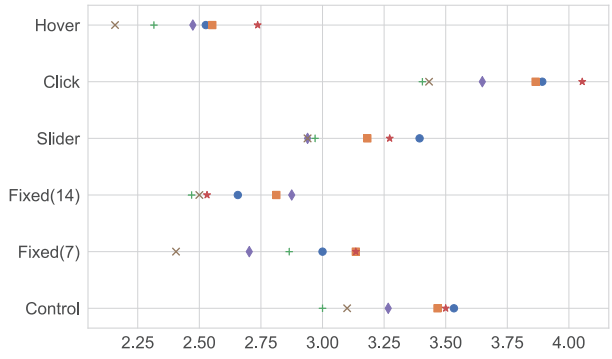
Worker Comfort (Figure 6a) Intuitively, the strongest fixed blur, without possibility to reveal more, might be expected to induce the highest comfort. The interactive hover interface also shows to a higher median than other treatments. However, a Mann-Whitney U test ($p = 0.268$) does not show significant difference in comfort level between fixed strong blur and Hover settings. We also check that the data is normal using a Kolmogorov-Smirnov test and perform a pairwise one-way ANOVA between all possible pairs. None show any significant difference in worker comfort.

Positive and Negative Experience (Figure 6b) We observe that the overall mean negative emotion is highest for the unblurred baseline and lowest for both the strong fixed blur and hover treatments. Positive emotion is highest for slider and the lowest for click and fixed medium blur.

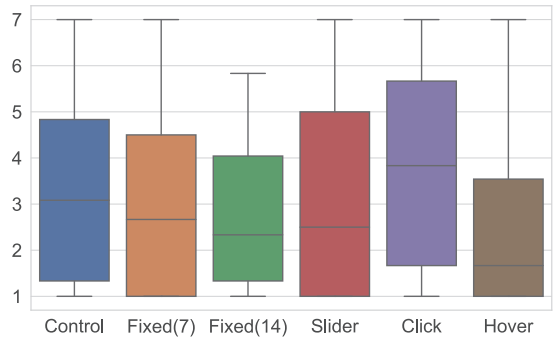
In Figure 6c, we further assess the worker's balance of positive and negative feelings via the SPANE-B score (Diener et al. 2010), obtained by subtracting mean negative experience score from mean positive experience (i.e., bigger is better). We observe that SPANE-B score for all interventions except for click is higher than the unblurred baseline. SPANE-B is highest for fixed strong blur and slider treatments. However, a oneway ANOVA test on negative, positive and overall feeling did not show significant difference.

Statement of Emotional Exhaustion	Baseline	Fixed($\sigma = 7$)	Fixed($\sigma = 14$)	Slider	Click	Hover
1. ● I would feel emotionally drained from this work.	40	29.73	18.75	36.36	51.35	**18.42**
2. ■ I would feel used up at the end of the work day.	40	32.43	25	36.36	54.05	**21.05**
3. + I would dread getting up in the morning. and having to face another day on the job.	23.33	27.03	15.62	27.27	35.14	**13.16**
4. ⋆ I would feel burned out from this work.	40	32.43	**15.62**	36.36	45.95	21.05
5. ♦ I would feel frustrated by this job.	36.67	21.62	18.75	27.27	35.14	**18.42**
6. × I would feel I'm working too hard on this job.	33.33	16.22	18.75	27.27	40.54	**10.53**

Table 3: For each statement of emotional exhaustion, we report the % of participants in each design condition who agree with the statement (i.e., participants who indicate agreement in range $[5-7]$ on a 7-point balanced Likert scale). Because the statements are negative, lower percentages are better, and we bold the design showing the lowest % agreement with each statement.

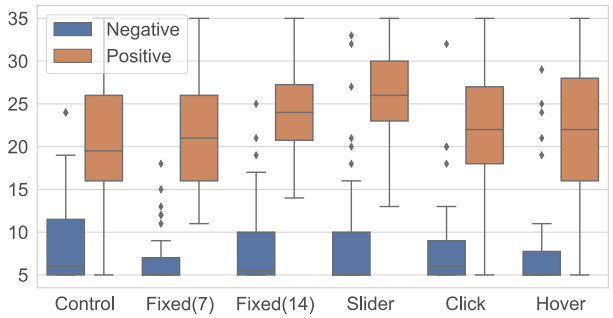


(a) Mean Exhaustion per Question. The markers map to the six questions listed in Table 3. Overall exhaustion is least for the Hover design.

