Human Perceptions of Sensitive Content in Photos

Yifang Li, Wyatt Troutman, Bart P. Knijnenburg and Kelly Caine
Clemson University
{yifang2,jwtrout,bartk,caine}@clemson.edu

Abstract

Before we can obfuscate portions of an image to enhance privacy, we must know what portions are considered sensitive. In this paper, we report results from a study aimed at identifying sensitive content in photos from a human-centered perspective. We collected sensitive photos and/or descriptions of sensitive photos from participants and asked them to identify which elements of the photo made each photo sensitive. Using this information, we propose an initial two-level taxonomy of sensitive content categories. This taxonomy may be useful to privacy researchers, online social network designers, policy makers, computer vision researchers and anyone wishing to identify potentially sensitive content in photos. We conclude by providing insights about how these results may be used to enhance computer vision approaches to protecting image privacy.

1. Introduction

Obfuscating sensitive elements of photos can be a step towards enhancing privacy while maintaining utility (e.g., [10, 13, 17, 29]). For example, Google Street View automatically blurs people's faces and vehicle license plates [13] in an attempt to hide what Google considers to be "personally identifiable data" (i.e., faces and license plate text) [11]. However, it is likely that there is other information in the images captured by street view that is "personally identifiable" or, perhaps more important from our perspective, is considered sensitive or private by those people whose images are captured.

A prerequisite to developing human-centered photo privacy protections is that we must first know what content *users* consider sensitive or private. However, most existing work approaches obfuscation of elements in photos from a security modeling or legal perspective, rather than a human-centered perspective. From the legal perspective, only predefined items such as social security number and date of birth are to be protected [25]. From the security modeling perspective, there must be a threat and an attacker [4]. From a human-centered perspective, though, privacy is far more complex than either the legal or security modeling perspectives suggest. For example, while an "immodest outfit" may not be a target for an attacker, nor covered under legal definitions of PII, it may very well be considered sensitive by users.

Therefore, to design the requirements for computer vision systems that identify and obfuscate sensitive elements in photos, we must have precise knowledge regarding *users'* perspective about sensitive content in photos. However, this knowledge is currently lacking.

To generate a systematic framework that is based on a human-centered understanding of privacy and content sensitivity to guide the recognition of content which could be obfuscated to improve privacy, we collected 98 data elements (e.g., photos) from 20 MTurk participants, and asked them to identify the sensitive content in each photo. Using this data, we created a two-level taxonomy of sensitive content categories, with detailed examples of each category (Table 2). This framework may be useful as a guideline when developing online photo privacy protection mechanisms. Specifically, this framework suggests what content should be obfuscated from a human-centered perspective. After describing what content is sensitive and why people are reluctant to share it, we provide insights on how computer vision may be able to automate identification of sensitive content and describe related challenges which may guide future computer vision research.

2. Background

In this section, we introduce the two important aspects of photo privacy protection: obfuscation method and sensitive content. Ongoing work is exploring methods to obfuscate sensitive content. In this work we focus on the prerequisite aspect—determining what content is sensitive.

Obfuscation, may be defined as "the production of noise modeled on an existing signal in order to make a collection of data more ambiguous, confusing, harder to exploit, more difficult to act on, and therefore less valuable" [7]. Obfuscating photos, which controls the information disclosure, has been adopted by both OSN (Online Social Network) users in the wild and researchers. For example,

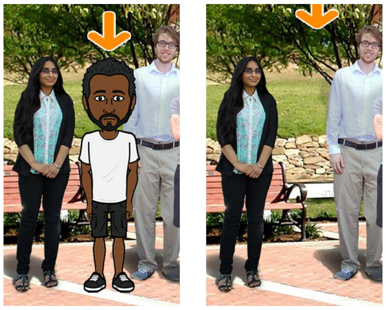

Figure 1. Obfuscation examples. Left: avatar. Right: inpainting.

privacy-conscious users apply mosaics on their vehicle license plates or add emojis on other irrelevant people's faces in a photo. However, users are often unaware of the potential sensitive content; that is to say, they may not know what content to obfuscate. On the other hand, researchers are turning to develop mechanisms to automatically recognize sensitive content and obfuscate it [17, 31].

Obfuscation method (what to use), and sensitive content (what to obfuscate) are two important considerations when developing an effective and usable photo protection mechanism. With adequate existing knowledge on the first aspect, the second aspect—sensitive content—needs more investigation.

