Detecting and Defending Against Seizure-Inducing GIFs in Social Media

Laura South
Northeastern University
Boston, Massachusetts
south.l@northeastern.edu

David Saffo
Northeastern University
Boston, Massachusetts
saffo.d@northeastern.edu

Michelle A. Borkin
Northeastern University
Boston, Massachusetts
m.borkin@northeastern.edu

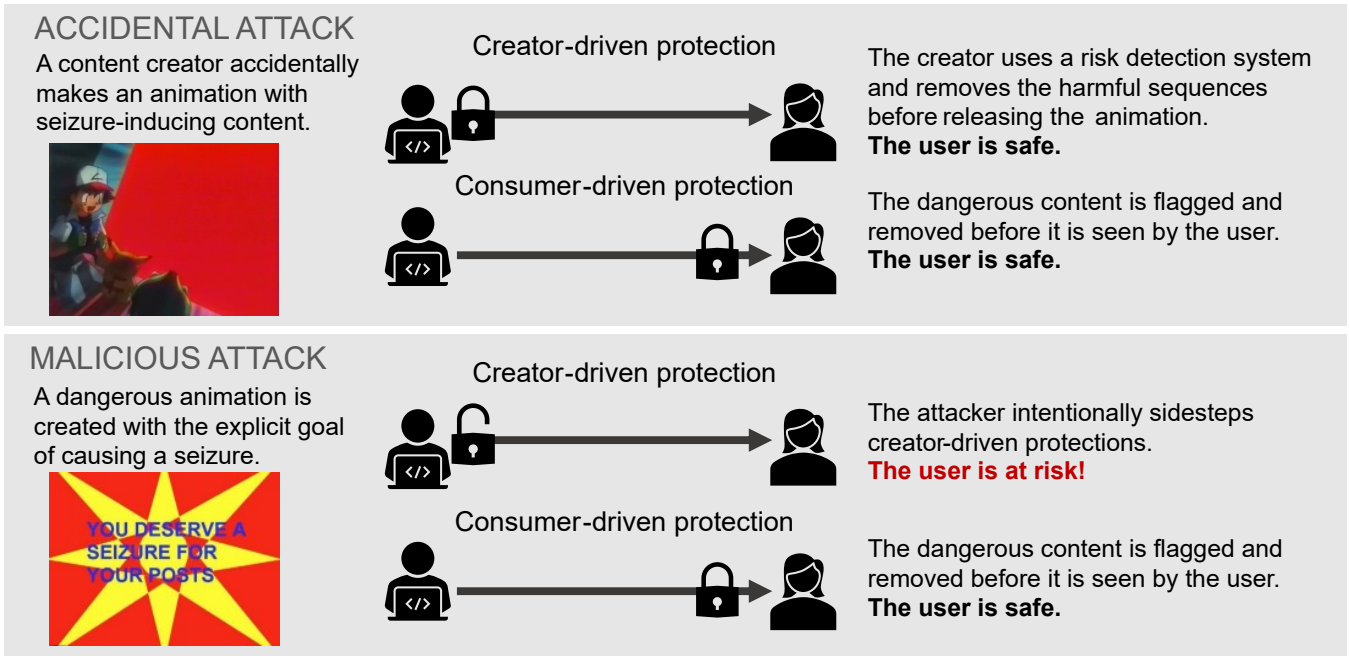

Figure 1: Flashing and strobing GIFs can cause seizures and even death when viewed by people with photosensitive epilepsy (PSE). *Creator-driven* systems rely on content creators to actively check their work for seizure-inducing content before releasing it online, leaving photosensitive individuals vulnerable when creators avoid such protections out of ignorance or malice. *Consumer-driven* systems protect users in both accidental and malicious attack scenarios.

ABSTRACT

Despite recent improvements in online accessibility, the Internet remains an inhospitable place for users with photosensitive epilepsy, a chronic condition in which certain light stimuli can trigger seizures and even lead to death in extreme cases. In this paper, we explore how current risk detection systems have allowed attackers to take advantage of design oversights and target vulnerable users with photosensitivity on popular social media platforms. Through interviews with photosensitive individuals and a critical review of existing systems, we constructed design requirements for consumer-driven protective systems and developed a prototype browser extension for actively detecting and disarming potentially seizure-inducing GIFs and videos. We validate our system with a comprehensive dataset of simulated GIFs and GIFs collected from social media. Finally, we conduct a novel quantitative analysis of the prevalence of seizure-inducing GIFs across popular social media platforms and contribute recommendations for improving online accessibility for individuals with photosensitivity.All study materials are available at https://osf.io/5a3dy/.

CHI '21, June 03–05, 2021, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

CCS CONCEPTS

• **Human-centered computing** → *Accessibility systems and tools*; **Empirical studies in accessibility**.

KEYWORDS

accessibility, photosensitive epilepsy, GIFs, human-computer interaction

ACM Reference Format:
Laura South, David Saffo, and Michelle A. Borkin. 2021. Detecting and Defending Against Seizure-Inducing GIFs in Social Media. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Epilepsy, a chronic disorder characterized by recurrent seizures, is one of the most common neurological conditions in the world, affecting approximately 65 million people [34]. Between 2 and 14% of those with epilepsy will experience seizures triggered by specific visual stimuli [19]. This condition is called **photosensitive epilepsy (PSE)**. Children and adolescents are more likely than adults to have an abnormal response to light stimulation, and females (60%) are more affected than males (40%) [18]. The consequences of encountering seizure-inducing patterns or light sequences can be severe for people with photosensitivity; even when no seizure occurs, flashing or flickering content can cause debilitating migraines or other symptoms. In extreme cases, seizures can lead to the sudden, unexpected death of someone with epilepsy (SUDEP) [29].

In 1997, an episode of the television show *Pokémon* infamously caused hundreds of children in Japan to experience seizures and migraines [44]. This incident led to increased interest in photosensitive epilepsy and photosensitive triggers, including several studies defining thresholds for dangerous flashes, transitions to and from saturated red, and repeated patterns [7, 19, 25, 46]. Automatic detection software aimed at avoiding accidental exposure to triggering content began to appear in the mid-2000s, implementing the newly-defined thresholds for photosensitive risk factors. Two well-established systems for automatic detection of photosensitive risk factors are the Photosensitive Epilepsy Analysis Tool [1], or PEAT, and the Harding Flash and Pattern Analyser [2], or FPA. Both tools are **creator-driven** systems because the person creating content is responsible for checking that their work does not contain seizure-inducing material before it is released into the world. Creator-driven protection makes sense in response to *Pokémon*-style accidental attacks; if the *Pokémon* creators had checked their work for dangerous sequences before broadcasting the episode, no viewers would have been affected.

Creator-driven protections alone are no longer sufficient when attackers can use social media to directly target photosensitive individuals with triggering content. In December 2016, journalist Kurt Eichenwald recieved a message on Twitter from an anonymous account containing a GIF with bright flashing colors and the text "You deserve a seizure for your posts". Eichenwald, who has frequently spoken and written about his experience with photosensitive epilepsy, immediately collapsed and began to seize upon opening the message [27]. In 2008, malicious hackers posted hundreds of flashing strobe animations to a forum hosted by the Epilepsy Foundation, a nonprofit dedicated to supporting people with epilepsy and their family members [37]. A similar attack occurred in November 2019, when hackers began posting strobing GIFs from the official Epilepsy Foundation Twitter account, causing seizures and migraines for hundreds of the account's followers with photosensitivity [2]. A group of six users targeted the Epilepsy Society in a coordinated attack in May 2020, sending triggering GIFs to accounts associated with the Society, including one strobing GIF sent as a reply to a tweet celebrating 263 days without a seizure [10]. Creator-driven protections, such as PEAT and the FPA, do not protect photosensitive users in malicious attacks where the content creator deliberately chooses to evade protective mechanisms (Figure 1). To prevent malicious attacks, photosensitive risk detection systems must be **consumer-driven**; they must actively protect the user as they browse, rather than relying on content creators to police their work before it is released. Only one consumer-driven system currently exists: a browser extension called EpilepsyBlocker[3]. However, we find that EpilepsyBlocker falls short when directly compared to PEAT, an established creator-driven detection systems (Section 5), by failing to detect several dangerous sequences known to cause seizures. Creating an effective consumer-driven system is challenging because it requires engagement with potential users in the photosensitive epilepsy community, careful development, and rigorous testing.

In this paper, we explore web accessibility for photosensitive individuals across social media. Through interviews with photosensitive individuals and a review of existing photosensitive risk detection systems, we establish design requirements for a consumer-driven risk detection system to actively protect against seizure-inducing content. We use these design requirements to develop PhotosensitivityPal, a prototype Chrome extension to detect and defuse seizure-inducing GIFs and videos online. We evaluate the effectiveness of PhotosensitivityPal with three datasets: GIFs simulated to include seizure-inducing sequences, GIFs randomly collected from Twitter and Tenor GIF Keyboard, and potentially dangerous GIFs manually collected across social media. Finally, we use PhotosensitivityPal to conduct the first study of the prevalence of seizure-inducing GIFs across social media, producing estimates for the overall prevalence of triggering GIFs, the prevalence of individual risk factors (i.e., flashes, red transitions, repeated patterns), and the differences in prevalence between social media platforms.

In this paper we present the following novel **contributions:**

(1) A framework for characterizing protective systems for people with photosensitivity as creator-, platform-, or consumer-driven.
(2) PhotosensitivityPal, a prototype browser extension for active detection of photosensitive risk factors in GIFs and videos.
(3) Three validation datasets, including 150 GIFs simulated with flashes, red transitions, and repeated patterns below and above established safety thresholds for photosensitive epilepsy, 200 GIFs randomly collected from Twitter and Tenor, and 137 potentially dangerous GIFs manually collected online.
(4) A study of seizure-inducing GIFs in the wild, including the first quantitative assessment of the prevalence of dangerous content on popular social media sites.

2 RELATED WORK

To contextualize our work, we first discuss existing research in the areas of (1) photosensitive epilepsy within the broader accessibility movement in HCI, and (2) systems for defending against seizure-inducing content.

[1] https://trace.umd.edu/peat
[2] https://www.hardingfpa.com/
[3] https://www.epilepsyblocker.com/

2.1 Accessibility

Browser extensions for accessibility: Browser extensions have historically been successful at improving web accessibility for users with a range of disabilities. Twitter A11y [21] and Caption Crawler [23] provide automated image descriptions to help users with visual impairments, while Firefixia [16] can modify webpage design to help people with dyslexia. Lexi [6] and Anita [35] automatically simplify page text to help users with cognitive or reading impairments. In this work, we contribute a browser extension to improve web accessibility for people with photosensitivity through active detection of seizure-inducing GIFs and videos.

