Pitfalls of Affective Computing

How can the automatic visual communication of emotions lead to harm, and what can be done to mitigate such risks?

Martin Cooney, Sepideh Pashami, Anita Sant'Anna, Yuantao Fan, Sławomir Nowaczyk
Center for Applied Intelligent Systems Research (CAISR), Halmstad University,
Halmstad, Sweden
martin.cooney@hh.se,sepideh.pashami@hh.se,anita.santanna@hh.se,yuantao.fan@hh.se,slawomir.nowaczyk@hh.se

ABSTRACT

What would happen in a world where people could "see" others' hidden emotions directly through some visualizing technology? Would lies become uncommon and would we understand each other better? Or to the contrary, would such forced honesty make it impossible for a society to exist?

The science fiction television show Black Mirror has exposed a number of darker scenarios in which such futuristic technologies, by blurring the lines of what is private and what is not, could also catalyze suffering. Thus, the current paper first turns an eye towards identifying some potential pitfalls in emotion visualization which could lead to psychological or physical harm, miscommunication, and disempowerment. Then, some countermeasures are proposed and discussed – including some level of control over what is visualized and provision of suitably rich emotional information comprising intentions – toward facilitating a future in which emotion visualization could contribute toward people's well-being.

The scenarios presented here are not limited to web technologies, since one typically thinks about emotion recognition primarily in the context of direct contact. However, as interfaces develop beyond today's keyboard and monitor, more information becomes available also at a distance – for example, speech-to-text software could evolve to annotate any dictated text with a speaker's emotional state.

CCS CONCEPTS

• **Information systems** → *Sentiment analysis*; • **Security and privacy** → *Human and societal aspects of security and privacy*; • **Human-centered computing** → *Visualization systems and tools*;

KEYWORDS

Affective computing; emotion visualization; Black Mirror; privacy; ethics; intention recognition

ACM Reference Format:
Martin Cooney, Sepideh Pashami, Anita Sant'Anna, Yuantao Fan, Sławomir Nowaczyk. 2018. Pitfalls of Affective Computing: How can the automatic visual communication of emotions lead to harm, and what can be done to mitigate such risks?. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191611

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191611

1 INTRODUCTION

Research in affective computing, emotion recognition, and sentiment analysis aims to improve people's well-being by enabling computers and robots to better make decisions and serve, through awareness of people's emotions [19]. Emotions can be recognized, with varying degrees of accuracy, from various signals, including facial expressions, gestures, and voices, using wearables or remote sensors (e.g., galvanic skin response, brain machine interfaces, and cameras) [14, 20, 22]. As more and more of such sensors begin to appear in our surroundings, the capability to recognize emotions can potentially become something available globally, irrespective of physical closeness, thus affecting also web technologies.

Recognition results can be not only used by computers to plan appropriate actions, but also visualized for humans. For example, in our previous work we explored the usage of a robot to paint based on emotions detected in a person using a brain machine interface and thermal camera, as shown in Fig. 1. Such a system could one day be useful to help people with autism, depression, trauma, or alexithymia (difficulty recognizing one's emotions) to communicate and investigate their emotions. Other examples, which did not require a robot, have shown emotions through movements or colored lights in kinetic clothing [16, 18].

In the current paper we envision a future in which such technologies perform with high accuracy and are widespread, so that people's emotions can typically be seen by others. We note that humans are generally adept at recognizing certain types of emotions, but the process can certainly be facilitated and well-hidden emotions can require technology to expose (in fact, people might often not be *fully* aware of their own emotions, much less those of others). Such emotion visualization could offer a number of benefits in many areas, e.g., for artistic expression; facilitating empathizing and emotional resonance; detecting dangerous situations rapidly; and exposing potentially bias in legal workers such as judges or jurors, or falsehood in relationships. However, despite good intentions, such capabilities also present significant dangers, in exposing a person's inner, honest, and potentially vulnerable state for others to see.