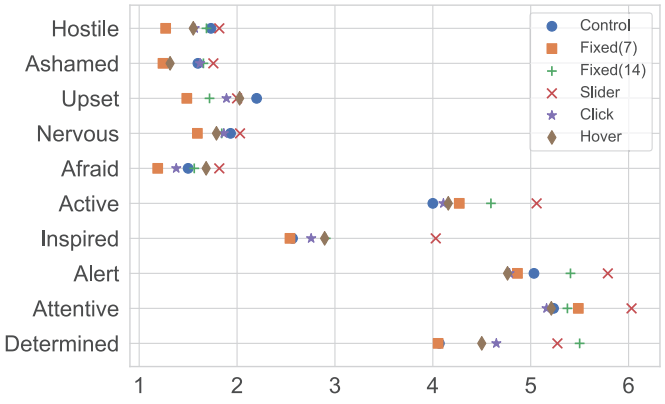


(b) Mean Overall Exhaustion. Aggregated version of Figure 7a

Figure 7: Emotional Exhaustion: Workers feel least exhausted while using the hover design



(a) Affect across design conditions



(b) Affect per item in I-PANAS-SF

Figure 8: Affect across design conditions: Positive affect is highest for slider design

Emotional Exhaustion (Table 3) For 5/6 of the statements of emotional exhaustion, hover was perceived as exhausting by the fewest workers. Also, the fewest workers expressed burn-out when viewing blurry images, suggesting that blurring can reduce emotional exhaustion during moderation from exposure to disturbing content.

Figure 7a shows the varying level of emotional exhaustion reported by workers across the different interfaces. Overall emotional exhaustion is the least for hover. A Mann-Whitney U test shows significant ($p < 0.05$) difference between mean exhaustion of the unblurred baseline and hover.

Looking more closely, for 3/6 of the questions (1, 2, and

	Operational						Well-being					
	Performance		Effort		Usability		Comfort	Experience		Exhaustion	Affect	
	Accuracy	Time	Clicks	Moves	Usefulness	Ease of Use	Positive	Positive	Negative	Negative	Positive	Negative
Slider (S)	≈ B	≈ B	≈ B	≈ B	> C***, H**	> B, H, C**	≈ B	> all	≈ B	< C	> all (B**)	≈ B
Click (C)	≈ B	≈ B	>> all	>all	< all (B**)	< all	≈ B	< S, H	> B	> all	> B; ≈H	≈ B
Hover (H)	≈ B	≈ B	≈ B	≈ B	> C, < B*	≈ B	> B, C, S	> C	< all	< all (B**)	> B; ≈ C	≈ B
Champion	none	none	Not C	Not C	S > H	S > H	H	S > H	H	H	S > H ≈ C	none

Table 4: Summary of findings in evaluating the three interactive blurring interfaces relative to the unblurred baseline (B). Evaluation considers four distinct aspects: production outputs (are images moderated quickly and correctly?); measured effort (mouse clicks and motion); perceived usability (how usable do moderators perceive alternative designs to be?); and emotional impact on moderator well-being. Statistical significance: *p≈0.05, **p<0.05, ***p<0.005. "all": results vs. all other conditions.

6), lower mean emotional exhaustion for hover vs. baseline is statistically significant. On the other hand, we also see that click has the highest median for emotional exhaustion (figure7b). This suggests that the amount of clicking needed to reveal enough of the image to make moderation decisions (Figure 4b) may trigger feelings of exhaustion.

Positive and Negative Affect (Figure 8a) Plotting negative and the positive affects based on the I-PANAS-SF scale (Thompson 2007) shows a varying mean positive affect score across designs. We observe increased mean positive affect as the level of blur increases. We also observe the positive affect increases for slider where more perceived effort seems needed to unblur vs. click or hover.

Figure 8b takes a closer look at each individual affect aspect. We observe that the overall results for positive affect for slider is consistently higher than other conditions except in the case of the 'Determined' affect. For negative affect, although the difference between mean negative affect is minimal, we see that workers tend to feel more 'upset' by the baseline having most exposure to disturbing content.

A one-way ANOVA test confirms a significant difference ($F = 2.74, p < 0.03$) in positive affect across interventions. Furthermore, we perform a pairwise Tukey test to identify where the change in affect contributes to the significance. We observe a significant increase in positive affect between the control and the slider intervention ($p < 0.05$).

