Selected Papers of #AoIR2019:
The 20th Annual Conference of the
Association of Internet Researchers
Brisbane, Australia / 2-5 October 2019

THE ETHICS OF EMOTION IN AI SYSTEMS

Luke Stark
Microsoft Research

Jesse Hoey
University of Waterloo

Introduction

Computational analyses of psychological and behavioral data pertaining to human emotional expression have a surprisingly long history (Dror, 2009), an underappreciated diversity of methods (Boehner, DePaula, Dourish, & Sengers, 2007; Schröder, Hoey, & Rogers, 2016), and an increasingly critical role in social machine learning (ML) and artificial intelligence (AI) applications (Davies, 2017; Stark, 2018). Speculative and science fiction is replete with questions regarding the emotional lives of artificial beings. Yet contemporary, quotidian, narrow AI/ML technologies are most frequently used by social media platforms for modeling and predicting human emotional expression as signals of interpersonal interaction and personal preference (Bucher, 2016). These analytics are now being deployed in domains as varied as hiring, personal health and wellness, customer service, border security, and, as part of a broader category of "digital phenotypes" (Jain, Powers, Hawkins, & Brownstein, 2015), in digitally mediated mental health treatment and prevention (Brandt & Stark, 2018). While the ethical and social impacts of ML/AI systems have of late become major topics of both public discussion and academic debate (Barocas & Selbst, 2016; boyd & Crawford, 2012; Johnson, 2018), the ethical dimensions of AI/ML analytics for emotional expression have been under-theorized in these conversations.

Emotional Expression and Machine Analysis

Here, we make several contributions to the emerging critical literature on the ethics of AI/ML systems. The paper first taxonomizes systems for tracking and modeling human emotional expression through the types of data they collect (D. Kim, Frank, & Kim, 2014). Many such systems seek to model emotional interactions through "digital

Suggested Citation (APA): Stark, L and Hoey, J. (2019, October 2-5). *The Ethics of Emotion in AI Systems.* Paper presented at AoIR 2019: The 20th Annual Conference of the Association of Internet Researchers. Brisbane, Australia: AoIR. Retrieved from http://spir.aoir.org.

phenotyping," the analysis of biosignals, including optical data (such as facial movement, gait, or infrared emanation); audio data (such as the vocal tone and cadence) (Jain et al., 2015); haptic and physiological data (such as skin conductivity, blood flow, and body velocity) (Picard, 2000); others examine semantic signifiers of emotional expression, including written words, graphic means such as emoji and emoticons, and other representations of human feeling) (Alashri et al., 2016). As part of our review, we also describe a novel method for modeling human social and emotional interactions computationally: Bayesian Affect Control Theory or BayesACT (Schröder et al., 2016). Developed out of collaborations between sociologists and computer scientists, it combines affect control theory, a form of structural symbolic interactionism (Lively & Heise, 2004), with Bayesian probabilistic decision theory.

We draw on this taxonomy to consider the ethical challenges posed by ML/AI human emotion analysis (Desmet & Roeser, 2015). One such challenge is that data on human emotional expression imperfectly reflect human emotions themselves – complex, culturally specific yet broadly recognizable signals of core human subjectivity. As such, these data are representative of multiple elements of human subjective experience, both denoted imperfectly and quantified partially. Models extrapolated from these data claim a descriptive power that is also a prescriptive power, forcing individuals to adjust their own attitudes to conform to an "objective" measure of emotional expression that is in fact partial, artificial, and potentially detached from lived experience.

A second ethical challenge concerns the ways in which subjectivity and the status of the human individual are inextricably tied to longstanding theories of ethical decision-making. Given the ways emotion and other intuitive processes are understood to play a central role in ethical and moral judgments (Prinz, 2004), the digital remediation of emotional expression has the potential to shift normative frameworks for decision-making based on the values of technology firms, not of individuals as users and citizens.