2 POTENTIAL PITFALLS

In considering the potential dangers of emotion visualization, we were reminded of some scenarios in Black Mirror, a television series which investigates how weaknesses in human nature could drive

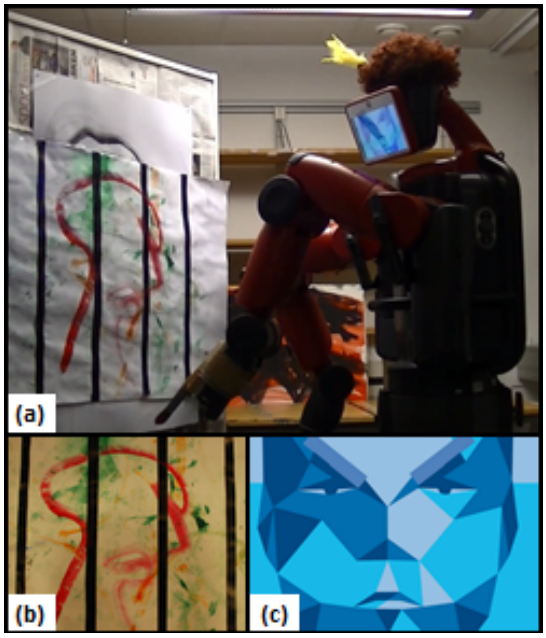

Figure 1: Example of emotion visualization: (a) A robot expresses anger through (b) color, lines, and composition, as well as (c) its facial expression, sounds, and gestures.

the misuse of future technologies [1]. In particular, episodes 2 and 3 of season 4, "Arkangel" and "Crocodile", depict bleak scenarios in which a capability to see into another person's mind invites trouble. In Arkangel, a mother's urge to protect her daughter by secretly monitoring her actions, and manipulating her perception and relationships, leads to her daughter developing psychological problems, and violence. In Crocodile, technology exposing a dark secret leads a person to feel anguish and fear, which pushes them towards committing a string of murders. In a similar way, we foresee how emotion visualization could exhibit "enantiodromia" (the propensity of a system to proceed toward its opposite), by seeking to help but in practice potentially leading to psychological or physical harm, miscommunication, or empowerment of computers at the expense of humans; or one group of humans against another.

2.1 Psychological harm

Emotion visualization could harm people psychologically by revealing information which people do not wish to reveal, or contributing to a weakening of social skills, moral values, and "lebensfreude" (joy of life). For example, exposure of potential emotional problems (e.g. dementia, depression, or obsessive compulsive disorders) for all to see might be undesirable for some persons due to social stigma; we imagine that some people could also try to restrain their emotions, and feel fear that some hidden feeling might be expressed. Indeed, control over the displaying of emotions is often lacking; emotions can be leaked, and acts of deception revealed, through facial microexpressions and body language, making such concealment highly difficult [10].

Furthermore, if it is believed that if there is no need to speak or move to convey emotions, the result could be a weakening of many social skills. Making "white lies" impossible could appear to remove a person's responsibility to be compassionate: for example a person might feel that "the system is to blame if the communication of my emotional state causes offense; it is not my fault". Likewise coerced honesty might result in some lack of development of a person's morals; people who receive less opportunities to exercise honesty might experience more difficulty learning how to do so. This relates to human nature in the sense that removing challenges can make humans weaker. Just as leg strength declines when a person is confined to using a wheelchair, or children who are not exposed early on to viruses and bacteria can become adults with a weakened immune system, people whose emotions are always in "plain sight" might end up being unable to make morally correct choices without the crutch of "forced honesty".

Removing uncertainty can also make life more monotonous. For example, Shakespeare's plays might have been less interesting if Hamlet's mental state, the relationship between Petruchio and Katherina, and the mutual attraction between Romeo and Juliet had been perfectly clear to all – in this last case, the families might have separated the lovers immediately; a much less exciting story.

2.2 Physical harm

Emotion visualization could also lead to physical harm. The ability to quickly detect threats and opportunities via negative or positive emotions could increase the incidence of fighting and violence, as well as promiscuity in dating scenarios and thereby sexually transmitted diseases. For example, a person could feel threatened by seeing their partner and a potential rival feel highly positive emotions toward one another, which could lead to anger and violence[3, 4].

Some individuals and groups might also seek to control others' emotions. This could be carried out mentally, e.g., in the case of an employer who might seek to restrict which emotions are displayed by service staff, but also physically, e.g., a bully might derive pleasure from seeing a victim's fear. In the latter case, however, we note that emotion visualization could also have a positive or neutral effect. Acknowledgement of emotional states in others by a bully could potentially interfere with attempts to dehumanize, increase feelings of guilt, and render more likely the probability that a victim receives empathy [12]. Or, visualization might not exert a strong effect at all – if "reading" emotions becomes easy and ubiquitous, it could lead to desensitization (in the way that people today can become accustomed to seeing disasters on television) [23].