2.1. Obfuscation Method

Prior work has identified the trade-off existing between the effectiveness and the user experience of obfuscations. For example, one study finds that blurring provides a good experience while it is not effective; a highly effective form of obfuscation, blocking, destroys photo aesthetic, thus it is not likable [24]. They also suggest two promising obfuscations – avatar (replacing the person with a cartoon avatar) and inpainting (completely removing a person) (Figure 1), which are both effective and provide a good user experience [24]. Another study further applies different obfuscations on scene elements in a photo and investigates their effectiveness and viewer experience. Their result shows applying silhouettes on objects performs well [14]. In brief, prior works have provided some effective and usable obfuscations options. The next step is to understand what content in a photo should be obfuscated, which is our focus.

2.2. Sensitive Content

To the best of our knowledge there is no work that systemically identifies and summarizes sensitive content in photos. Most photo obfuscation systems consider people's identity to be the highest priority sensitive content [10, 17, 29]. Aside from the identity of people, other elements may also reveal personal information (e.g., a phone screen showing intimate text messages may be unintentionally captured in a photo [21]). In ubiquitous computing situations, monitor screen and irrelevant persons in photos are considered sensitive [16], while photos containing text information, people's address, organization, and email lead to privacy concerns [3, 12]. OSN users are also concerned about some objects in photos and photo backgrounds [2].

Category	Citation	Research method
Identity	[1] [5] [17] [22]	Interview Focus group N/A Interview
Nudity	[27]	EU Data Protection Directive 95/46/EC, US Privacy Act of 1974, OSNs rules
Factors that harm impression management	[1] [16]	Interview In situ study
Factors that reveal personal information	[16] [3]	In situ study Previous news
Illegal	[5]	Focus group
Photo quality	[19] [22]	Survey Interview

Table 1. Sensitive content in prior work.

There is no existing framework that summarizes and categorizes sensitive photo content. Based on our review of existing literature, we have identified six categories of sensitive image content discussed in previous work (Table 1).

2.3. Limitations of prior work

There are two major limitations in prior work. First, these identified types of sensitive content are mostly coarse-grained categories, which do not provide enough information that can be used practically. For example, it is unclear what content on a computer monitor is sensitive (a personal bank account number is sensitive, but The New York Times' website may not be sensitive). Hence, they are unable to support photo privacy protection mechanisms in capturing most of the potential privacy threats in online photos. Second, some works use federal laws [25] and official OSN rules to predict sensitive content, instead of collecting real users' opinions about sensitive content in their own photos [27]. This means the mechanisms using their definition of 'sensitive content' may not meet users' real needs, which harms their applicability in a real-world usage scenario and reduces their power if applied on OSNs. Our work aims to bridge this gap by identifying sensitive content categories from real users' opinions about their own photos.

Figure 2. Sensitive content examples. Left: old houseware. Right: medical condition. To protect participants' privacy, these are representative images similar to those uploaded by participants.

3. Method

To understand what photo content is sensitive, we collected photos and/or descriptions of photos with sensitive content from MTurkers. We defined private for the purposes of this study as photos that participants do not want to share with 1) anyone, 2) family, 3) friends, 4) colleagues/classmates, and 5) acquaintances. For each photo a participant uploaded we asked them "What content in this photo do you consider sensitive?" Our analysis of sensitive content types is based on the answer to this question.

3.1. Participants

We recruited 20 participants (11 male, 9 female) via Amazon Mechanical Turk. Seventy-five percent of participants were aged 25-34. The majority were white (65%), while the other 35% were black, Asian, and Hispanic. At least 80% reported using the Internet and OSNs a few times a day, and 70% uploaded photos to OSNs several times a day, once a day, or several times a week. The other 30% uploaded photos a few times a month.

3.2. Procedure

After giving their informed consent, participants answered six demographic questions and two social network familiarity questions. Next, we asked participants to look at the photos stored on their phone and identify a photo that they considered "private" (photo 1). Once they identified such a photo we provided three options: 1) share the photo with us, 2) look for a similar photo online with similar content and share that photo with us, or 3) describe the photo in detailed text.

After uploading or describing the first photo, participants repeated this procedure four times. For each photo, a different context was provided. Specifically, we asked them to identify a photo they would not want to share with their family (photo 2), friends (3), colleagues/classmates (4), and acquaintances (5).

