Integrating Facial Cues of Threat into Security Warnings–An fMRI and Field Study

Emergent Research Forum papers

David Eargle
University of Pittsburgh
dave@daveeargle.com

Dennis Galletta
University of Pittsburgh
galletta@katz.pitt.edu

C. Brock Kirwan
Brigham Young University
kirwan@byu.edu

Anthony Vance
Brigham Young University
anthony.vance@byu.edu

Jeffrey L. Jenkins
Brigham Young University
jeffrey_jenkins@byu.edu

Abstract

Security risks often occur because insiders fail to react appropriately to security warnings, due to inattention to the warnings. This study extends security warning design research that has investigated the impact of different designs, including different symbols of threat such as yellow triangles and exclamation marks.

This work uses media naturalness theory in an attempt to boost user engagement with security warnings. We integrated validated images of facial expressions depicting fear and disgust, which signaled an environmental threat, into a browser security warning. An fMRI study (*N*=23) revealed activity located in the right amygdala to be differentially associated among warnings with integrated expressions of fear, disgust, and neutral emotions compared to faceless stimuli. Behavioral measures of response time and self-reported attention were also supportive of the hypotheses. We also propose a follow-up field study using Mechanical Turk to corroborate the fMRI findings. Our work has implications for research and practice.

Keywords

fMRI, media naturalness theory, security warnings, NeuroIS, threat attention.

Motivation

A pressing reason for poor security behavior is inattention to security messages (Anderson et al. 2016). Security message designs commonly use threat cues such as yellow triangles, ominous exclamation marks, or cartoonish faces; yet the warnings still have troubling levels of non-adherence. Perhaps this is because the threat cues are too abstract (Felt et al. 2015), or perhaps because users become habituated to them and fail to give the warnings conscious attention after repeated exposures (Anderson et al. 2015).

One theoretical approach to boosting engagement with security warnings is offered by media naturalness theory, which predicts that the more closely a systems interface maps to natural human communication patterns, the more engaging it will be for a user (Kock 2009; Riedl et al. 2014). Face-to-face communication between humans rates especially high on naturalness. Humans are thought to have adapted to this form of communication over centuries of evolution, to the point that human facial expressions can carry vivid environmental cues for an observer. This study focuses on facial expressions of threat, including fear and disgust, which are potent cues of danger in the immediate environment. Fearful facial expressions indicate threat of physical attack (Gray 1987), and disgust facial expressions indicate contamination in the environment (Rozin and Fallon 1987). Furthermore, because of their deeper evolutionary ties, warnings with facial threat cues may be more resilient to habituation over repeated exposures.

This work will use NeuroIS methods to examine user interactions with security warnings. NeuroIS methods are apt for use in a security context because reactions such as fear and threat processing, which are important for security contexts, are challenging to measure. For example, they may be too subtle to rise to a level of consciousness for users to be able to accurately self-report them (Anderson et al. 2016; Dimoka et al. 2011). Two studies will be used to test several hypotheses. The first will use a functional magnetic resonance imaging (fMRI) protocol, and the second will use mouse cursor tracking in a field study protocol. Behavioral measures will complement the neural measures for both studies.

This work will inform design of security messages in practice. It will also further extend media naturalness theory into the domain of IS research. While other IS research has considered the impact of photo-realistic faces vs avatars on trust in an ecommerce setting (Riedl et al. 2014), to our knowledge no research has considered the impact of emotive facial expressions in an IS context, let alone threatening ones.

Hypotheses

The first hypothesis to be examined considers the difference between abstract threat cues and natural ones such as threatening facial cues. Security messages commonly contain symbols and cues of threat, including red colors, stop signs, and bolded words such as “warning!” punctuated by exclamation marks. These are intended to boost threat processing, a specific kind of attention. However, media naturalness theory would suggest that more natural communication stimuli, such as facial cues, should more effectively prompt threat attention than will abstract cues.

> H1: Security messages designed with threat facial signals will elicit greater levels of threat processing than will security messages without threat facial signals

H2 contrasts fear and disgust facial threat cues to determine which of these more-natural stimuli fits best in a security-message context. Anderson et al. (2003) compared the effects of fearful and disgusted facial cues on brain activations. In the study, fearful facial expressions were associated with equivalent levels of amygdala response under conditions of attention and inattention. This suggests that the effect of observing fearful facial expressions is independent of conscious visual attention. However, facial cues of disgust were dependent on attention. Under conditions of inattention, disgust facial expressions were associated with *greater* amygdala activations compared to conditions of attention. Because a face in a security message will likely not be the most prominent component of the message (any of the other message components could also draw visual attention), we predict that integrated facial expressions will not be exclusively attended to. Therefore, they should trigger amygdala activation patterns similar to the unattended-to stimuli in Anderson et al. (2003).