2.3 Miscommunication

A risk of misunderstandings, unseen biases, and falsification also exists. For example, current emotion visualization systems typically assume a person feels one emotion at a time, and do not take into account the full complexity of human emotions, like that emotions are usually directed toward some referent and can coexist [13]; e.g., a person can feel angry when hearing that a loved family member was hurt without feeling anger toward that family member, and cry happy tears when hearing they are safe.

Emotional signals can also be ambiguous; for example, smiling is not always a sign of joy, and nodding can have many meanings [21]. Misunderstandings based on such ambiguity could lead to fighting or unhappiness (e.g. if a person appears to feel joy when another is in pain). Another potential problem is that, like humans, recognition algorithms can also have bias, which could negatively affect someone's life if emotions are used for some evaluation (like in the Black Mirror episode "Nosedive"[1]), and fuel self-fulfilling prophecies [17]; as an example of the latter, a person expecting an interaction to be awkward based on a system's prediction might behave different from usual, which could put others off-balance, leading to awkwardness and making it seem like a system had been correct in its estimation.

Emotion recognition is also a highly challenging task; errors can occur, and the "objectivity" of the output is by no means guaranteed. Thus, if a person's own inference about someone else's emotions differs from what is visualized, dissonance and distrust might be felt. Finally, a system could be hacked to make it seem as if a person genuinely feels a different emotion, which could benefit politicians, lawyers, or criminals.

2.4 Disempowering individuals

If it becomes difficult for some humans to conceal thoughts and emotions, robots and computers could also gain an advantage in persuading people, potentially also for commercial gain. Today interactions with computers are common, e.g. for social media or online shopping; this tendency could be further exacerbated if some people turn to robots as "safe" companions, to avoid sharing private emotions with other human beings [6, 11]. Such systems could be highly convincing: robots without emotions would not need to be careful with choosing words, thus presenting a persuasive air of confidence, free of distractions [15]. Moreover, emotions can also be used to deceive people to affect their decisions [9]. For example, organizations could use such systems to adapt prices presented, e.g. for online products, based on detecting a potential customer's emotional state.

Some potential dangers associated with such a scenario relate to society, trust, and equality. Robots could leverage emotions to deceive people into liking them in order to persuade, resulting in relationships which might not be genuine and potential reductions in a person's social contact with other humans [24]. Partially or completely false information (e.g., "fake news") could be transmitted without fear of having emotions betray the secret, undermining trust; moreover, scams targeting certain demographics such as the elderly could also check that a victim truly believes a story and use emotional feedback to be more convincing, potentially leading to increased suffering. And, presenting different prices to different people could be unfair and promote inequality.

3 STRATEGIES FOR AVOIDING PITFALLS

To minimize the risk that emotion visualization technology will be misused, we propose that some capabilities can be incorporated into systems, as shown in Fig. 2: controls (an off-button, spoofing, and cloaking) and intention recognition. For the former proposal, the core concept is that a person's emotions should only be visualized with their consent, to trusted persons: possibly, a feeling could

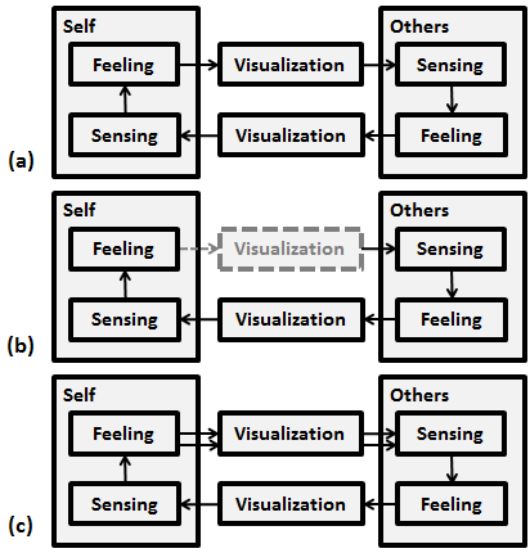

Figure 2: Potential strategies for avoiding pitfalls: (a) A "naive" model for emotion visualization. A person senses, affecting their emotions, which are in turn visualized by systems which they or others own, and influence others' emotions. (b) Controlling what is visualized. The person's visualization system can be turned off completely so nothing is shown, set to play back a sequence of false emotions (e.g., appearing cheerful for clients), or filter the input from recognition to only show some emotions. Other people's visualization systems can be blocked by artificially altering signals associated with emotions such as facial expressions, voice, and body language. (c) Conveying rich information. In scenarios in which misunderstandings could be highly undesirable, information about which emotions are being felt toward what referents can be conveyed.