4. Results

In total, we collected 98 data points. Of these, 91 were photos from participants. Sixty-five photos were personal photos and the remaining twenty-six are photos that participants found online that were similar to the personal photos they identified on their phone. Additionally, two participants opted to provide text descriptions for some of their photos, resulting in a total of seven text descriptions.

For each data point, participants answered the question "What content in the photo do you consider sensitive?" In this question, some participants also provided their thoughts and additional comments about the sensitive content they identified. We first confirmed that their description matched the content in the corresponding photo or text, then grouped the answers into seven main categories (see the two-level sensitive content categories in Table 2).

The categories that emerged from our data roughly align with the six sensitive content categories we derived from previous literature as described in the background section. However, our data suggests both expansion and refinement of these categories. Below we discuss each main category and how an overall categorization of sensitive content can be developed (see Table 2).

Consistent with previous work [1, 5], we found that participants in our study identified *identity of people* as sensitive content. Regarding *children*, despite the many benefits of mothers sharing children's photos (e.g., archiving childhood, receiving validation of motherhood [22]), participants in our study expressed concerns when answering the sensitive content question that their photos might be accessed by untrusted audiences, and one participant also stated that this was a decision that her son would make by himself when he was old enough.

Nudity or partial nudity is another common concern, including nudity of the photo owner, their partner, or unknown people in photos they download to their phones.

Also aligned with previous work, participants identified sensitive content congruent with extensive *impression management*, refusing to share photos with unflattering *appearance* (e.g., messy hair, inmodest outfit) and *behavior that may be interpreted in a negative way* (e.g., inappropriate joke) [1].

The next two categories refer to different types of *personal information*. Other than the *identity information* which previous work has mentioned [3, 12], we uncover that *medical conditions* are sensitive, indicating an unhealthy condition accordingly shows the person's weakness to photo viewers (right image in Figure 2). Moreover, viewers may infer disease, which could be highly private, based on revealed medical information. Another concern is content that exposes them being *LGBT* (e.g., attending gay pride with a same-gender partner). Users' *personal habits and interests* are considered sensitive as well: *"I do not*

Category	Sub-category	Sensitive content examples
People in the photo	- The identity of people (22) - Children (9)	- Photo owner's face, family member's face, cousin, ex-girlfriend - Young step son, photo owner's kids, young niece
Nudity	- Nudity or partial nudity (11)	- Wife/girlfriend nudity, photo owner nudity, shirtless, wife's Bikini
Impression management	- Appearance (5) - Facial expression (2) - Embarrassing shots (2) - Behavior/activity that may be interpreted in a negative way (4) - Low economic status (2) - Sensitive environment (3) - Private area at home (3) - Photo quality (1) - Pet's behavior (1)	- Messy hair, no makeup, inmodest outfit, bare feet, fat body - Bad facial expression, goofy face - Embarrassing shot of me sleeping in a chair - Modeling clothes and being vain, inappropriate/risque joke, photo owner is with a lot of food and looks bad - Low-quality food, old housewares - Party, club, bar - Bathroom, toilet, a messy corner - bad angle and being technically flawed - Dog peeing
Factual personal information	- Identity information (3) - Medical condition (7) - LGBT related (3) - Affiliation (1) - Phone screenshot (2) - Location (1)	- ID card, bank account, signature, home address, family trust - Hospital stay, head injury, vomiting, skin rash, grandma in hospital - Hang out with the same-gender partner, transvestite - Wearing army uniform - Text messages with a friend, screenshot with time on it - Attending a social event which reveals location
Subjective personal information	- Habit/interest (2)	- Exercise instruction, album cover, hang out with special friends
Could get me into trouble	- Illegal/inappropriate content (5) - Unauthorized/no permission (6)	- drinking, drunk, teenage illegal drinking, drug test result - supervisor's face, friend met online, family member, roommate, a person passed away and not able to ask him for permission
Personal moment	- Personal moment (3)	- Kissing, intimate and affectionate moment, date night

Table 2. Two-level sensitive content categories with examples. The number in parentheses represents the number of photos/text descriptions in this sub-subcategory.

want them to know the type of music I listen to as they may judge me."

Sharing *illegal content*, such as underage drinking and drug test result on OSN is risky, because they *"could have got in trouble for doing so."* We find that participants respect others' privacy when they are *unauthorized* to share a photo. This happens when a person in the photo *"asked me not to share it."*

5. Discussion

In the previous section we outlined **what** content is considered sensitive. In this section, we summarize **why** people are not willing to share various types types of sensitive content. We also discuss **how** computer vision can be applied to our findings, as well as the associated challenges.