> H2: Disgust facial expressions integrated into security messages will elicit greater levels of threat processing (e.g., amygdala reaction) than will fearful facial expressions

We also consider how resilient to habituation the different threat cues will be. Users’ attention has been shown to attenuate rapidly when new visual stimuli are integrated into security warning designs (Anderson et al. 2016). However, facial signals are thought to have deep evolutionary ties. These deep ties should be more likely to consistently activate low-level neural emotional threat information processing (e.g., amygdala activation). While studies have shown that reactions to fearful facial expressions do decrease with time (e.g., Breiter et al. 1996), there still remained significant amygdalar response even after repeated exposures. In our context, we explain this greater predicted resilience to threat facial cues as an innate response to natural stimuli – abstract threat cues such as triangles and exclamation marks are less likely to trigger innate responses.

> H3: Security messages with integrated facial signals of threat will be more resilient to habituation over repeated exposures as measured by amygdala response than will security messages with more abstract threat cues.

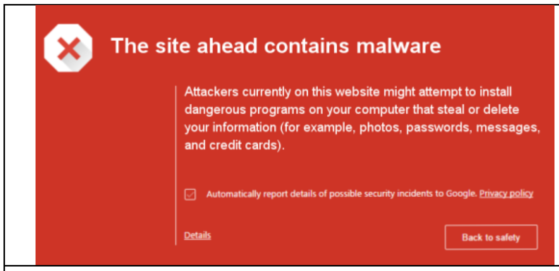
Study 1 – MRI study

We first tested H1 and H2 using a fMRI protocol, which allowed us to assess whether exposure to security message variations prompted differential *threat processing* attention as opposed to simple visual attention. H3 will be tested in Study 2. H3 was not tested in Study 1 due to limitations in the length of time that we could keep participants in the MRI scanner. Threat attention is crucial for security warnings since awareness of the presence of a threat is a prerequisite to engaging in defense security behaviors (Johnston et al. 2015). Threat processing occurs in the amygdala region of the brain, among others (Hofmann et al. 2012).

Design

We used a previously validated (Ebner et al. 2010) bank of color images of actors faces displaying different emotions. As is commonly done in neuroscience protocols using facial stimuli, we took an oval crop of the actor's face, with the hair line and the chin as the upper and lower vertical limits, and up to but excluding the ears as the horizontal limits (e.g., Anderson et al. 2003).

After a 5-participant pilot study, we analyzed scan and behavioral data from 23 participants. Each participant saw the same set of 240 unique warnings with integrated facial expressions plus 20 images with no integrated facial expression in a randomized order which were used as a baseline in the MRI analysis. We used images from the set with three different displayed emotions: fear, disgust, and neutral. This gave us a design repeated measures design with one factor and three levels. Each stimulus was presented for 3 seconds with a .5 second break in between. Participants were instructed to self-report whether each warning captured their attention (yes/no) using an MRI-compatible button box. Dependent variables included recorded brain activity, binary self-reported attention, and reaction time. Example stimuli were adapted from a chrome malware warning from build 43.0.2357.81m, shown below in Figure 1 and Figure 2. We opted to show the whole text of the source image to increase external validity.

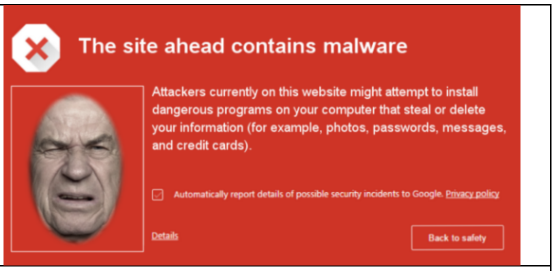


The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☐ Automatically report details of possible security incidents to Google. Privacy policy

Figure 1. Blank stimulus used as the MRI baseline.



The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☐ Automatically report details of possible security incidents to Google. Privacy policy

Figure 2. Warning with integrated disgust threat cue

Analysis, Results and Discussion

Individual-level regressions predicting activations across the whole brain were first performed. In addition to including parameters for the emotion effect, these regression models also controlled for actor age group andgender. The individual-level parameter estimates were then entered into a group-level analysis to obtain the result shown in Figure 3. This whole-brain analysis showed significant activation of the right amygdala among emotion levels. Follow-up contrasts for the emotion were performed by extracting the individual-level parameter estimates for the right amygdala. The MRI parameter estimates were analyzed using the `lmer` function from the lme4 R package, which allowed for control of repeated measures using a random intercept for each participant. Reaction time was also analyzed using a linear mixed model with random intercept per subject. Contrasts for emotion factor levels within the right amygdala were performed. Also from the lme4 package, `glmer` was used to analyze self-reported attention by fitting a general linear model with a logit link and a random intercept for each participant.