Web accessibility studies: Many researchers have conducted automated and manual studies of web accessibility, often based on the standards described in the Web Content Accessibility Guidelines (WCAG) [11]. Although Guideline 2.3 in the WCAG 2.0 discourages use of seizure-inducing content on webpages, large-scale studies of web accessibility have often not focused on adherence to this guideline. Some of these studies explicitly discuss guidelines related to users with visual impairments [1, 32, 38], while others implement checkpoints defined in WCAG 1.0, an earlier version of the guidelines with less focus on cognitive and neurological disabilities [24, 30]. Park et al. pointed out the additional challenge of automatically assessing adherence for WCAG guidelines that involve analyzing images and colors, rather than text [36]. Gleason et al. examined GIFs pulled from social media and found that many lack effective alternative text descriptions, rendering them inaccessible for users with visual impairments [20]. In this paper, we contribute a novel study of online accessibility for people with photosensitivity when interacting with GIFs on four popular social media platforms.

Weaponizing seizure-inducing content: Conti et al. first identified the possibility of attackers weaponizing seizure-inducing sequences to target users with photosensitivity in a 2005 paper about security vulnerabilities in mission-critical information visualization systems [14]. Little attention has been paid since then to the potential for malicious attacks against people with PSE, despite recent high-profile incidents on social media [2, 27, 42]. In this paper, we extend Conti et al.'s initial formulation of cyberattacks targeting users with photosensitivity by documenting instances of attacks taking place on social media and developing a framework for consumer-driven protection against malicious attacks.

2.2 Photosensitive risk detection systems

Three forms of defending against seizure-inducing content: In this section, we contribute a novel framework for viewing protection systems for users with photosensitivity. We define and provide examples of three forms that defensive systems can take: creator-driven, platform-driven, and consumer-driven. Each form has its own benefits and drawbacks; an ideal online ecosystem would include defensive systems of all three forms.

Creator-driven solutions place responsibility on content creators to ensure that their work does not contain dangerous sequences before it is viewed by others. PEAT and Harding FPA are the two main examples of creator-driven systems. Both systems analyze video files offline and produce a binary pass/fail response. PEAT labels borderline videos as "caution (pass)" or "caution (fail)", although the system ultimately categorizes these borderline videos as "pass" or "fail", respectively. PEAT identifies dangerous flashes and red transitions, but does not detect dangerous repeated patterns. In this paper, we compare PhotosensitivityPal's accuracy against PEAT's because it is the most established photosensitive risk detection system currently available. We do not use the Harding FPA to test our system because the FPA is meant to analyze broadcast television and video games rather than web content and because the FPA is commercial software that is not free to access for testing.

Platform-driven solutions are implemented by the websites and apps that deliver content from creators to consumers. In response to the multiple malicious attacks that have taken place on Twitter, the company encourages photosensitive users to disable autoplay on their accounts so that videos and GIFs only play after a cue from the user [43]. In December 2019, Twitter announced that they would no longer allow users to post animated GIFs with .APNG file extensions because this file format was able to bypass autoplay settings [45]. Twitter took another step towards protecting photosensitive users in July 2020, when they banned search terms related to seizures or epilepsy on the GIF keyboard that users see when composing tweets [43]. Despite Twitter's recent progress, platform-driven solutions remain sparse and uneven between websites. Facebook and Reddit allow users to turn off autoplay, but this feature is not explicitly connected to defending against seizure-inducing content. Other popular social media sites, such as Instagram, Tumblr, and TikTok, do not allow users to turn off autoplay at all. The first major platform-driven protection from a GIF repository website occurred in September 2020, when GIPHY banned all searches with keywords related to photosensitive epilepsy [15]. Tenor GIF Keyboard, another popular GIF repository, does not provide any platform-driven layers of protection for photosensitive users. In this paper, we focus on the design, implementation, and evaluation of consumer-driven solutions for protecting users with photosensitivity, but the potential impact of comprehensive platform-driven solutions should not be ignored and is discussed in greater detail in Section 7.2.

Consumer-driven solutions analyze content as it arrives in front of the user, blocking anything with seizure-inducing sequences. EpilepsyBlocker[4] is a browser extension that tests all GIFs and videos encountered in the browser and returns a binary label of "dangerous" or "safe". To the best of our knowledge, EpilepsyBlocker is the only currently available consumer-driven solution for defending against seizure-inducing content. However, EpilepsyBlocker's creators provide no evidence from an evaluation to prove that the system can accurately detect seizure-inducing sequences. In Section 5, we find that EpilepsyBlocker fails to accurately flag the dangerous GIFs in our validation datasets. EpilepsyBlocker also does not allow users to mitigate dangerous content with low-contrast or low-saturation filters, a feature that people with photosensitivity advocated for in our user interviews (Section 3). In this work, we contribute design requirements for consumer-driven risk detection systems based on discussion with people with photosensitivity. We also contribute PhotosensitivityPal, a prototype implementation of a consumer-driven system.

Evaluating detection systems: Standardized ground-truth datasets are critical for evaluating photosensitive risk systems, but few such benchmarks currently exist. In 2016, Alzubaidi et al. released the

[4] https://www.epilepsyblocker.com/

Pattern Inducer, which can be used to generate short video clips with flashes [3]. Flashing sequences produced by the Pattern Inducer can vary according to four features: flashing rate, flashing area vs. viewed area, location of flash within viewed area, and flashing duration. The Pattern Inducer falls short of producing sequences with repeated patterns, flashes with varied luminance differences, and saturated red transitions, despite empirical proof that all three features can determine the seizure-inducing potential of a sequence [19]. In this work, we contribute a comprehensive benchmark dataset of videos with repeated patterns, flashes, and saturated red transitions.

Methods for detecting dangerous sequences: Automatic systems for detecting seizure-inducing content use one of two approaches: rule-based or machine learning. The clearly defined guidelines for flashes and red transitions established in the WCAG 2.0 lend themselves well to rule-based approaches; as a result, many currently available detection systems use rule-based approaches (PEAT, EpilepsyBlocker, Harding FPA). Alzubaidi et al. proposed a parallelized implementation of rule-based detection of flashes and achieved significant speed improvements compared to a serial alternative [3]. Barbu, Banda, & Katz created a deep learning algorithm for removing seizure-inducing flashes and patterns from videos [5]. Ensuring the accuracy of detection systems using a machine learning approach remains a challenge due to the lack of standardized ground-truth validation datasets. Barbu et al.'s system was able to seemingly disarm well-known examples of seizure-inducing GIFs and videos (e.g., GIFs from the infamous *Pokémon* episode and GIFs used in malicious attacks), although their evaluation relied on qualitative judgements from non-photosensitive participants who were presented with pairs of videos and asked to select which had fewer flashes. Non-photosensitive individuals might not be aware of the precise thresholds for seizure-inducing sequences and sequences that were found to have "fewer flashes" could still be hazardous to some photosensitive users. In this paper, we use a rule-based approach to detect seizure-inducing sequences with flashes, red transitions, and repeated patterns.

3 USER INTERVIEWS

In order to learn more about accessibility and safety for people with photosensitivity (including PSE) on social media, we conducted a series of semi-structured interviews with people with photosensitivity. These interviews were used to inform our design requirements for consumer-driven risk detection systems (Section 4.1). Recruitment and informed consent materials for our interviews were approved as an exempt study by our Institutional Review Board (IRB).

Participant demographics: Five participants were recruited via posts on the public r/epilepsy[5] subreddit discussion board and a Facebook group related to photosensitivity. All participants were over 18 years old and had experienced photosensitivity symptoms at some point in their lives. Our sample size was necessarily limited by the small pool of potential participants who have photosensitivity, are active in online support communities, and are physically able to participate in a phone or video call. Of the nine people who initially signaled interest, six responded to further contact and ended up scheduling interviews. One participant experienced a seizure minutes before our scheduled call and decided to withdraw from the study without completing an interview, leaving us with a final sample size of five.

The demographics and social media use of all study participants is presented in Table 1: four participants were female and one was male, and participant ages ranged from 19 to 50 (average age of 34.6). Four participants had been diagnosed with photosensitive epilepsy and one participant had been diagnosed with photophobia manifesting with photosensitivity symptoms (i.e., migraines and focal seizures triggered by bright light stimuli). Interviews were conducted remotely via video or phone call and lasted approximately 30 minutes. Participants were compensated $15 for their time. A full list of interview questions is provided in the Supplemental Material.

Overall safety and accessibility: We asked participants to begin by describing their level of comfort and safety when using social media. Three participants (P1, P4, P5) felt generally safe online. P1 and P5 credited their comfort to trusting friends and family not to send triggering content ("I have good friends and if they were to send me an email it would always be kind" - P1) or to their personal sensitivity levels ("Mine is really hard to trigger compared to other people with PSE. I would have to stare at it for a while before anything bad would actually happen." - P5). P4 acknowledged that he did not feel equally comfortable on all social media platforms ("TikTok is probably more risky than Facebook because it is only video, whereas Facebook has lots of reading and photos, and few videos." - P4). The other two participants (P2, P3) expressed discomfort with GIFs and videos on social media. P2 reported feeling "spotty and dizzy" after encountering triggering GIFs online and expressed that she would be "out in a seizure on the floor in a second" if she had encountered similar GIFs before starting her current medication. P3 explained that she had been sent a strobing GIF by an anonymous user on Reddit two months earlier, echoing the same malicious attack structure described in Section 1:

> "I was on Reddit doing my normal scrolling and I got a direct message from someone I didn't recognize. My Reddit was glitching out, it didn't show that they sent a GIF or anything, so I hit accept chat and I opened it up and it's this black and white GIF that's flashing like there's no tomorrow. I just had to throw my phone across the room because it was so intense. I had to get my mom to pick my phone up and delete it, just so I could pick it up again cause I'm just extremely sensitive to that. I had a killer migraine for the rest of the day from it, but if I hadn't responded so quickly I probably would have been a lot worse than that."

Other than her recent attack, P3 had not had any issues with Reddit but had struggled with encountering accidentally triggering content on Facebook: "Recently they started autoplaying videos, so I've come across more that aren't intentionally supposed to trigger photosensitivity, but just how the videos are made, they can. I've come across a couple of those on Instagram and Facebook, which it's not very fun because yes I can scroll past it, but in those two seconds that I looked at that video it can set off my photosensitivity and cause a killer migraine that sometimes is so painful that I start crying."

[5]https://www.reddit.com/r/Epilepsy

ID	Age	Gender	Photosensitivity diagnosis	Social media used
P1	50	F	Photophobia	Facebook
P2	21	F	Photosensitive epilepsy	Facebook, YouTube
P3	19	F	Photosensitive epilepsy	Instagram, Facebook, Reddit
P4	42	M	Photosensitive epilepsy	Facebook, TikTok, Twitter
P5	41	F	Photosensitive epilepsy	Reddit, Facebook, YouTube

Table 1: Demographics of interview participants, including age, gender, photosensitivity diagnosis, and forms of social media regularly used.

Protective measures: Twitter has encouraged photosensitive users to disable video autoplay to avoid accidentally viewing dangerous GIFs or videos [43]. Facebook and Reddit allow users to disable autoplay as well, although they do not explicitly recommend it as a solution for photosensitive users. Because autoplay disabling is one of the few platform-driven solutions supported by multiple sites, we asked participants if they use this feature regularly. We also asked participants to describe what other measures they take to avoid encountering dangerous content online.