5.1. Reasons for not sharing

The categories outlined in our results section reflect four main reasons for not sharing a photo. The first reason is to maintain a good impression. People use OSN as a tool to manage their impression, by tuning themselves from the actual self to the ought self (a person's representation of the attributes that others believe he or she should have) [15]. Hence, they are willing to share photos that emphasize socially desirable characteristics [9], but avoid those that may harm their impression, such as the examples in the *impression management* and *environment* categories.

The second reason is personal, family, and property safety. People have realized that due to the complexity of OSNs, it becomes a hotbed of various crimes. For example, children's photos may be accessed by online predators [33]. Combined with location leakage (e.g., home address), children may even become offline victims. OSN could also facilitate online fraud and identity theft attacks by collecting the user's name, email address, social security number, or bank account information [6, 26]. Safety is mainly a concern in the *factual personal information* and *people* categories.

Third, people avoid sharing photos which could get them into trouble. On the one hand, they will not show evidence on OSNs that they have violated the law. One participant showed us a photo of a positive drug test for marijuana.

She is currently nursing a baby and this photo indicates she smoked right after the baby was born. Sharing this photo online may result in losing custody of her baby. On the other hand, they try to avoid social tension engendered by unauthorized sharing. We find that people generally show their respect for others' privacy concerns if they are asked not to share. However, if photo owners are not aware of others' concerns, the multi-party sharing conflict is still an issue, since OSN users have less control over themselves in group photos uploaded by other people other than untagging [5].

5.2. Opportunities and challenges for computer vision

Existing computer vision mechanisms have achieved almost human level accuracy on *people* recognition using neural networks [30, 32]. Regarding *children*, one work developed an automatic system that detects images containing children in different poses [18]. Combined with effective and usable obfuscations introduced in [24] such as avatar and inpainting, the privacy concerns of the people in the photo should be sufficiently addressed. Nudity, another common sensitive content, can be successfully detected by skin detection and image zoning [28].

With respect to *impression management*, existing technologies enable facial expression detection [8]. After recognizing the unflattering facial expression, they can adjust some facial features to make the photo more flattering. However, for other *impression management* cases—for example *appearance*—the detection is not straightforward. *Appearance* is very subjective and highly dependent on each person and their context. Sometimes the boundary between good and bad content is blurred (e.g., a vintage style outfit vs. an inmodest outfit). Researchers in computer vision should develop more advanced models to distinguish them.

Similarly, for two personal information categories (*factual personal information* and *subjective personal information*), some content can already be accurately identified and then obscured (e.g., texts in different contexts [23]), while challenges remain in automatically detecting *personal habits and interests*. People's tastes and associated privacy concerns are divergent. Hence, the future mechanisms should be more customized for individual users. Take album covers as an example: ideally, the system learns from a user's public music preferences (e.g., followed bands, liked musicians) to get an understanding of what they prefer to show to their audience, then filter out and obfuscate any accidentally captured album covers in a photo that show obviously discrepant tastes.

Regarding the subcategory *illegal/inappropriate content*, computer vision mechanisms should have the ability to accurately capture certain objects, such as beer cans and drug test strips. Once identified, systems can either replace a beer can with a coke can, or suggest not to make it public.

6. Limitations and Future Work

While we categorized the sensitive content in photos, there are other factors that may affect the sensitivity, such as photo recipient, sharing context, sharing purpose, and the preferences of users. These factors might even interact with the content (i.e., content that is innocuous in one context may be considered sensitive in another context). One potential way to address this issue is with user-tailored privacy solutions that address individual differences and predict users' privacy preferences based on their known characteristics to provide personalized privacy settings [20]. In our next study, we plan to study one of these factors in depth: the recipient. By investigating users' sharing preference of each sensitive category with different recipient groups, we will get a more complete picture of the requirements for human-centered privacy protections for photos.

7. Conclusion

To develop an effective photo privacy protection system, the first thing designers need to know is what content is considered sensitive and therefore needs to be obscured. However, until now, there has been no systematic framework for identifying sensitive content in photos that was based on data rather than anecdote. In this work, we present a framework based on a human-centered study that identifies sensitive content in photos. This framework may benefit privacy, online social network, and computer vision researchers, and provide insights about how computer vision may more effectively and efficiently enhance photo privacy.

8. Acknowledgements

This research was supported by the National Science Foundation under grant no.1527421. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank our colleagues from the HATlab for suggestions that improved this study. We are also grateful to the people who participated in this study.