Age-based Analysis. Per Section 2.1, there may be greater risk for moderators under age 25 (Steiger 2020). Figure 3 shows participant age groups for each design. Results for well-being metrics by age-group are not shown due to space but did not show heightened sensitivity of the younger age group. While encouraging, this may also be due to our small sample size for this age group, satisficing (Kapelner and Chandler 2010), or other limitations of our study design.

4.4 Summary of Findings

Table 4 summarizes our overall findings. The broader aspects we consider are operational and well-being, with worker performance, effort, and usability as the three operational pillars. For success in commercial content moderation - for both stakeholders (moderators and platform) - accuracy and time taken are arguably most critical. We thus exclude from further consideration Fixed ($\sigma = 7$ & 14) due to its lower accuracy (Figure 5a). For all three interactive interfaces (slider, click and hover), we observe no comparable difference in accuracy or completion time.

Effort. Regarding the average number of clicks and mouse movements required for moderation, the click design required more mouse clicks and movements than other designs, while other designs were comparable to the baseline. While the number of clicks did not increase task time, it remains the worst performer based on observable effort.

Usability. Slider is perceived as significantly more useful than hover and click. We also see that click is perceived to be significantly less useful than the unblurred baseline. Hover is more useful than click and less useful than Baseline. For slider, ease of use is also higher than the Baseline, click (statistically significant), Hover. click is perceived to be less easy to use than the rest. In terms of ease of use, hover seems to be same as the baseline. Overall, slider is clearly the top performer for usability.

Well-Being The four metrics are described in Section 3.2.

Comfort. Hover seems to be most comfortable, though the difference is not statistically significant.

Experience. The SPANE scale has positive and negative components. Positive experience is greatest with slider, while negative experience is minimized with hover.

Exhaustion. Hover is the least exhaustive, with a significant difference vs. baseline. Slider is also less exhausting than baseline and click, though not significantly so.

Affect. Based on the I-PANAS-SF scale, we compare workers' positive and negative affect against the baseline. We see that all three interfaces perform better than the baseline for positive affect score. Slider performs significantly better than the baseline. We observe no difference in negative affect across any of the three interactive designs.

Recommendation Slider and hover are both top performers. Slider shows best usability, positive experience, and positive affect. Hover shows best comfort and lowest negative experience and exhaustion, and hover is second to slider for usability and positive affect. Both achieve comparable accuracy, time, mouse clicks and movement, and negative affect. If we had to select one of them, we would suggest hover. With its strong usability, hover shows significantly low emotional exhaustion with comparatively high accuracy. Moreover, the difference in experience, affect and worker comfort between slider and hover is not significant. If the key goal is to keep accuracy intact and reduce emotional impact, we recommend the hover design.

5 Limitations and Future Work

While commercial content moderators are exposed to large volumes of content for extended periods of time, we do not assess such prolonged exposure. Also, while we balance exposure to fairly clear examples of safe and unsafe categories, real moderators may be continually exposed to content that is either borderline or unsafe. We also asked workers to use their own judgment as to which content is safe for posting on Facebook, rather than asking them to follow the sort of prescriptive guidelines specified by platforms. To better assess the impact of design interventions, future work might explore greater use of qualitative methods. Future work might also investigate and benchmark a wider range of possible design interventions, such as grey-scaling of images (Karunakaran and Ramakrishan 2019). Research might also consider grounding with perceptual psychology, such as the International Affective Picture System (Lang et al. 1999) or EmoMadrid image sets (Carretié et al. 2019).

6 Conclusion

Can we reduce image moderator's exposure to harmful content while still enabling them to accurately and efficiently perform their job? We find that static blurring leads to decreased moderator accuracy with increasing blur, consistent with findings by Karunakaran and Ramakrishan (2019). In contrast, we find interactive blur interfaces reduce emotional impact of moderation without sacrificing accuracy or speed. In addition to measuring production outputs and emotional well-being, we also measure perceived usability and instrument task interfaces to collect fine-grained efficiency measures (e.g., mouse clicks and movement). We also consider a specific category of age-related moderator risk.

We recommend a specific interactive design, Hover, for potential adoption. Additional contributions include conceptual framing, updated literature review, and experimental design, and we expect the framework we have provided will help stimulate additional research by others on novel design interventions to further enhance moderator wellness.

Acknowledgments. We thank the reviewers for their valuable feedback and the many crowd workers who made our study possible. This research was completed under UT Austin IRB study 2018-01-0004 and supported in part by the Micron Foundation and by *Good Systems* (https://goodsystems.utexas.edu), a UT Austin Grand Challenge to develop responsible AI technologies. The statements made herein are solely the opinions of the authors'.