only be visualized by a person who offers their own feelings to be visualized, as in a mutual exchange of emotions by friends. For privacy in other scenarios, a person's emotion visualization could be turned off, false emotions could be expressed, or some emotions not visualized. Emotions could be hidden from others' systems; for example, by shrouding facial expressions, like adversarial patches which can trick object detection algorithms [5], or by causing clothing to automatically move to inject noise into gestures so that their emotional significance is unclear. Thus, for instance, a person with depression would not be forced to expose their condition to total strangers, employees could avoid being continually monitored and afraid of losing their jobs for some emotional slip-up, and undesired consequences from recognition mistakes in challenging conditions such as low illumination could be mitigated.

It is, of course, a challenge how to ensure those capabilities from a technological perspective: regulations that enforce them can be subverted. And, there is a risk that this will evolve into an "arms race" between better and better spoofing systems being continuously developed, as better and better recognition algorithms evolve. This

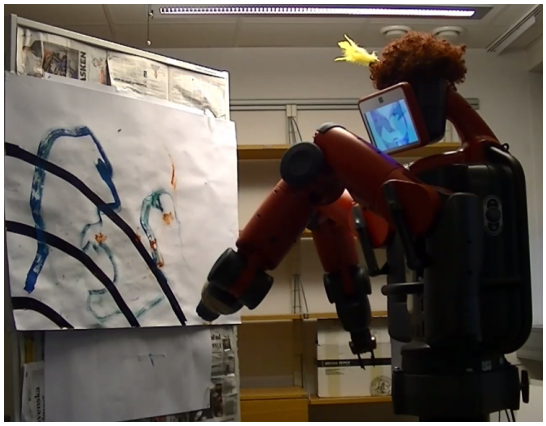

Figure 3: Example of our robot attempting to convey some enriched emotional information, comprising anger (face on the left), fear (face on the right), sadness (via blue color), and generally negative valence (via some descending "mood lines").

could lead to a world where the richest and most technologically advanced are the only ones who can hide their true emotions.

Another approach, to reduce the risk of misunderstandings, might involve intention recognition, i.e., recognizing why and towards what target a person expresses emotions. In previous work, we have proposed an initial approach for intention inference using a "monosemy" metric based on the term frequency-inverse document frequency (TF-IDF) heuristic to quantify the degree to which an individual behavior is typically associated with underlying intentions; in conjunction with leveraging temporal structure of activities, co-occurring signals in other modalities, and multiple observations [7, 8]. However, much work remains to be conducted on this topic, which has been described as "virtually unexplored" [25]. Fig. 3 shows an attempt to concurrently convey multiple emotions. Referents could also be shown together with emotions in some structured manner; for example, in the case of a person crying happy tears, a feeling of joy that a loved one is safe, and arousal due to some avoided danger (e.g., a fire, or car accident), could be visualized as the foreground and background of an image.

If such technological countermeasures are impossible or impractical, an alternative could be to enact legislation to allow emotion visualization for medical, therapeutic, or private purposes only, under sufficient regulation. For example, under privacy regulation such as the General Data Protection Regulation (GDPR) [2], one's emotional state could be regarded as personal information and therefore subject to protections and transparencies which are already available today. We believe that such considerations of what problems can occur and how they can be avoided, will enable emotion visualization to contribute positively to people's well-being.

4 ACKNOWLEDGEMENTS

The authors received funding for this work from the Swedish Knowledge Foundation (CAISR 2010/0271 and Sidus AIR no. 20140220). We thank all those who kindly contributed comments and thoughts!