(1) *Autoplay disabling*: None of the participants used autoplay disabling features as a protective measure. P1 was not aware of the feature. Two participants (P2, P3) chose not to use it because they did not find its performance to be sufficiently reliable ("It's very basic. It doesn't always filter properly. It's there. It doesn't work half the time, but it's there. Sometimes it will turn itself back on." - P2; "I did use it on Twitter briefly, but I shut it off because it was often very off. Like sometimes it blocked things that were nowhere near what it was supposed to block, so I turned it off." - P3). Two participants (P4, P5) chose to enable autoplay despite the risks because they felt autoplay was a necessary part of engaging with the website ("Wouldn't use [autoplay disabling]. For TikTok it is part of the experience." - P4) or because they feel that the content they seek out online is unlikely to include triggers ("I'm not a gamer so I don't watch the types of materials that would have that sort of [content]. I'm watching the cute cat videos." - P5).

(2) *Adjusting room and screen brightness*: Three participants (P1, P4, P5) described making environmental changes to minimize risk. P4 and P5 make sure to only watch videos in brightly lit rooms, while P1 and P4 lower the brightness and contrast on their devices.

(3) *Avoiding videos with disclaimers or warnings*: P5 felt comfortable relying on disclaimers and warnings about triggering content ("I definitely avoid videos where any kind of a warning is given, a disclaimer at the beginning of the video, I definitely don't watch anything like that. So far it's worked for me, I haven't come across any videos where I thought they should have put a warning on this. But I'm not as avid of a user as a lot of people who are a bit younger than me." - P5). P2 also relied on warnings, but pointed out that these warnings are not always effective, as exemplified by Weird Al Yankovic's "Everything You Know Is Wrong" music video: "Basically, even though he does have the warning underneath, the comments are really just people being like 'thanks Weird Al, you ruined my childhood, you almost gave me a seizure."

(4) *Relying on friends and family to filter content*: P3 described relying heavily on friends to help her avoid flashing content, even avoiding the TikTok app altogether at their suggestion ("I don't have a TikTok account because apparently there's a lot of videos around there with this filter with rainbow flashing lights in the background. My friends told me about that and I don't touch it and I avoid that completely."). Despite her friends' help, P3 still feels vulnerable online ("Occasionally I'll get told 'Hey don't go on this page on Instagram cause they post a lot of videos with flashing lights,' so I'll avoid that, but besides that I don't really have tools because i haven't really found one that really works. It's kind of like 'Let's hope I don't come across anything too stimulative'.")

Estimating the proportion of dangerous content: Participants were asked to estimate the proportion of triggering content they encounter online. P4 and P5 found the proportion to be very low (1% or less and 2-5%, respectively), while P3 estimated that 5-10% of GIFs and videos triggered a photosensitive response. P2 found that the proportion of dangerous content she encountered varied depending on how active she was on the Internet: "[the proportion] used to be closer to 50-60%, but recently it's been closer to 15-30% because I'm not really on social media much recently except for my own Facebook pages". P1 was unable to provide a percentage, but estimated that she encountered triggering content online about once a week.

Usefulness of a browser extension: Participants were asked to rate how useful a browser extension to block triggering content would be in their daily lives and what features would be most important to them in such a system. Two participants (P1, P4) did not feel that an extension was necessary in their lives because they are unlikely to encounter dangerous content ("I think if I was like a gamer, that would be really important. I'm 50, I'm not playing games." - P1) or because they do not view social media through web browsers ("If TikTok would let you select a seizure safety mode that would filter the videos that you don't want – that would be great. Nobody uses TikTok on the web browser." - P4). Three participants (P2, P3, P5) felt that an extension would be beneficial for their online safety ("I would absolutely use it, and recommend it to all my friends who also have epilepsy to use." - P2; "The second I found out about something like that I would use it immediately. No hesitation." - P3; "It's not something I would deem as something I have to have in my life, but if it were available I would absolutely use it." - P5).

Mitigating dangerous content: While participants emphasized that the most important feature of a browser extension would be its ability to accurately block dangerous content ("As long as it can

filter out the things that really mess with my head, I'd be fine with even just that. And videos that are strictly, just pure strobing effects, if it gets rid of that completely, I'm down." - P2), they also agreed that users should be able to view content labelled as dangerous in a safe manner if they choose.

Our user interviews revealed that while some people with photosensitivity feel generally safe on social media, others often encounter triggering content that makes them feel vulnerable online. We found that while none of the participants regularly use platform-driven protections on social media, such as autoplay disabling, most felt that a consumer-driven browser extension would be beneficial to their safety. In the next section, we will establish design requirements based on the findings of our interviews and introduce PhotosensitivityPal, a prototype consumer-driven protection system.

4 PHOTOSENSITIVITYPAL SYSTEM

In this section we introduce design requirements for a consumer-driven protection system and describe PhotosensitivityPal, a prototype browser extension for actively detecting and defending against seizure-inducing GIFs and videos.

4.1 Design requirements

We established four design requirements for a consumer-driven photosensitive risk detection system, based on the results of our interviews and a critical review of existing photosensitive risk detection systems:

(1) **Active**: The system should run actively on the user's device, automatically checking for photosensitive hazards in all content they encounter.
(2) **Defensive**: Until the system can determine that an item is safe, it should be assumed dangerous and blocked from the user's view.
(3) **Mitigating**: The system should allow the user to safely view content labelled as dangerous with a range of mitigation strategies, such as low-contrast and grayscale filters.
(4) **Flexible**: The system should be able to handle the wide variety of animation and video file formats found on the Internet and respond defensively if an unrecognized file format is encountered.

These design requirements will inform the development of our prototype consumer-driven system, PhotosensitivityPal.

4.2 Detecting photosensitive risk factors

Determining the seizure-inducing potential of a sequence can be split into three distinct tasks: detecting flashes, detecting transitions to and from saturated red, and detecting repeating patterns.

4.2.1 Detecting flashes. Flash detection is implemented according to the specifications described in Guideline 2.3.1 in the WCAG 2.0 [11]. Flashes are dangerous if more than three occur within one second, so flashes need to be counted in discrete segments of frames, each representing one second of the animation. If the animation is shorter than one second, the maximum number of flashes is calculated proportionately. For example, a GIF lasting 0.5 seconds is considered dangerous if it includes more than one flash, while a GIF lasting 0.7 seconds is allowed up to two flashes. According to WCAG 2.0, a general flash consists of a pair of opposing changes in relative luminance where the luminance difference between states is at least 10% of the maximum relative luminance and the relative luminance of the darker state is below 0.8. We first obtain relative luminance values for each pixel in two adjacent frames using the relative XYZ colorspace, producing a grayscale image. To identify flashing segments of the image, as illustrated in Figure 2, we isolate pixels that have relative luminance below 0.8 in one of the two adjacent frames (**Dark threshold**). We also identify pixels that change in relative luminance between the two frames by more than 10% of the brightest pixel in either frame (**Luminance difference**). Flashing pixels that meet both criteria are found through a bitwise AND between **Dark threshold** and **Luminance difference**. If more than 25% of the total pixels in each frame are flashing, a light-dark or dark-light luminance shift has occurred. A flash is recorded if two opposing luminance shifts occur within the same one-second interval.

4.2.2 Detecting transitions to and from saturated red. Saturated red transition (i.e., red flash) detection is implemented according to the specifications described in Guideline 2.3.1 in the WCAG 2.0 [11]. Red flashes are harmful if more than three occur in any one-second interval. As with detecting general flashes, we count red flashes in all possible frame segments lasting one second. This includes GIF looping behavior, where the last frame is immediately followed by the first frame again. If the animation is shorter than one second, the maximum number of red flashes is calculated proportionately. According to WCAG 2.0, a red flash is a pair of opposing transitions involving saturated red where $RedRatio \geq 0.8$ (**Equation 1**) for either or both frames involved in the transition and the change in $PureRed$ (**Equation 2**) between the two frames is greater than 20. As shown in Figure 3, we isolate pixels with $RedRatio \geq 0.8$ in either or both frames (**RR**) and pixels with a change in $PureRed \geq 20$ (**PR**). Red flashing pixels that meet both criteria are found through a bitwise AND between **RR** and **PR**. If more than 25% of the total pixels are flashing, an opposing transition has occurred. If two opposing transitions are observed in the same one-second block of frames, a red flash is present. A GIF or video is labelled dangerous if the number of red flashes observed in a one-second segment exceeds the safety threshold.

$$RedRatio = \frac{R}{R+G+B} \tag{1}$$

$$PureRed = \begin{cases} (R-G-B) \times 320 & \text{if } R-G-B > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

4.2.3 Detecting repeated patterns. Guideline 2.3.1 in the WCAG 2.0 does not provide a standard for detecting dangerous repeated patterns. We use the results of Wilkins et al.'s empirical study on pattern-induced seizures to build an effective detection system for common forms of seizure-inducing patterns. Wilkins et al. found that potentially harmful patterns contain "clearly discernible stripes where there are more than five light-dark pairs of stripes in any orientation" [46]. Stripes may be "parallel or radial, curved or straight, and may be formed by rows of repetitive elements such as polka

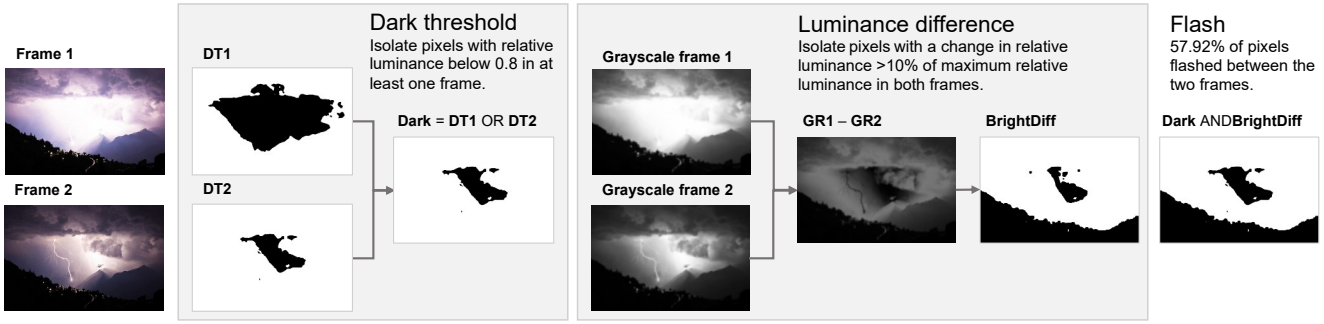

Figure 2: Dangerous flashes between two adjacent frames in a GIF are detected by isolating pixels that have relative luminance below 0.8 in at least one frame (Dark threshold) and pixels that change in relative luminance by more than 10% of the maximum relative luminance across both frames (Luminance difference). If more than 25% of pixels meet both criteria, a light-dark or dark-light transition has occurred. A pair of light-dark and dark-light transitions together create a flash.