References

[1] A. Adams, S. J. Cunningham, and M. Masoodian. Sharing, privacy and trust issues for photo collections. 2007. 2, 3

[2] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 357–366. ACM, 2007. 2

[3] T. Aura, T. A. Kuhn, and M. Roe. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 41–50. ACM, 2006. 2, 3

[4] J. Bau and J. C. Mitchell. Security modeling and analysis. *IEEE Security & Privacy*, 9(3):18–25, 2011. 1

[5] A. Besmer and H. Richter Lipford. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1563–1572. ACM, 2010. 2, 3, 5

[6] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web*, pages 551–560. ACM, 2009. 4

[7] F. Brunton and H. Nissenbaum. *Obfuscation: A user's guide for privacy and protest*. 2015. 1

[8] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 915–920. IEEE, 2011. 5

[9] M. D. Dorethy, M. S. Fiebert, and C. R. Warren. Examining social networking site behaviors: Photo sharing and impression management on facebook. *International Review of Social Sciences and Humanities*, 6(2):111–116, 2014. 4

[10] L. Du, M. Yi, E. Blasch, and H. Ling. Garp-face: Balancing privacy protection and utility preservation in face de-identification. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014. 1, 2

[11] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in google street view. In *IEEE International Conference on Computer Vision*, 2009. 1

[12] L. Geng, L. Korba, X. Wang, Y. Wang, H. Liu, and Y. You. Using data mining methods to predict personally identifiable information in emails. In *International Conference on Advanced Data Mining and Applications*, pages 272–281. Springer, 2008. 2, 3

[13] Google Street View. Image acceptance and privacy policies, 2018. Retrieved March 07, 2018 from https://www.google.com/streetview/privacy/. 1

[14] R. Hasan, E. Hassan, Y. Li, K. Caine, D. J. Crandall, R. Hoyle, and A. Kapadia. Viewer experience of obscuring scene elements in photos to enhance privacy. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2018. 2

[15] E. T. Higgins. Self-discrepancy: a theory relating self and affect. *Psychological review*, 94(3):319, 1987. 4

[16] R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia. Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, pages 1645–1648. ACM, 2015. 2

[17] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 781–792. ACM, 2015. 1, 2

[18] M. Islam, P. A. Watters, and J. Yearwood. Real-time detection of childrens skin on social networking sites using markov random field modelling. *Information Security Technical Report*, 16(2):51–58, 2011. 5

[19] S. Kairam, J. Kaye, J. A. Guerra-Gomez, and D. A. Shamma. Snap decisions?: How users, content, and aesthetics interact to shape photo sharing behaviors. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 113–124. ACM, 2016. 2

[20] B. P. Knijnenburg. Privacy? i cant even! making a case for user-tailored privacy. *IEEE Security & Privacy*, 15(4):62–67, 2017. 5

[21] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Enhancing lifelogging privacy by detecting screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4309–4314. ACM, 2016. 2

[22] P. Kumar and S. Schoenebeck. The modern day baby book: Enacting good mothering and stewarding privacy on facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1302–1312. ACM, 2015. 2, 3

[23] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. Yuille, and C. Koch. Adaboost for text detection in natural scene. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 429–434. IEEE, 2011. 5

[24] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction*, 1(2), 2017. 2, 5

[25] E. McCallister, T. Grance, and K. A. Scarfone. Guide to protecting the confidentiality of personally identifiable information (pii). Technical report, 2010. 1, 2

[26] T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *Journal of Economic Perspectives*, 23(3):3–20, 2009. 4

[27] T. Orekondy, B. Schiele, and M. Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3706–3715. IEEE, 2017. 2

[28] C. Santos, E. M. dos Santos, and E. Souto. Nudity detection based on image zoning. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 1098–1103. IEEE, 2012. 5

[29] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. *arXiv preprint arXiv:1711.09001*, 2017. 1, 2

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 5

[31] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, and G.-J. Ahn. Towards pii-based multiparty access control for photo sharing in online social networks. In *Proceedings of the 22nd*

ACM on Symposium on Access Control Models and Technologies, pages 155–166. ACM, 2017. 2

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 5

[33] J. Wolak, D. Finkelhor, K. J. Mitchell, and M. L. Ybarra. Online" predators" and their victims: myths, realities, and implications for prevention and treatment. *American psychologist*, 63(2):111, 2008. 4