REFERENCES

[1] [n. d.]. Black Mirror episode explanation. ([n. d.]). en.wikipedia.org/wiki/Black_Mirror
[2] [n. d.]. EU General Data Protection Regulation (GDPR). ([n. d.]). https://www.eugdpr.org/
[3] [n. d.]. Fear-Based Anger Is the Primary Motive for Violence. ([n. d.]). https://www.psychologytoday.com/blog/wicked-deeds/201707/fear-based-anger-is-the-primary-motive-violence
[4] Robert Agnew. 2006. General strain theory: Current status and directions for further research. *Taking stock: The status of criminological theory* 15 (2006), 101–123.
[5] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. 2017. Adversarial Patch. *ArXiv e-prints* (Dec. 2017). arXiv:cs.CV/1712.09665
[6] Joseph Bullington. 2005. 'Affective' Computing and Emotion Recognition Systems: The Future of Biometric Surveillance?. In *Proceedings of the 2Nd Annual Conference on Information Security Curriculum Development (InfoSecCD '05)*. ACM, New York, NY, USA, 95–99. https://doi.org/10.1145/1107622.1107644
[7] Martin Cooney, Shuichi Nishio, and Hiroshi Ishiguro. 2015. Affectionate interaction with a small humanoid robot capable of recognizing social touch behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2015), 19.
[8] Martin D Cooney, Shuichi Nishio, and Hiroshi Ishiguro. 2015. Importance of touch for conveying affection in a multimodal interaction with a small humanoid robot. *International Journal of Humanoid Robotics* 12, 01 (2015), 1550002.
[9] Roddy Cowie. 2015. Ethical Issues in Affective Computing. (2015). http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199942237.001.0001/oxfordhb-9780199942237-e-006
[10] Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* 32, 1 (1969), 88–106.
[11] Thomas Grote and Oliver Korn. 2017. Risks and Potentials of Affective Computing. An Interdisciplinary View on the ACM Code of Ethics. In *CHI 2017 workshop on Ethical Encounters in HCI*.
[12] Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review* 10, 3 (2006), 252–264.
[13] Jeff T Larsen, A Peter McGraw, and John T Cacioppo. 2001. Can people feel happy and sad at the same time? *Journal of personality and social psychology* 81, 4 (2001), 684.
[14] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 443–449. https://doi.org/10.1145/2818346.2830593
[15] Kohei Ogawa, Koichi Taura, and Hiroshi Ishiguro. 2012. Possibilities of androids as poetry-reciting agent. In *RO-MAN, 2012 IEEE*. IEEE, 565–570.
[16] Masaru Ohkubo, Miki Yamamura, Hiroko Uchiyama, and Takuya Nojima. 2014. Breathing clothes: artworks using the hairlytop interface. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*. ACM, 39.
[17] Catherine O'Neill. 2016. Weapons of Math Destruction. *How Big Data Increases Inequality and Threatens Democracy* (2016).
[18] Rebeccah Pailes-Friedman. 2015. BioWear: a kinetic accessory that communicates emotions through wearable technology. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 627–633.
[19] Rosalind W Picard. 1995. Affective Computing-MIT Media Laboratory Perceptual Computing Section Technical Report No. 321. *Cambridge, MA* 2139 (1995).
[20] Rosalind W. Picard and Jennifer Healey. 1997. Affective wearables. *Personal Technologies* 1, 4 (01 Dec 1997), 231–240. https://doi.org/10.1007/BF01682026
[21] Isabella Poggi, Francesca D'Errico, and Laura Vincze. 2010. Types of Nods. The Polysemy of a Social Signal.. In *LREC*.
[22] Anas Samara, Maria Luiza Recena Menezes, and Leo Galway. 2016. Feature Extraction for Emotion Recognition and Modelling Using Neurophysiological Data. In *Ubiquitous Computing and Communications and 2016 International Symposium on Cyberspace and Security (IUCC-CSS), International Conference on*. IEEE, 138–144.
[23] Erica Scharrer. 2008. Media exposure and sensitivity to violence in news reports: Evidence of desensitization? *Journalism & Mass Communication Quarterly* 85, 2 (2008), 291–310.
[24] Amanda Sharkey and Natalie Wood. 2014. The Paro seal robot: demeaning or enabling. In *Proceedings of AISB*, Vol. 36.
[25] Alessandro Vinciarelli, Hugues Salamin, and Maja Pantic. 2009. Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, 42–49.