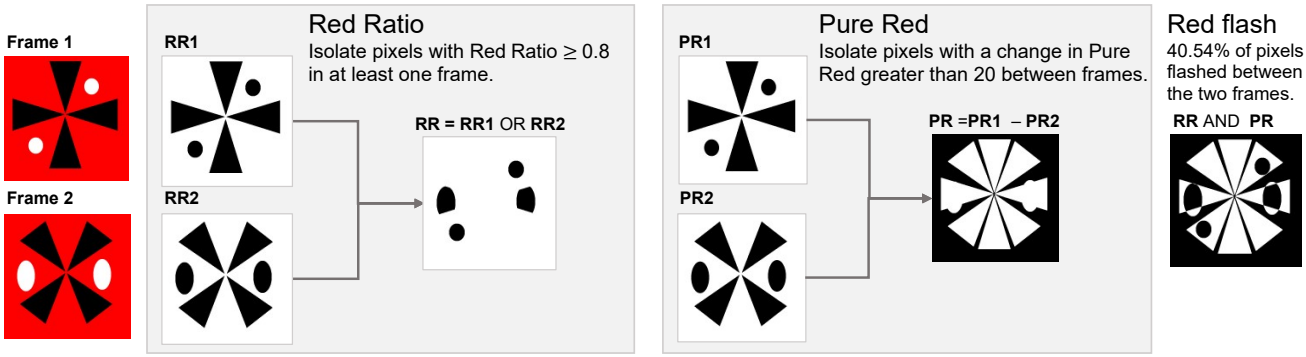

Figure 3: Dangerous red transitions between two adjacent frames in a GIF are detected by isolating pixels with Red Ratio (Equation 1) greater than 0.8 in at least one frame and pixels with a change in Pure Red (Equation 2) greater than 20 between frames. If more than 25% of pixels meet both criteria, a dangerous red transition has occurred.

dots". Wilkins et al. define a stripe in terms of a luminance difference, noting that "a luminance difference of <3 cd/m^2 is likely to affect <15% of patients" and can be used as a general threshold for characterizing a "safe" stripe. The final characteristic proposed by Wilkins et al. in defining a harmful striped pattern is the luminance of the brightest stripe in the pattern; the authors recommend that stripes with luminance greater than 50cd/m^2 be considered dangerous if all the other thresholds are exceeded. The area covered by stimuli is less important for repeated patterns than flashes and red transitions; Wilkins et al. note that "the proportion of patients affected by five-stripe pairs is similar for patterns that occupy the entire screen and those that occupy only a quarter of the screen" and conclude that the five-stripe limitation provides "adequate protection without specifying pattern size".

We use two approaches for identifying dangerous repeated patterns: one for patterns created by linear stripes (Figure 4) and one for patterns constructed from repeated discrete elements (Figure 5). Both approaches begin by extracting relative luminance for each pixel in the frame, producing a grayscale image. We threshold to isolate pixels with relative luminance > 50 cd/m^2 (i.e., potential bright stripes) and relative luminance < 47 cd/m^2 (i.e., potential dark stripes). Once we have identified areas of the image with the potential to contain harmful luminance differences, we need to locate the individual shapes that form the pattern, if one is present. To locate straight line segments, we first isolate areas where luminance changes sharply with Canny edge detection [12] and then apply the probabilistic Hough line transform [28] to get a list of lines present in the image that could outline the edges of individual stripes. Line segments are matched up with their nearest neighbors. Paired line segments with opposing luminance values are counted as stripe pairs. We locate discrete elements with sharp luminance differences using OpenCV's BlobDetector [8]. Patterns created by many repeated elements, such as polka dots, are characterized by many similar elements positioned at regular repeated intervals. We detect repeated element patterns by grouping elements by size and by distance to the nearest same-size neighbor, coupled with the heuristic rule that the distance to the nearest neighbor should be less than 3 times the area of the element itself. If the area occupied by elements that fit within this constraint is greater than 25% of the total area and elements are arranged into more than five rows or columns, the GIF is considered dangerous.

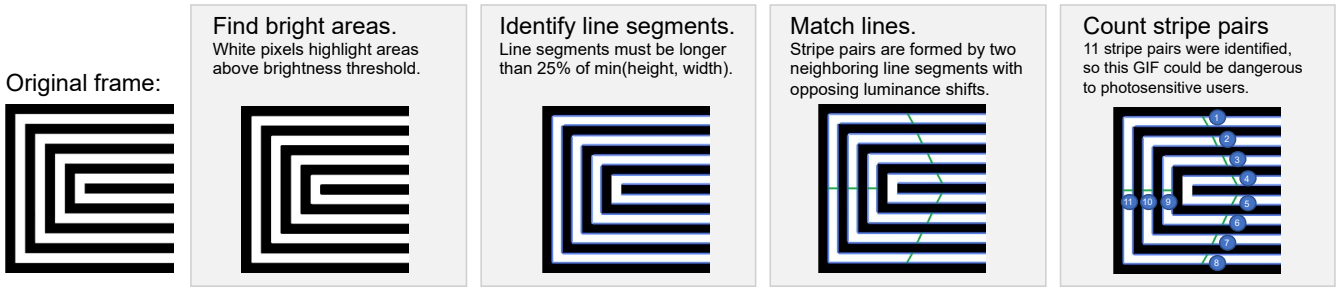

Figure 4: Dangerous patterns with linear stripes are detected by locating areas with luminance greater than 50 cd/m^2 (i.e., potential bright stripes) and areas with luminance below 47 cd/m^2 (i.e., potential dark stripes). We identify straight line segments with probabilistic Hough line transform and count the number of adjacent line segments with matching light-dark or dark-light luminance shifts. If more than five pairs covering more than 25% of the total area are found, the pattern is likely to be dangerous.

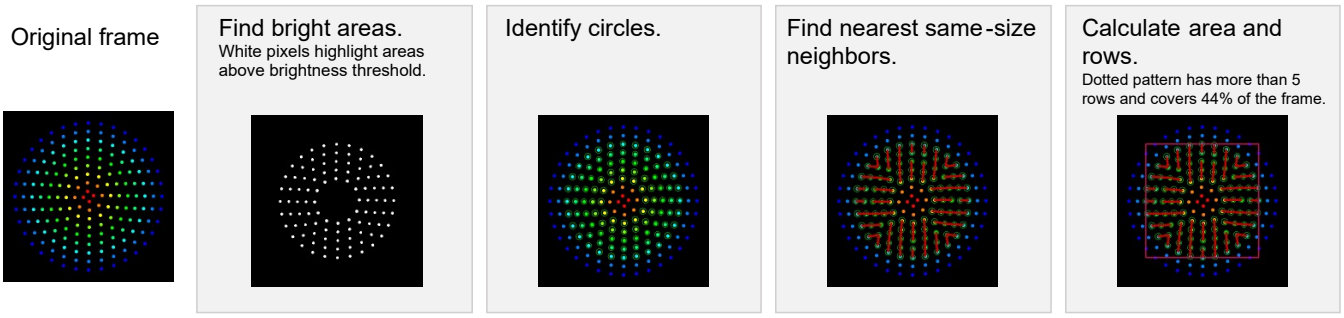

Figure 5: Dangerous dotted patterns are detected by first locating areas of high luminance and identifying circular shapes. Shapes are grouped by area and distance to nearest same-size neighbor. If the area covered by the grouped shapes exceeds 25% of the total area and contains more than five rows, the pattern is likely to be dangerous.

4.3 Mitigating dangerous content

Interview participants in Section 3 emphasized the importance of being able to override the system's decision and safely view content labelled as dangerous. In our system users have several options once a GIF is determined to be dangerous. A message is displayed summarizing which photosensitive risk factors were detected, along with options for mitigation strategies. The default is an outright block on the dangerous content, but users have the option to apply filters to view potentially dangerous content more safely (Figure 6). Flashes and patterns become dangerous when there are sharp changes in luminance between frames (flashes) or between elements within a frame (repeated patterns). By reducing the difference in luminance, low-contrast filters can make GIFs with dangerous flashes and repeated patterns safer to view. Similarly, grayscale and low-saturation filters can be used to make GIFs with saturated red transitions safer to view by removing color or lowering saturation values.

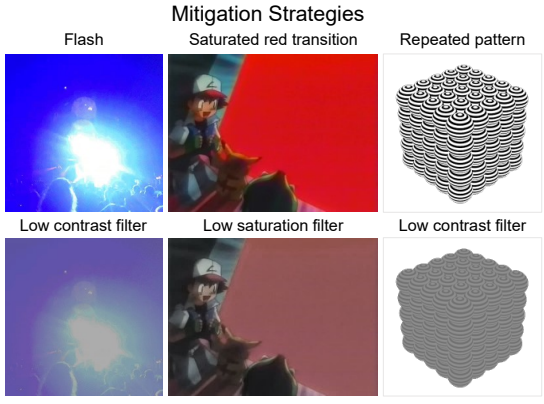

Figure 6: Users can choose to view dangerous content safely through low-contrast and low-saturation filters.

4.4 Implementation

The prototype risk detection system combines a browser extension and a backend server to identify and test animated images. The server is implemented in Python using OpenCV [8] and ImageIO [40] to analyze GIFs frame-by-frame. Upon loading a new webpage, the browser extension looks for all elements in the DOM with an <img> or <video> tag, stores their URLs for later analysis, and temporarily blocks content from view. If an alt-text description is provided, it is inserted into the placeholder to provide a summary of the content while it is blocked.

5 EVALUATION

The accuracy of PhotosensitivityPal was evaluated on three datasets: simulated, randomized, and potentially dangerous. The simulated dataset allows us to test performance on generated GIFs with a known ground truth value, while the randomized and potentially dangerous datasets allow us to assess performance on GIFs collected "in the wild". Two existing risk detection systems (PEAT and EpilepsyBlocker) were also tested on the same datasets to find potential inconsistencies among the three systems and to see if the systems were able to reach a consensus on the randomized and potentially dangerous datasets, where the ground truth is not known. Because the datasets include many potentially seizure-inducing GIFs that could be used in future malicious attacks, we will not post the datasets publicly. However, the datasets will be made available in a protected-access repository and researchers and developers will be able to apply individually for access.

5.1 Simulated dataset

The first dataset we use to evaluate the three risk detection systems includes GIFs simulated to include flashes, saturated red transitions, and repeated patterns. By testing the systems on simulated GIFs with a known ground truth, we assess how well they can identify sequences that are known to exceed thresholds for seizure-inducing content established in empirical studies[11, 19, 46].