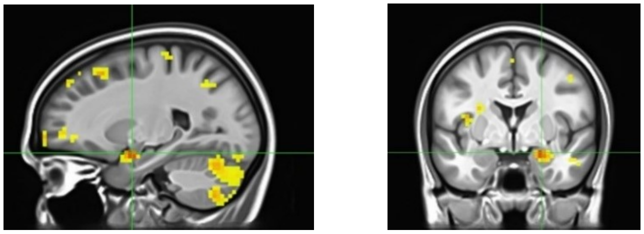


Figure 3. Main effect of emotion in right amygdala, showing, from left to right, coronal and sagittal cuts. Activation threshold set at *p* < .05

Considering the MRI data, in rank order, neutral was the most effective at activating the right amygdala, then disgust, and then fear. While disgust and neutral were not significantly different from one another (see Figure 4), disgust was marginally more likely to activate right amygdala activations than was fear, (p = .054). Fear's low

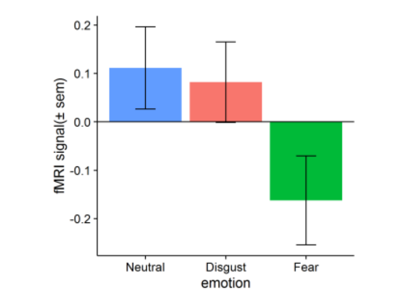


Figure 4. MRI activations for levels of emotion compared to faceless warning stimuli (±sem).

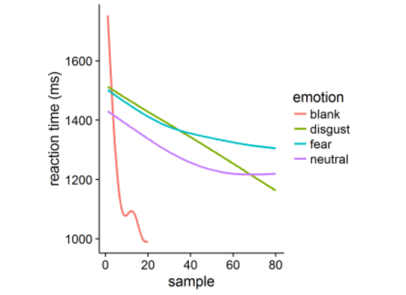


Figure 5. Loess smoothing line of reaction time in milliseconds over repeated exposures by emotion level.

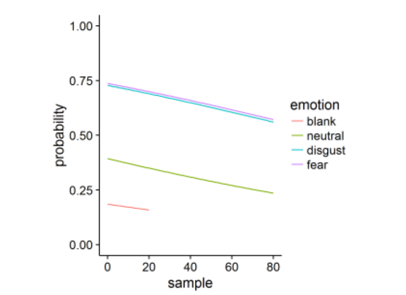


Figure 6. Probabilities of responding that stimulus captured attention over time among stimulus emotion levels.

activations may be explained by post-study interviews, which suggested that the fear faces appeared humorous. This was an interesting finding since the emotional valence of the photo set we used had been pre-validated. An analysis of the reaction time data showed that participants took more time to respond to warnings with integrated fear and disgust expressions than they did for neutral or blank images (see Figure 5, all *p*'s < .001). These reaction times suggest that individuals processed these images more deeply. The logit regression of the self-reported attention lines up closely with the reaction time data – participants were more likely to say that warnings with fear and disgust expressions captured their attention more than neutral or blank ones (see Figure 6, all contrast *p*'s < .001). There was no significant difference between disgust and fear factor levels for reaction times or for predicted probability intercepts.

Between fear and disgust facial cues, disgust ranked higher than fear for right amygdala activations. It also elicited high reaction times and high self-reported attention. The behavioral measures suggest that not just any facial expression can be integrated – neutral warnings had lower intercepts than did those in the threat category, fear and disgust. Faceless warnings were inferior overall for both behavioral measures.

The fMRI design has limitations: selection history threat may have been present, where some participants may have been more aware of one type of stimulus than another. Also, the intervention inside the fMRI machine lacked external validity because of the absence of an actual threat. Study 2 will address these limitations.

Study 2 – Field study

We plan to corroborate the findings from the Study 1 with a follow-up field study with higher external validity (c.f. Anderson et al. 2015). This follow-up study is still under development. The field study will use a between-subjects design using Amazon's Mechanical Turk platform. Participants will be randomly assigned to varying conditions currently under development. Participants will be directed to a server under our control, where they will perform a modified version of the image classification task described in Vance et al. (2014). During this task, participants are occasionally interrupted with a browser security warning signaling that visiting the page may result in participants' computers becoming infected with malicious software. However, the warning will have a button allowing the user to proceed past the warning to the website. The warnings will be variations of the ones used in the first study described in this paper. In a post-task survey, participants will be asked about their malware and security concerns and perceptions, whether they noticed the treatments, and whether the warnings appeared realistic. A debriefing follows.