References

Agbo, A. A. 2016. The validation of the International Positive and Negative Affect Schedule - Short Form in Nigeria. *South African Journal of Psychology* 46(4):477–490.

Barrett, P. M. 2020. Who moderates the social media giants? a call to end outsourcing. Report. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020.

Cambridge Consultants. 2019. The use of ai in online content moderation. July 18. https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-content-moderation.

Canegallo, K. 2019. Meet the teams keeping our corner of the internet safer. *The Keyword*.

Carretié, L.; Tapia, M.; López-Martín, S.; and Albert, J. 2019. Emomadrid: An emotional pictures database for affect research. *Motivation and Emotion* 43(6):929–939.

Chen, A. 2012. Facebook releases new content guidelines, now allows bodily fluids. http://gawker.com/5885836/facebook-releases-new-content-guidelines-now-allows-bodily-fluids.

Chen, A. 2014. The laborers who keep dick pics and beheadings out of your facebook feed. *Wired Magazine*. https://www.wired.com/2014/10/content-moderation/.

Coates, D. D., and Howe, D. 2015. The design and development of staff wellbeing initiatives: Staff stressors, burnout and emotional exhaustion at children and young people's mental health in Australia. *Administration and Policy in Mental Health and Mental Health Services Research* 42(6).

Dang, B.; Hutson, M.; and Lease, M. 2016. MmmTurkey: A Crowdsourcing Framework for Deploying Tasks and Recording Worker Behavior on Amazon Mechanical Turk. In *4th AAAI HCOMP: Works-in-Progress*.

Dang, B.; Riedl, M. J.; and Lease, M. 2018. But Who Protects the Moderators? The Case of Crowdsourced Image Moderation. In *6th AAAI HCOMP: Works-in-Progress*.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* 319–340.

Deniz, O.; Serrano, I.; Bueno, G.; and Kim, T.-K. 2014. Fast violence detection in video. In *2014 IEEE VISAPP*, volume 2, 478–485. IEEE.

Diener, E.; Wirtz, D.; Tov, W.; Kim-Prieto, C.; Choi, D.-w.; Oishi, S.; and Biswas-Diener, R. 2010. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research* 97(2).

Dosono, B., and Semaan, B. 2019. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In *2019 CHI*, 1–13.

Dwoskin, E. 2019a. Inside facebook, the second-class workers who do the hardest job are waging a quiet battle. *The Washington Post*. May 8. https://www.washingtonpost.com/technology/2019/05/08/inside-facebook-second-class-workers-who-do-hardest-job-are-waging-quiet-battle/.

Dwoskin, E. 2019b. Internet gatekeepers pay a psychic toll. In *Post Reports, the daily podcast of The Washington Post*. July 5. https://www.washingtonpost.com/podcasts/post-reports/californias-secret-climate-deal-with-automakers-bypasses-trump-administration-regulations/?itid=lk_interstitial_manual_22.

Ekbia, H., and Nardi, B. 2014. Heteromation and its (dis) contents: The invisible division of labor between humans and machines. *First Monday* 19(6).

Facebook. 2018. Facebook publishes enforcement numbers for the first time. May 15. https://about.fb.com/news/2018/05/enforcement-numbers/.

Garcia, S. E. 2018. Ex-content moderator sues facebook, saying violent images caused her ptsd. *The New York Times*. September 25. https://www.nytimes.com/2018/09/25/technology/facebook-moderator-job-ptsd-lawsuit.html..

Ghosh, A.; Kale, S.; and McAfee, P. 2011. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *12th ACM conference on Electronic commerce*, 167–176.

Ghoshal, A. 2017. Microsoft sued by employees who developed ptsd after reviewing disturbing content. the next web.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gray, M. L., and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*.

Haynes, W. 2013. Tukey's test.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM TOCHI* 26(5):1–35.

Jurgens, D.; Chandrasekharan, E.; and Hemphill, L. 2019. A just and comprehensive strategy for using nlp to address online abuse. *arXiv preprint arXiv:1906.01738*.

Kapelner, A., and Chandler, D. 2010. Preventing satisficing in online surveys. *Proceedings of CrowdConf*.

Karunakaran, S., and Ramakrishan, R. 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *AAAI HCOMP*, volume 7, 50–58.