5.1.1 Generating flashes. Flashing GIFs were simulated according to flash frequency (3 and 5 flashes per second), area of flash relative to total frame area (10, 25, and 50%), relative luminance of darker color involved in the flash (0, 100, or 210 pixel values), and the difference in relative luminance between colors involved in the flash (20, 50, and 100 pixel values). Figure 7 summarizes the features used to simulate flashes and the performance of EpilepsyBlocker, PEAT, and PhotosensitivityPal on each GIF. PEAT labelled 45 out of 54 GIFs correctly. Of the nine GIFs mislabelled by PEAT, six were false positives and three were false negatives. The prototype system labelled 52 GIFs correctly and produced four false positives. EpilepsyBlocker was unable to produce a result for any of the simulated flashing GIFs. For performance metrics on the simulated flashing GIFs dataset, see Table 2.

5.1.2 Generating red transitions. GIFs with transitions to and from saturated red (i.e., red flashes) were simulated according to flash frequency (3 and 5 flashes per second), area of flash (10, 25, and 50% of total frame area), Red Ratio (0.5 and 1.0, see Equation 1), and Pure Red (10 and 30, see Equation 2). Figure 8 summarizes the features used to simulate saturated red transitions and the performance of EpilepsyBlocker, PEAT, and PhotosensitivityPal on each GIF. PEAT and the prototype system both performed better on the simulated GIFs with saturated red transitions than the other two risk factors (Table 2). PEAT labelled all simulated GIFs with red flashes correctly except for one GIF, which had red flashes that took up only 10% of the total area but was incorrectly labelled as dangerous. The prototype system labelled all red flash simulated GIFs correctly. EpilepsyBlocker was unable to produce a result for any of the red flash simulated GIFs.

5.1.3 Generating repeated patterns. GIFs with repeated patterns were simulated according to the following features (Figure 9): shape (linear stripes, radial stripes, and dotted lines), movement (stationary or animated), stripe count (5 and 7 stripes), luminance of light stripe (40 and 60 cd/m^2), and difference in luminance between light and dark stripes (3, 10, and 40 cd/m^2). PEAT did not accurately detect any of the dangerous repeated pattern GIFs, while the prototype system was able to accurately label all simulated pattern GIFs except for two combination of features: 9 stationary radial stripes with bright stripe luminance of 60 cd/m^2 and luminance differences of 10 and 40 cd/m^2. EpilepsyBlocker was unable to produce a result for any of the simulated GIFs with repeated patterns. For performance metrics on the simulated repeated patterns dataset, see Table 2.

Flash			
Measure	Formula	PEAT	PhotosensitivityPal
Accuracy	(TP + TN) / Total	0.83	0.94
Recall	TP / (TP + FN)	0.67	1.0
Precision	TP / (TP + FP)	0.5	0.75
Miss-rate	FN / (TP + FN)	0.33	0
Saturated red			
Measure	Formula	PEAT	PhotosensitivityPal
Accuracy	(TP + TN) / Total	0.96	1.0
Recall	TP / (TP + FN)	1.0	1.0
Precision	TP / (TP + FP)	0.67	1.0
Miss-rate	FN / (TP + FN)	0	0
Repeated patterns			
Measure	Formula	PEAT	PhotosensitivityPal
Accuracy	(TP + TN) / Total	0.83	0.97
Recall	TP / (TP + FN)	0	0.83
Precision	TP / (TP + FP)	-	1.0
Miss-rate	FN / (TP + FN)	1.0	0.17

Table 2: Performance metrics were calculated for PEAT and PhotosensitivityPal when analyzing simulated GIFs with flashes, saturated red, and repeated patterns. Higher accuracy, recall, and precision is better, while a lower miss-rate is preferred. Measures could not be calculated for EpilepsyBlocker because the system was not able to analyze any of the simulated GIFs.

5.1.4 Performance metrics. We calculated four metrics to compare the performance of the three risk detection systems on simulated GIFs: accuracy, recall, precision, and miss-rate (Table 2). True positive (TP) and true negative (TN) indicate that a dangerous GIF was labelled as dangerous and a safe GIF was labelled as safe, respectively. A false positive (FP) means that a safe GIF was mistakenly labelled dangerous and a false negative means that a dangerous GIF was mistakenly labelled safe. *Accuracy* measures the overall number of correct labels, *recall* measures the number of correctly labelled dangerous GIFs relative to the number of truly dangerous GIFs, and *precision* measures the number of correctly labelled dangerous GIFs relative to the total GIFs labelled dangerous by each system. *Miss-rate*, or false positive rate, measures the number of GIFs mistakenly labelled safe relative to the total number of GIFs labelled safe. Higher accuracy, recall, and precision are desirable,

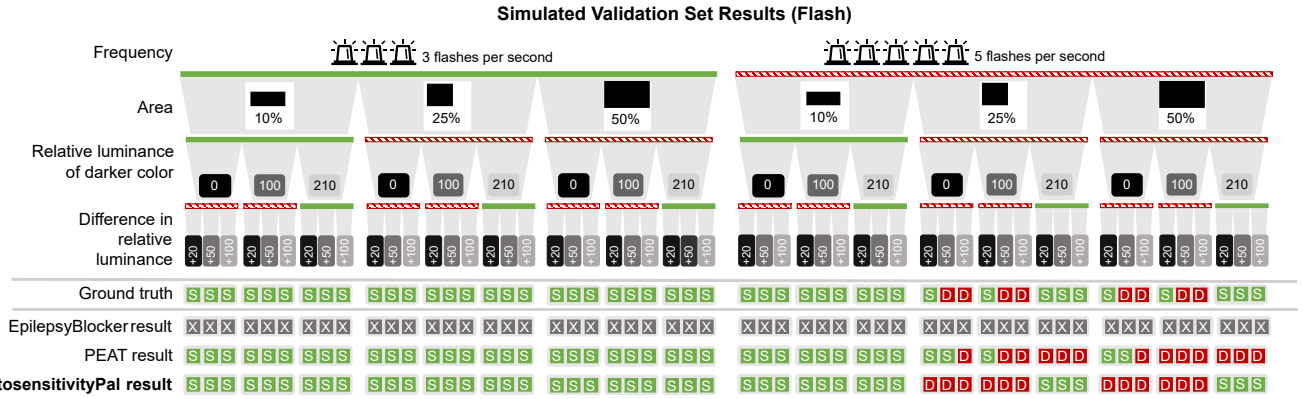

Figure 7: Flashing GIFs were generated according to four characteristics that determine seizure-inducing potential: flash frequency (three or five flashes per second), flash area (10%, 25%, or 50%), the relative luminance of the darker color involved in the flash (0, 100, or 210 pixel values), and the difference in relative luminance between colors involved in the flash (20, 50, and 100 pixel values). Ground truth labels and results from PEAT and prototype systems are indicated with red (dangerous) squares D and green (safe) squares S. Gray squares X indicate that the system could not analyze the GIF.

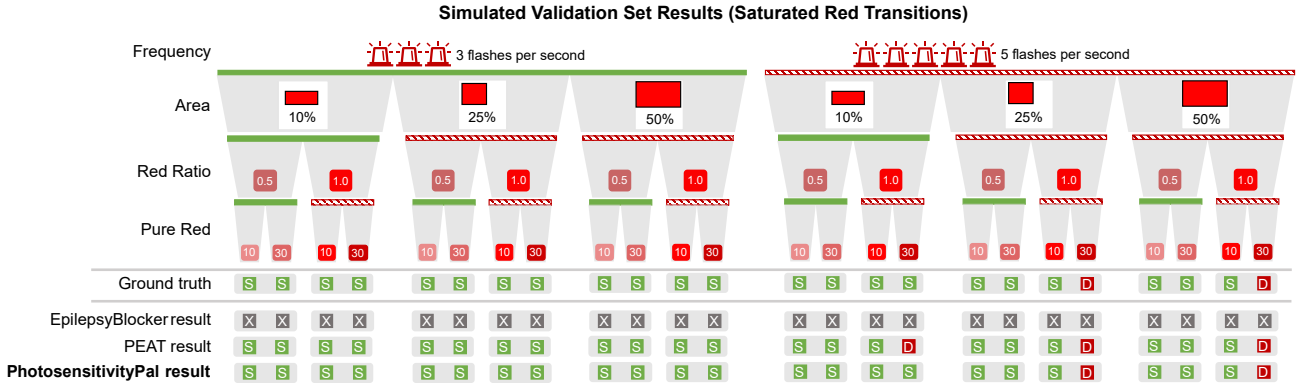

Figure 8: GIFs with saturated red transitions were generated according to four characteristics that determine seizure-inducing potential: red flash frequency (three or five red flashes per second), red flash area (10, 25, or 50%), red ratio (0.5, 1.0), and difference in pure red between states (10, 30). Ground truth labels and results from PEAT and prototype systems are indicated with red (dangerous) squares D and green (safe) squares S. Gray squares X indicate that the system could not analyze the GIF.

while a lower miss-rate is preferred. In this application, the consequences of false negatives are very serious, so a low miss-rate is particularly important.

5.1.5 Summary. We compare the performance of EpilepsyBlocker, PEAT, and PhotosensitivityPal in Table 2. Because the EpilepsyBlocker extension was created to detect GIFs on webpages with a prescribed structure (i.e., only searching for GIFs nested within a specific element), the system was unable to detect and analyze any of the simulated GIFs, producing an error message instead of a valid risk assessment.PhotosensitivityPal had higher accuracy, recall, and precision than PEAT for all three categories. PhotosensitivityPal also had a lower miss-rate than PEAT in all simulations, including a miss-rate of 0 for flashes and saturated red (i.e., no dangerous GIFs were mistakenly labelled as safe). PEAT performed best on the saturated red dataset and worst on repeated patterns.

5.2 Randomized dataset

The simulated dataset evaluation established that PhotosensitivityPal is able to detect dangerous content that precisely reflects established thresholds. To evaluate the system's effectiveness "in the wild", we collected 200 random GIFs from Twitter and Tenor (100 GIFs from each source). Each GIF was tested with PhotosensitivityPal, PEAT, and EpilepsyBlocker. Because PEAT can only analyze .AVI files, all .GIF and .MP4 files were converted to .AVI prior to testing using FFmpeg [6], an open source image processing framework. Image resolution and colors were preserved as much as possible during conversion through manual examination of each converted file. The prototype system and EpilepsyBlocker were both tested on the original .GIF or .MP4 files. Figure 10 (right)

[6] https://ffmpeg.org/

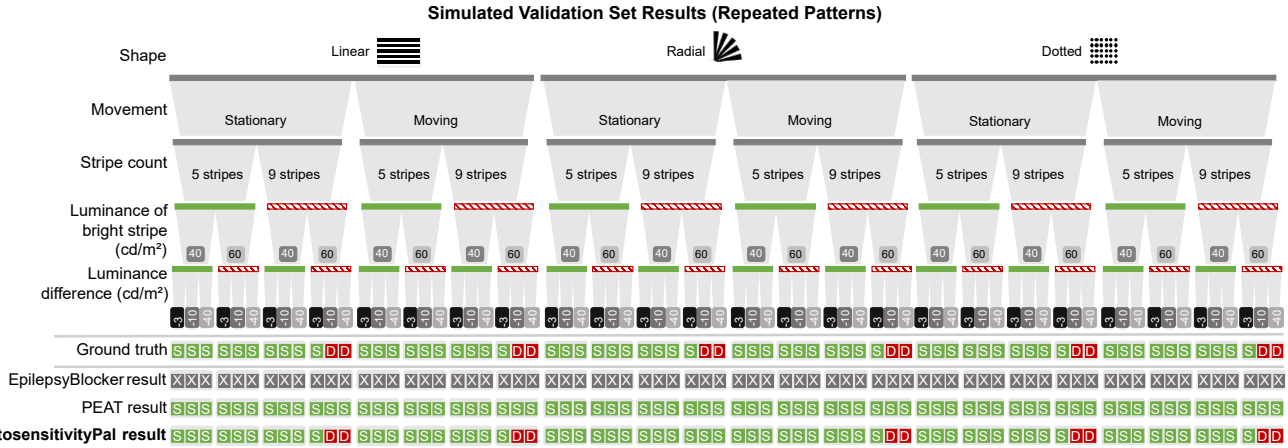

Figure 9: GIFs with dangerous repeated patterns were generated according to pattern shape (linear, radial or dotted), movement (stationary or moving), stripe count (5 or 9 stripes), luminance of brightest stripe (40 or 60 candelas/m^2), and luminance difference between light and dark stripes (3, 10, or 40 candelas/m^2). Ground truth labels and results from PEAT and prototype systems are indicated with red (dangerous) squares D and green (safe) squares S. Gray squares X indicate that the system could not analyze the GIF.

shows examples of two GIFs (one safe and one dangerous) in our randomized dataset.