Emotions and Ethical Tensions in AI

Based on these analyses, we elucidate how a focus on emotional expression as a component of ML systems highlights conceptual tensions within current AI/ML ethics discourses. These include, first, how a focus on emotional expression as a component of AI/ML analysis demonstrates the broader tendency of AI/ML research to perform *de facto* human subject research without attendant awareness of or attention to the ethical complexities of such experimentation (Brandt & Stark, 2018). Second, attempts to quantify and standardize measures of emotional expression illustrate the conceptual difficulty in constituting shared ethical or normative guidelines around AI/ML systems because of what Nagel terms "the fragmentation of value" (Nagel, 1979) and the challenge of developing widely shared intersubjective norms.

Finally, the analytics of emotional expression highlight human emotion's centrality not just to ethical AI/ML systems, but also to these system's broader mediating effects on social and political community and cohesion through their everyday use. We ground the fourth and last portion of the paper – preliminary recommendations around both policy and design -- in recent applied work on virtual agents in two areas: cognitive assistive

technologies for persons with dementia that are functionally and emotionally aligned with their target users, and facilitator agents in social networks aimed at promoting efficient and inclusive group processes. We draw on these case studies to consider the ethical implications of emotional AI in practice.

References

Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., & Desouza, K. C. (2016). An analysis of sentiments on facebook during the 2016 U.S. presidential election (pp. 795–802). Presented at the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE. http://doi.org/10.1109/ASONAM.2016.7752329

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671–732. http://doi.org/10.15779/Z38BG31

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How Emotion is Made and Measured. *International Journal of Human-Computer Studies*, *65*, 275–291. http://doi.org/10.1016/j.ijhcs.2006.11.016

boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

Brandt, M., & Stark, L. (2018). Exploring Digital Interventions in Mental Health: A Roadmap. In A. Shaw & D. T. Scott (Eds.), *Interventions* (pp. 167–182). Bern, Switzerland: Peter Lang.

Bucher, T. (2016). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. http://doi.org/10.1080/1369118X.2016.1154086

Davies, W. (2017). How are we now? Real-time mood-monitoring as valuation, *10*(1), 34–48. http://doi.org/10.1080/17530350.2016.1258000

Desmet, P. M. A., & Roeser, S. (2015). Emotions in Design for Values. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design* (pp. 203–219). Dordrecht: Springer. http://doi.org/10.1007/978-94-007-6970-0_6

Dror, O. E. (2009). Afterword: A Reflection on Feelings and the History of Science. *Isis*, *100*(4), 848–851.

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Publishing Group*, *33*(5), 462–463. http://doi.org/10.1038/nbt.3223

Johnson, D. G. (2018). AI, agency and responsibility: the VW fraud case and beyond. *AI & Society*, *0*(0), 0–0. http://doi.org/10.1007/s00146-017-0781-9

Kim, D., Frank, M. G., & Kim, S. T. (2014). Emotional display behavior in different forms of Computer Mediated Communication. *Computers in Human Behavior*, *30*, 222–229. http://doi.org/10.1016/j.chb.2013.09.001

Lively, K. J., & Heise, D. R. (2004). Sociological Realms of Emotional Experience. *American Journal of Sociology*, *109*(5), 1109–1136. http://doi.org/10.1086/381915

Nagel, T. (1979). The Fragmentation of Value. In *Mortal Questions* (pp. 128–141). Cambridge, UK: Cambridge University Press.

Picard, R. W. (2000). Affective Computing. Cambridge, MA: The MIT Press.

Prinz, J. J. (2004). Introduction: Piecing Passions Apart. In *Gut Reactions: A Perceptual*

Theory of Emotion (pp. 1–11). Oxford, UK: Oxford University Press.
Schröder, T., Hoey, J., & Rogers, K. B. (2016). Modeling Dynamic Identities and Uncertainty in Social Interactions. *American Sociological Review*, *81*(4), 828–855. http://doi.org/10.1177/0003122416650963
Stark, L. (2018). Algorithmic Psychometrics and the Scalable Subject. *Social Studies of Science*, *48*(2), 204–231.