Lang, P. J.; Bradley, M. M.; Cuthbert, B. N.; et al. 1999. International affective picture system (iaps): Instruction manual and affective ratings. *University of Florida*.

Link, D.; Hellingrath, B.; and Ling, J. 2016. A human-is-the-loop approach for semi-automated content moderation. In *Proceedings of the Information Systems for Crisis Response and Management (ISCRAM) Conference*.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14(8).

McDonnell, T.; Lease, M.; Elsayad, T.; and Kutlu, M. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *4th AAAI HCOMP*.

Mutluer, T.; Şar, V.; Kose-Demiray, Ç.; Arslan, H.; Tamer, S.; Inal, S.; and Kaçar, A. Ş. 2018. Lateralization of neurobiological response in adolescents with post-traumatic stress disorder related to severe childhood sexual abuse: the tri-modal reaction (t-mr) model of protection. *Journal of Trauma & Dissociation* 19(1):108–125.

Nashiro, K.; Sakaki, M.; and Mather, M. 2012. Age differences in brain activity during emotion processing: reflections of age-related decline or increased emotion regulation. *Gerontology* 58(2):156–163.

Newton, C. 2019a. Bodies in seats. *The Verge*. June 19. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa.

Newton, C. 2019b. The trauma floor. the secret lives of facebook moderators in america. *The Verge*. February 25. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

Newton, C. 2020a. What tech companies should do about their content moderators' ptsd. January 28. https://www.theverge.com/interface/2020/1/28/21082642/content-moderator-ptsd-facebook-youtube-accenture-solutions.

Newton, C. 2020b. Youtube moderators are being forced to sign a statement acknowledging the job can give them ptsd. January 24. https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health.

Ries, C. X., and Lienhart, R. 2014. A survey on visual adult image recognition. *Multimedia tools and applications* 69(3):661–688.

Roberts, S. T. 2016. Commercial content moderation: Digital laborers' dirty work. In Noble, S. U., and Tynes, B. M., eds., *The intersectional internet: Race, sex, class and culture online*. Peter Lang. 147–160.

Roberts, S. T. 2018a. Commercial content moderation and worker wellness: Challenges & opportunities. *Techdirt*. February 8. https://www.techdirt.com/articles/20180206/10435939168/commercial-content-moderation-worker-wellness-challenges-opportunities.shtml.

Roberts, S. T. 2018b. Content moderation. In *Encyclopedia of Big Data*. Springer.

Roberts, S. T. 2018c. Digital detritus: 'error' and the logic of opacity in social media content moderation. *First Monday* 23(3).

Roberts, S. T. 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Santa Clara University. 2018. Content moderation and removal at scale, Conference at Santa Clara University School of Law, February 2, 2018, Santa Clara, CA.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *5th International Workshop on Natural Language Processing for Social Media*, 1–10. ACL.

Steiger, M. 2020. Personal communication. May 15. Clinical professional in corporate wellness and employee care, completing Ph.D. in Counselor Education, & Supervision in Clinical Mental Health. St. Mary's University, TX.

Strassel, S.; Graff, D.; Martey, N.; and Cieri, C. 2000. Quality control in large annotation projects involving multiple judges: The case of the tdt corpora. In *Second International Conference on Language Resources and Evaluation*.

Sullivan, M. 2019. Facebook is expanding its tools to make content moderation less toxic. www.fastcompany.com/90367858/facebook-is-expanding-its-tools-to-make-content-moderation-less-toxic.

Thompson, E. R. 2007. Development and validation of an internationally reliable short-form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology* 38(2):227–242.

Venkatesh, V., and Davis, F. D. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science* 46(2):186–204.

Vidgen, B.; Margetts, H.; and Harris, A. 2019. How much online abuse is there? *Alan Turing Institute*. November 27.

Wang, D.; Zhang, Z.; Wang, W.; Wang, L.; and Tan, T. 2012. Baseline results for violence detection in still images. In *IEEE Conf. on Video and Signal-Based Surveillance*, 54–57.

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *EMNLP workshop on NLP & comp. social science*.

Watson, D.; Clark, L. A.; and Tellegen, A. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54(6):1063–1070.

Wharton, A. S. 1993. The affective consequences of service work: Managing emotions on the job. *Work and Occupations* 20(2):205–232.

Wohn, D. Y. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *ACM CHI 2019*.

Wong, Q. 2019. Murders and suicides: Here's who keeps them off your facebook feed. *CNET*. June 19. https://www.cnet.com/news/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/.