Almost all of the GIFs included in the randomized dataset were found to be safe by both PEAT and the prototype system (183, 91.5%). Ten GIFs (5%) were found to be dangerous by both systems. Of the seven GIFs for which PEAT and the prototype did not reach consensus, two caused PEAT to crash and return no response (both were labelled safe by the prototype). Two GIFs were labelled safe by PEAT and dangerous by the prototype. Although PEAT returns a binary safe/dangerous result, borderline items that are labelled safe are given an additional label of "caution (pass)", indicating that they are close to being labelled dangerous. Both of the GIFs labelled safe by PEAT and dangerous by the prototype received "caution (pass)" labels. Three GIFs were labelled dangerous by PEAT and safe by the prototype. EpilepsyBlocker was unable to produce a result 181 out of the 200 GIFs in the randomized dataset. Out of the 100 GIFs collected from Tenor, 81 produced a 403 error and did not progress past an "Analyzing GIF..." message. The remaining 19 GIFs from Tenor were all labelled safe by EpilepsyBlocker, including one GIF that was labelled dangerous by PEAT and the prototype. None of the 100 GIFs collected from Twitter were acknowledged as GIFs or tested for photosensitive risk factors by the EpilepsyBlocker extension.

Summary: PhotosensitivityPal and PEAT agreed on the majority of the GIFs in the randomized dataset (91.5%), indicating that our system can perform equally well as an established risk detection system on GIFs found "in the wild" without a known ground truth.

5.3 Potentially dangerous dataset

The simulated dataset allowed us to test the performance on ground truth examples, while the randomized dataset ensured that the system agreed with an accepted existing implementation, PEAT, when testing random GIFs across the Internet, most of which were innocuous. These validation sets give little information on how the system will perform on dangerous GIFs like those used in the malicious attacks described in Section 1. Such GIFs are rare online, and often are not enshrined in GIF repositories such as Tenor or GIPHY. Malicious GIFs appear on Twitter or Reddit during attacks, while accidentally dangerous GIFs are occasionally posted on the same social networks inadvertently, but these instances are rare and often not preserved, as dangerous content is deleted or accounts are banned. As a result, it is difficult to reliably collect dangerous GIFs using random social media APIs. The people who create, collect, and deploy malicious GIFs congregate on less popular forums, such as 4chan and SomethingAwful, where threads celebrating recent malicious attacks overflow with strobing and flashing GIFs, often accompanied by derogatory language targeting people with photosensitive epilepsy. We manually explored such sites in an effort to construct a dataset of potentially dangerous seizure-inducing GIFs. Two authors independently began the search for potentially dangerous GIFs the same way a hypothetical malicious attacker might seek out harmful content: by searching for "epilepsy gif" and "seizure gif" on Google Images and following potentially dangerous content to its source to find similar items. GIFs were added to the potentially dangerous validation dataset if they met the following two inclusion criteria:

(1) The GIF includes flashes, transitions to or from saturated red, or repeated patterns that appear to exceed safe thresholds (as defined by Harding et al., 2005) [25].
(2) References to "seizure" or "epilepsy" appear in the GIF itself or in accompanying text.

These criteria allow the potentially dangerous dataset to include GIFs from both accidental attacks (where users try to alert others about posts containing harmful content) and malicious attacks (where users discuss and share seizure-inducing GIFs). Figure 11

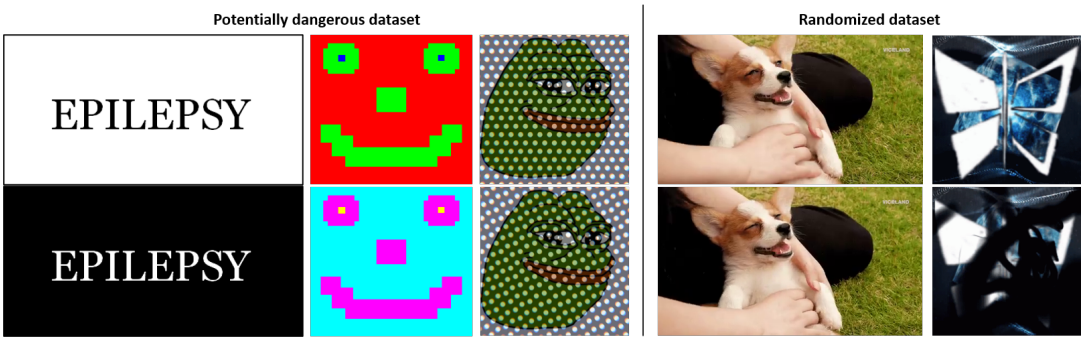

Figure 10: Examples of GIFs from the potentially dangerous (left) and randomized (right) validation sets.

contains an example of each type of attack identified during manual collection of potentially dangerous GIFs. Manual collection of course introduces the potential for sampling bias, however the nature of these GIFs inhibits automated collection. Over the course of our collection process, we found harmful GIFs on personal websites, in hidden directories, and in comment threads randomly interspersed between harmless memes, all of which would be difficult or impossible to discover with automated collection techniques alone. Although it is possible that not all of the GIFs collected in this manner will be genuinely harmful to photosensitive users, the potentially dangerous dataset comes closer to representing the actual GIFs used to harass individuals with photosensitive epilepsy in past incidents than the simulated dataset or the generally safe randomized dataset. We identified 137 potentially dangerous GIFs that met our inclusion criteria. All 137 GIFs were tested with the prototype system, PEAT, and EpilepsyBlocker. Once again, GIFs with .GIF and .MP4 extensions were converted to .AVI to be tested by PEAT.

PhotosensitivityPal labelled 121 out of the 137 GIFs (88%) as dangerous. EpilepsyBlocker was unable to analyze 118 of 137 GIFs and labelled all of the remaining 19 GIFs as safe. PEAT struggled to detect the most egregious examples of strobing GIFs in the potentially dangerous dataset because PEAT does not account for looping videos. In many of the GIFs in the potentially dangerous dataset, creators took advantage of GIF looping behavior by using only two frames to produce a rapid flicker effect. Such GIFs appear to be harmless because of their short duration, but become hazardous when looped indefinitely. When analyzing the original GIFs collected in the potentially dangerous dataset, PEAT and PhotosensitivityPal agreed on only 38 out of 137 items (27%). We summarize PEAT and PhotosensitivityPal's performance in Table 3. To account for looping behavior, we constructed extended versions of each GIF in the dataset by duplicating the GIF's frames until we reached a total duration of two seconds. These extended GIFs were then analyzed with PEAT. PEAT detected far more dangerous sequences when analyzing the extended GIFs: 91 GIFs that were labelled safe by PEAT originally were identified as dangerous when looped for two seconds. We summarize PEAT and PhotosensitivityPal's performance on the extended GIFs in Table 4. Of the eight GIFs labelled safe by PEAT even when looped but dangerous by PhotosensitivityPal, four included repeated patterns, a photosensitive risk factor that PEAT does not explicitly look for in videos. PEAT labelled 7 looped GIFs as dangerous that were found to be safe by PhotosensitivityPal, indicating some differences in sensitivity between the two systems despite their agreement on the majority of the GIFs in the potentially dangerous dataset.

Summary: PhotosensitivityPal and PEAT agreed on labels for only 38 out of 137 GIFs in the original potentially dangerous dataset. When the GIFs were extended to replicate looping behavior, the number of GIFs with consensus between the two systems rose to 117 (85%). This highlights the importance of accounting for looping behavior in GIFs, particularly when analyzing malicious GIFs that repeat a small number of frames many times to produce rapid flickering and strobing effects.

6 SOCIAL MEDIA STUDY

Social media platforms are used to communicate with friends, learn about current events, and even further careers, but stories like Eichenwald's (Section 1) indicate that these mundane activities can be dangerous for people with photosensitivity. As social media users becomes more reliant on visual content [31, 33], the likelihood of encountering harmful visual stimuli rises. Until now no research has been conducted to investigate the prevalence of GIFs and videos with potentially seizure-inducing strobes and patterns on popular websites. To answer this open question, we tested GIFs posted on social media sites (Twitter and Tumblr) and GIF repository sites underlying popular messaging services like Facebook Messenger and Slack (GIPHY and Tenor GIF Keyboard) for seizure-inducing content with PhotosensitivityPal. These four sources were chosen because they are popular, each with hundreds of thousands of active users[7] [22, 39, 41], and because they represent different levels of responsiveness towards the issue of accessibility for people with photosensitivity: while Twitter and GIPHY have introduced a handful of platform-driven solutions to limit the spread of dangerous seizure-inducing GIFs (Section 2.2), Tumblr and Tenor have not taken public steps to improve safety for photosensitive users on their platforms. Although Tumblr has fewer users than huge social media sites such as Twitter, Facebook, and Instagram, the platform has a unique relationship with GIFs. Many attribute the GIF's sudden resurgence in the mid-2010s after decades of declining popularity to adoption by Tumblr users [4, 26], making it a natural fit for studying the prevalence of seizure-inducing GIFs on

[7] https://www.tumblr.com/about

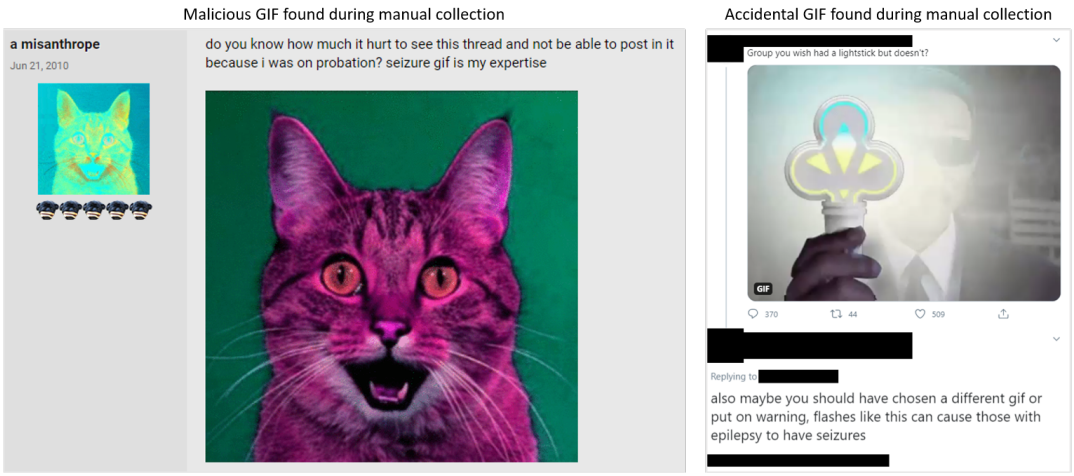

Figure 11: Examples of malicious (left) and accidental (right) dangerous GIFs discovered during manual collection of potentially dangerous GIFs online.