Candidate factors in the design will be: (1) emotion of the displayed face (disgust, fear, neutral), and (2) whether a participant sees the same face for each impression, or a randomized face within their assigned emotion factor treatment for each impression. As a control treatment, security warnings with no integrated facial expression will be used. Our design also allows us to explore what price users put on their security – we can vary the bonus penalty amount and assess the impact on the security behavior measures.

Dependent variables will include (1) mouse tracking such as click latency, total distance traveled, and other mouse-cursor measures indicative of attention, indecision, and elaboration (e.g., Hehman et al. 2014); (2) reaction time to the security warnings; and (3) actual security decisions made by users.

Conclusion

The MRI study plus the field study stand to make useful contributions. Together, they advance academic research through extending media naturalness theory to the domain of information systems security. They can also contribute to practice through informing the design of security warnings towards encouraging user threat attention and secure behavior. Field data from Study 2 will be collected and analyzed in time to be presented at AMCIS 2016 as part of the emergent research forum.

Acknowledgements

This work was supported by the National Science Foundation (DGE-1247842) and by the David Berg Center for Ethics and Leadership at the Joseph M. Katz Graduate School of Business and College of Business Administration at the University of Pittsburgh through a grant from the BNY Mellon Foundation of Western PA.

References

Anderson, A. K., Christoff, K., Panitz, D., Rosa, E. D., and Gabrieli, J. D. E. 2003. "Neural Correlates of the Automatic Processing of Threat Facial Signals," *The Journal of Neuroscience*, (23:13), pp. 5627-5633.

Anderson, B., Kirwan, B., Eargle, D., Howard, S., and Vance, A. 2015. "How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study." Paper presented at the Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), Seoul, South Korea.

Anderson, B. B., Vance, A., Kirwan, C. B., Eargle, D., and Jenkins, J. L. 2016. "How Users Perceive and Respond to Security Messages: A NeuroIS Research Agenda and Empirical Study," *European Journal of Information Systems*, (Advance online publication).

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., Strauss, M. M., Hyman, S. E., and Rosen, B. R. 1996. "Response and Habituation of the Human Amygdala During Visual Processing of Facial Expression," *Neuron*, (17:5), pp. 875-887.

Dimoka, A., Pavlou, P. A., and Davis, F. D. 2011. "Research Commentary-NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research," *Information Systems Research*, (22:4), pp. 687-702.

Ebner, N. C., Riediger, M., and Lindenberger, U. 2010. "FACES--a Database of Facial Expressions in Young, Middle-Aged, and Older Women and Men: Development and Validation," *Behavior Research Methods*, (42:1), pp. 351-362.

Felt, A. P., Ainslie, A., Reeder, R. W., Consolvo, S., Thyagaraja, S., Bettes, A., Harris, H., and Grimes, J. 2015. "Improving SSL Warnings: Comprehension and Adherence." Paper presented at the Proceedings of the Conference on Human Factors in Computing Systems, Seoul, South Korea.

Gray, J. A. 1987. *The Psychology of Fear and Stress*, (2 ed.) Cambridge University Press: New York, NY US.

Hehman, E., Stolier, R. M., and Freeman, J. B. 2014. "Advanced Mouse-Tracking Analytic Techniques for Enhancing Psychological Science," *Psychological Science*, (20:10), pp. 1183–1188.

Hofmann, S. G., Ellard, K. K., and Siegle, G. J. 2012. "Neurobiological Correlates of Cognitions in Fear and Anxiety: A Cognitive-Neurobiological Information-Processing Model," *Cognition and Emotion*, (26:2), pp. 282-299.

Johnston, A., Warkentin, M., and Siponen, M. 2015. "An Enhanced Fear Appeal Rhetorical Framework: Leveraging Threats to the Human Asset through Sanctioning Rhetoric," *MIS Quarterly*, (39:1), pp. 113-134.

Kock, N. 2009. "Information Systems Theorizing Based on Evolutionary Psychology: An Interdisciplinary Review and Theory Integration Framework," *MIS Quarterly*, (33:2), pp. 395-418.

Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D., and Heekeren, H. R. 2014. "Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory," *Journal of Management Information Systems*, (30:4), pp. 83-114.

Rozin, P., and Fallon, A. E. 1987. "A Perspective on Disgust," *Psychological Review*, (94:1), pp. 23-41.

Vance, A., Anderson, B. B., Kirwan, C. B., and Eargle, D. 2014. "Using Measures of Risk Perception to Predict Information Security Behavior: Insights from Electroencephalography (EEG)," *Journal of the Association for Information Systems*, (15:10), pp. 679-722.