Potentially dangerous dataset (original, not looped)

		PEAT			
		Safe	Dangerous	No response	Total
PhotosensitivityPal	Safe	15	**1**	0	16
	Dangerous	**93**	23	5	121
	Total	108	24	5	137

Table 3: PEAT and PhotosensitivityPal reached consensus on only 38 out of 137 GIFs collected in the potentially dangerous dataset when GIFs were analyzed in their original form.

Potentially dangerous dataset (looped)

		PEAT			
		Safe	Dangerous	No response	Total
PhotosensitivityPal	Safe	9	**7**	0	16
	Dangerous	**8**	108	5	121
	Total	17	115	5	137

Table 4: When GIFs were extended to replicate looping behavior, PEAT and PhotosensitivityPal agreed on 117 out of 137 GIFs. PEAT found 91 GIFs that were previously labelled as safe to be dangerous when accounting for looping behavior.

social media. The social media study aims to answer three research questions:

RQ1: What proportion of GIFs posted online possess potentially seizure-inducing photosensitive risk factors?

RQ2: Does the proportion of dangerous GIFs vary significantly across the four sources included in this study (Twitter, Tumblr, Tenor, and GIPHY)?

RQ3: What proportion of GIFs posted online contain each of the three primary photosensitive risk factors (flashes, red transitions, and patterns)?

6.1 Methodology

To answer RQ1, we construct a 99% confidence interval for the π_{overall}, the overall proportion of GIFs containing seizure-inducing sequences. We use the more conservative Clopper-Pearson method [13], which is based on the exact binomial distribution rather than a Normal approximation (e.g., the Wald method) [17]. We determined through a power analysis that the sample must include at least 1168 GIFs to achieve 80% power with a 99% confidence interval. RQ2 involves looking for differences in the proportion of dangerous GIFs across four difference sources. We use the Chi-squared goodness of fit test on the null hypothesis H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4$ and the alternative hypothesis H_A: at least one π_i differs, where π_1, π_2, π_3, and π_4 are the proportion of GIFs with dangerous content on Twitter, Tumblr, Tenor, and GIPHY, respectively. We determined through a power analysis that to detect a small effect size with 80% power at $\alpha = 0.01$ our sample size must be at least 1546. To answer RQ3 we need to compare the proportion of dangerous GIFs across the three risk factors (flashes, red transitions, and patterns). We construct three 99% confidence interval with the Clopper-Pearson method,

one for each risk factor. Because the joint confidence intervals will be calculated simultaneously, we use the Bonferroni correction for multiple comparisons, leading to three individual intervals with $1-\frac{\alpha}{3} = 1-0.0033 = 0.9967 = 99.67\%$ confidence and an overall joint confidence of 99%. Power analysis tells us that to ensure 80% power at 99.67% confidence, we must collect at least 1429 GIFs. Because the marginal cost of testing additional GIFs is minimal in this study and greater sample sizes can lead to greater power, we increased the total sample size to 2000, with 500 GIFs collected from each source. Study preregistration can be found at https://osf.io/5a3dy/. Metadata and risk assessments for all 2000 GIFs can be found in Supplemental Material.

6.2 Results

RQ1: Of the 2000 GIFs tested for seizure-inducing content, 163 were labelled dangerous by PhotosensitivityPal (8.15%). The 99% confidence interval for the overall proportion of GIFs containing potentially seizure-inducing sequences is 6.65% to 9.85%.

RQ2: Tenor produced the most dangerous GIFs (50, 10%), followed by Tumblr (42, 8.4%), Twitter (38, 7.6%), and GIPHY (33, 6.6%). Although small differences were observed in the number of dangerous GIFs detected from each source, we did not find sufficient evidence supporting a difference in the proportion of dangerous GIFs across Twitter, Tumblr, Tenor, and GIPHY to reject the null hypthesis. The Chi-squared goodness of fit test produced a test statistic of 3.798 (df = 3, p-value = 0.2842).

RQ3: To answer RQ3, we constructed three simultaneous confidence intervals for the proportion of GIFs containing dangerous flashes, red transitions, and repeated patterns. Flashes were the most common photosensitive risk factor in our sample, occurring in 94 GIFs (4.70%). The 99% confidence interval for flash prevalence was 3.35% to 6.38%. Repeated patterns appeared in 67 GIFs (3.35%) with a 99% confidence interval of 2.22% to 4.80%. With only 12 instances out of 2000 GIFs (0.6%), red transitions were the least common photosensitive risk factor. The 99% confidence interval for red transition prevalence is 0.12% to 1.36%. Several GIFs collected in the study possessed more than one photosensitive risk factor; dangerous flashes and repeated patterns appeared jointly in six GIFs, while dangerous flashes and red transitions appeared in four.

In summary, our social media study found at least one dangerous photosensitive risk factor in 8.15% of GIFs collected randomly online. Dangerous flashes were most common, occurring at more than 7 times the rate of dangerous saturated red transitions. This result is close to the average of estimates given by participants in our user interviews when asked to estimate the proportion of triggering GIFs and videos they encounter online (Section 3), although individual estimates given in interviews ranged from 1 to 30%. We did not find a difference in the likelihood of encountering triggering GIFs across the four platforms. Even though sites like Twitter and GIPHY are aware of the dangers of seizure-inducing GIFs and have implemented a handful of platform-driven solutions to limit exposure to triggering content, GIFs posted to their platforms are equally likely to contain sequences harmful to users with photosensitivity.

7 DISCUSSION

7.1 Implications

The results of our study have implications for online accessibility guidelines (i.e., the WCAG, see Section 2.1), people with photosensitivity, and HCI researchers. The WCAG currently discourages flashes and red transitions but does not explicitly discourage repeated patterns, despite the broad consensus that repeated patterns are a common photosensitive risk factor. Our social media study demonstrates that dangerous repeated patterns are present in online content, but this risk factor is not explicitly discouraged in the WCAG. We recommend that the WCAG update all guidelines related to PSE to include dangerous repeated patterns as defined by Wilkins et al. [46]. Additionally, we recommend the WCAG explicitly state that GIFs with a duration less than one second and fewer than three flashes per second can still cause seizures when looped. Many of the dangerous GIFs collected for our evaluation (Section 5) took advantage of rapid looping to create strobing effects that would not be flagged by the current WCAG guidelines. Widespread implementation of protection systems would have a profound effect on the well-being of photosensitive individuals, both in terms of independence (i.e., no longer needing to forgo certain platforms or ask friends or family to screen content for them) and interdependence (i.e., depending on social media platforms to protect them in the same way the average user depends on Twitter or Facebook to protect them from viewing disturbing content). Finally, our work has implications for future research in HCI. Our social media survey demonstrates how inaccessible much on the Internet remains to people with PSE, despite a recent increased interest in accessibility within tech. We hope that our study encourages further research into how the web can be made safer for people with less well-known disabilities, such as photosensitivity. We also recommend that HCI and accessibility researchers keep photosensitivity in mind when designing virtual experiments. For example, a researcher can make their study more accessible for those with photosensitivity by allowing participants to conduct interviews on the phone rather than through video calls or by allowing for screen breaks during experiments.

7.2 The potential for platform-driven solutions

This paper has largely focused on the design, development, and evaluation of consumer-driven systems for protecting users with photosensitivity online. While consumer-driven systems have the advantage of being platform-agnostic and hypothetically able to protect users on any website they visit, they require the user to actively choose to enable the consumer-driven protection on their device. People who are unaware that they have photosensitivity are still vulnerable to dangerous content, because they will likely not seek out consumer-driven protection. If designed and implemented correctly, platform-driven solutions have the potential to dramatically improve online safety for people with photosensitive epilepsy by protecting people who do not use consumer-driven protection. Our user interviews indicated that current platform-driven solutions are lacking: participants felt that autoplay disabling features were not reliable and interfered too much with their social media experience. GIPHY and Twitter have introduced keyword-based

restrictions on GIF searches, banning terms (e.g., "epilepsy", "photosensitivity", "seizure") , but dangerous GIFs can still be found on both platforms by searching for slightly different terms, such as "strobe", "flashing light", or "optical illusion". Banning these alternative keywords is not feasible because they could also be used for harmless searches by other users. To overcome this problem, platforms could use an automatic detection system to label GIFs as safe or potentially dangerous for users with photosensitivity. If this information were added to GIF metadata in Tenor or GIPHY's repositories, then warnings about a GIF's content could follow that GIF whenever it is shared in a message or on social media. Some potential platform-driven solutions more straightforward: For example, two interview participants described a particularly triggering video filter on TikTok that causes the background color to quickly shift between several highly saturated shades. Participants agreed that TikTok should allow users to automatically block posts that use triggering filters. By listening to users with photosensitivity and allowing them to block triggering content through platform-driven systems, websites and apps can improve accessibility and safety for many users simulataneously. The effectiveness of platform-driven solutions is limited by each platform's awareness of the dangers of seizure-inducing content and their willingness to take action. While social media platforms and GIF repositories cannot be held solely responsible for accidental and malicious exposure to seizure-inducing GIFs, their roles as arbiters of digital content give them a unique opportunity to make online platforms safer for people with photosensitivity through platform-driven protection systems.

7.3 Future research directions

Updated empirical studies of photosensitive risk factors Although empirical studies have identified specific risk thresholds (i.e., flash rate, area of the flash, luminance difference between elements in a repeated pattern) for people viewing triggering content on television sets [19, 46], much remains to be understood about the relationship between photosensitivity and other devices, such as laptops and smartphones. For example, there is already at least one documented case of an individual accidentally triggering a seizure while taking a selfie with a smartphone in a darkened room [9]. Some researchers have suggested that differences in sensitivity caused by holding mobile devices closer to the face are offset by their smaller screen size [46], but this claim has not been verified through empirical study. Without further study on photosensitive triggers on smaller screens, the effectiveness of rule-based classification systems such as PhotosensitivityPal could be limited on laptops and mobile devices.

Detecting photosensitive risk factors with machine learning The rule-based approach PhotosensitivityPal uses to detect triggering sequences is well-suited for identifying flashes and red transitions because there are clear guidelines defining the characteristics of a dangerous flash or red transition. Machine learning could be a better approach for identifying repeated patterns because there are far more variations of potentially dangerous repeated patterns. Our dataset of simulated and collected GIFs are the first step towards building a training database for machine learning algorithms.

Classifying accidental and malicious attacks Distinguishing between accidental and malicious attacks requires analyzing the full context of how a GIF or video was posted online. For example, a GIF with flashes that is posted in an innocuous tweet may be accidental, but the same GIF included in a tweet that disparages people with epilepsy and tags the Epilepsy Foundation should be considered malicious. The potentially dangerous GIFs identified during the social media study (Section 6) were not categorized as accidental or malicious because there is currently no automated way of determining malicious intent in seizure-inducing online content. Automating the process of distinguishing between accidental and malicious seizure-inducing GIFs could lead to a better understanding of how dangerous content migrates through the digital ecosystem and remains a promising area for future research.

8 CONCLUSION

Recent incidents of malicious attackers sending intentionally strobing and flashing GIFs to people with photosensitive epilepsy have demonstrated that social media platforms can be dangerous and inaccessible for users with photosensitivity. In this paper, we contribute a novel framework characterizing systems for defending against seizure-inducing content as creator-driven, platform-driven, or consumer-driven. The Internet's current reliance on creator-driven protections has left users with photosensitivity vulnerable to seizure-inducing GIFs created by people who accidentally or intentionally evade creator-driven protections. Through a series of interviews with people with photosensitivity, we constructed design requirements for effective consumer-driven systems and developed PhotosensitivityPal, a prototype consumer-driven browser extension for detecting and mitigating seizure-inducing GIFs and videos. To address the current lack of standardized evaluation datasets for photosensitive risk detection systems, we contribute three datasets: 150 GIFs simulated to include dangerous flashes, red transitions, and patterns, 200 GIFs randomly collected from Twitter and Tenor GIF Keyboard, and 137 potentially dangerous GIFs manually collected from social media, forums, and personal websites. We found through our social media study that approximately 8% of GIFs posted on Twitter, Tumblr, Tenor, and GIPHY contain sequences that could cause seizures when viewed by someone with photosensitive epilepsy. While consumer-driven systems can help people with photosensitivity avoid encountering accidentally or intentionally triggering content, platform-driven systems have the potential to protect users who are not yet aware that they have photosensitivity as well as users who are not aware of consumer-driven systems. People with photosensitive epilepsy have historically been underrepresented in accessibility research, despite the serious and even deadly consequences of encountering seizure-inducing content online. The work described in this paper, including our user interviews and the design, development, and evaluation of PhotosensitivityPal, is the first step towards addressing this gap and improving safety and accessibility for users with photosensitive epilepsy on the Internet.

ACKNOWLEDGMENTS

We thank our participants for their contributions to this work. We would also like to thank Bruce Draper, Ross Beveridge, Magy Seif El-Nasr, Michail Schwab, and John Alexis Guerra-Gomez their support

and feedback. This research is supported in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1451070.

REFERENCES

[1] Amaia Aizpurua, Simon Harper, and Markel Vigo. Exploring the relationship between web accessibility and user experience. *International Journal of Human-Computer Studies*, 91:13–23, 2016.

[2] Jackie Aker. Epilepsy Foundation files criminal complaint and requests investigation in response to attacks on Twitter feed. *Epilepsy Foundation*, December 2019.

[3] Mohammad A Alzubaidi, Mwaffaq Otoom, and Abdel-Karim Al-Tamimi. Parallel scheme for real-time detection of photosensitive seizures. *Computers in biology and medicine*, 70:139–147, 2016.

[4] Saeideh Bakhshi, David A Shamma, Lyndon Kennedy, Yale Song, Paloma De Juan, and Joseph'Jofish' Kaye. Fast, cheap, and good: Why animated GIFs engage us. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 575–586, 2016.

[5] Andrei Barbu, Dalitso Banda, and Boris Katz. Deep video-to-video transformations for accessibility with an application to photosensitivity. *Pattern Recognition Letters*, 2019.

[6] Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, 2018.

[7] CD Binnie, J Emmett, P Gardiner, GFA Harding, D Harrison, and AJ Wilkins. Characterizing the flashing television images that precipitate seizures. *SMPTE journal*, 111(6-7):323–329, 2002.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] PM Brna and KG Gordon. "Selfie-epilepsy": A novel photosensitivity. *Seizure*, 47:5–8, 2017.

[10] Mattha Busby. Malicious tweets targeting epilepsy charity trigger seizures. *The Guardian*, May 2020.

[11] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, and Gregg Vanderheiden. Web Content Accessibility Guidelines (WCAG) 2.0. *WWW Consortium (W3C)*, 2008.

[12] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:679–698, 1986.

[13] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[14] Gregory Conti, Mustaque Ahamad, and John Stasko. Attacking information visualization system usability: overloading and deceiving the human. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 89–100, 2005.

[15] Paige Dawkins. Epilepsy society welcomes giphy's prompt action to reduce online risk to people with epilepsy. *Epilepsy Society*, September 2020.

[16] Vagner Figueredo de Santana, Rosimeire de Oliveira, Leonelo Dell Anhol Almeida, and Marcia Ito. Firefixia: An accessibility web browser customization toolbar for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–4, 2013.

[17] Keith Dunnigan. Confidence interval calculation for binomial proportions. In *MWSUG Conference, Indianapolis, IN*, 2008.

[18] Giuseppe Erba. Shedding light on photosensitivity, one of epilepsy's most complex conditions. *Epilepsy Foundation*, 2006.

[19] Robert S Fisher, Graham Harding, Giuseppe Erba, Gregory L Barkley, and Arnold Wilkins. Photic- and pattern-induced seizures: a review for the Epilepsy Foundation of America Working Group. *Epilepsia*, 46(9):1426–1441, 2005.

[20] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. Making GIFs accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10, 2020.

[21] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. Twitter A11y: A browser extension to make twitter images accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[22] Shannon Greenwood, Andrew Perrin, and Maeve Duggan. Social media update 2016. *Pew Research Center*, 11(2):1–18, 2016.

[23] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. Caption Crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[24] Stephanie Hackett, Bambang Parmanto, and Xiaoming Zeng. Accessibility of internet websites through time. In *Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility*, pages 32–39, 2003.

[25] Graham Harding, Arnold J Wilkins, Giuseppe Erba, Gregory L Barkley, and Robert S Fisher. Photic-and pattern-induced seizures: Expert consensus of the Epilepsy Foundation of America Working Group. *Epilepsia*, 46(9):1423–1425, 2005.

[26] Jialun "Aaron" Jiang, Casey Fiesler, and Jed R Brubaker. 'The Perfect One': Understanding communication practices and challenges with animated GIFs. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20, 2018.

[27] Cecilia Kang. A tweet to Kurt Eichenwald, a strobe, and a seizure. Now, an arrest. *The New York Times*, 2017.

[28] Nahum Kiryati, Yuval Eldar, and Alfred M Bruckstein. A probabilistic hough transform. *Pattern Recognition*, 24(4):303–316, 1991.

[29] Robert Kloster and Torstein Engelskjøn. Sudden unexpected death in epilepsy (SUDEP): a clinical perspective and a search for risk factors. *Journal of Neurology, Neurosurgery & Psychiatry*, 67(4):439–444, 1999.

[30] Rui Lopes, Daniel Gomes, and Luís Carriço. Web not for all: a large scale study of web accessibility. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, pages 1–4, 2010.

[31] Farhad Manjoo. Welcome to the post-text future. *The New York Times*, February 2018.

[32] Jennifer Mankoff, Holly Fait, and Tu Tran. Is your web page accessible? A comparative study of methods for assessing web page accessibility for the blind. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–50, 2005.

[33] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. "With most of it being pictures now, I rarely use it": Understanding Twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516, 2016.

[34] Anthony K Ngugi, Christian Bottomley, Immo Kleinschmidt, Josemir W Sander, and Charles R Newton. Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. *Epilepsia*, 51(5):883–890, 2010.

[35] Gustavo Paetzold and Lucia Specia. Anita: An intelligent text adaptation tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 79–83, 2016.

[36] Seong Je Park, Young Moo Kang, Hyung Rim Choi, Soon Goo Hong, and Yang Suk Kang. Development of image and color evaluation algorithm for the web accessibility evaluation tools. In *Asia-Pacific Conference on Computer Human Interaction*, pages 389–395. Springer, 2008.

[37] Kevin Poulsen. Hackers assault epilepsy patients via computer. *Wired*, March 2008.

[38] Christopher Power, André Freire, Helen Petrie, and David Swallow. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*, pages 433–442, 2012.

[39] Audrey Schomer. Google buys Tenor to power GIFs. *Business Insider*, March 2018.

[40] Steven Silvester, Anthony Tanbakuchi, Paul Müller, Juan Nunez-Iglesias, Mark Harfouche, Almar Klein, Matt McCormick, OrganicIrradiation, Arash Rai, Ariel Ladegaard, Antony Lee, Tim D. Smith, Ghislain Antony Vaillant, jackwalker64, Joel Nises, rreilink, Hugo van Kemenade, Chris Dusold, Felix Kohlgrüber, Ge Yang, Graham Inggs, Joe Singleton, Maximilian Schambach, Michael Hirsch, Miloš Komarčević, NiklasRosenstein, Po-Chuan Hsieh, Zulko, Chris Barnes, and Addison Elliott. imageio/imageio v0.9.0, July 2020.

[41] Shasta Smith. GIPHY selects Oracle Data Cloud to measure viewability across billions of GIFs. *Oracle Press Release*, September 2019.

[42] Nicola Swanborough. Epilepsy Society sees worst ever bullying attack on Twitter. *Epilepsy Society*, May 2020.

[43] Nicola Swanborough. Epilepsy Society welcomes Twitter's ban of GIF search terms. *Epilepsy Society*, July 2020.

[44] Takeo Takahashi and Yasuo Tsukahara. Pocket monster incident and low luminance visual stimuli: Special reference to deep red flicker stimulation. *Pediatrics International*, 40(6):631–637, 1998.

[45] @TwitterSupport. "We recently found a bug that lets you add multiple animated images to a tweet using animated PNG files. APNGs ignore our safeguards and can cause performance issues for the app and your device. Today we're fixing the bug which will no longer allow apngs to animate when tweeted.", December 2019. Retrieved 17 Sept 2020 from https://twitter.com/TwitterSupport/status/1209171739374014465.

[46] Arnold Wilkins, John Emmett, and Graham Harding. Characterizing the patterned images that precipitate seizures and optimizing guidelines to prevent them. *Epilepsia*, 46(8):1212–1218, 2005.