A review of Elizabeth Wilson's *Affect and Artificial Intelligence*

Ho Manh Tung

Ritsumeikan Asia Pacific University

Beppu, Oita, Japan

November 1, 2020

In *Affect and Artificial Intelligence,* Elizabeth A. Wilson analyzes "the early affective networks within which mid-twentieths-century computational devices were anticipated and built" and argues that emotion lies that the foundation of building smart machines. She also seeks to make the computational objects more engaging for humanities scholars by studying the affective components (curiosity, contempt, anger, sadness, etc.) of the works of the three brilliant early computer scientists and mathematicians, Alan Turing (1912-1954), Walter Pitts (1912-1969), and Joseph Weizenbaum (1923-2008), in developing their respective machines.

Krzywoszynska (2012) succinctly points out that the most interesting argument in the book is how AI, as a concept, becomes constructed in an emotionless and impassive way. Indeed, very early on in the book, Wilson provides a juxtaposition of two models of thinking about AI by quoting from Alan Turing's famous *Computational Machinery and Intelligence* paper (1950): the chess-playing model and the child-like model. Through numerous biographical examples, Wilson convincingly argues that most commentaries about AI are dominated by the chess-playing model, which is all about abstraction, number-crunching, and symbols-manipulation. And this is the result of our reluctance to admit the role of emotion in smart machines' life.

There are two problems pointed out by the Stahnisch (2011) and Krzywoszynska (2012). First, Wilson's (2010) book suffers the most from the lack of a concluding remark. And second, Krzywoszynska (2012) and Stanisch (2011) point out that Wilson's (2010) book is lightly referenced, and its sheer brevity does not match the formidable size of its topic. I agree with these points but add three points that I think can make the argument stronger in the book.

First, it is a glaring missed opportunity that Wilson did not engage with the literature on affective computing given Rosalind Picard had invented the field since the late 1990s (Picard, 2000). This field aims to teach a machine the ability to read, track, classify, and even express human emotions. Perhaps, it is the most direct attempt to endowing the "*ability to feel*" to a machine. By 2003, this field achieved the accuracy rate of recognizing the so-called eight *basic emotions* postulated by Ekman (1999) of 81% (Picard, 2003). In recent years, companies that sell AI products based on this approach claim a 95% accuracy rate in recognizing emotion. However, this is very much up for debate, given the rise of new theories and empirical results in emotion study (Heaven, 2020).

Based on Wilson's arguments in the book, I think she would argue even for affective computing, the assumption of the eventual success of the brute-force calculation approach on top of big data is too strong. This way of thinking is often called "*behaviorism*," "the doctrine that psychology can only, and should only be, the science of behavior, not of minds; that it can only measure and predict relationships between people's external circumstances ('stimuli') and their observed behaviors ('responses')" (Deustch, 2011, pp.157-158).

I believe different theories of emotions, i.e., different modes of explaining how emotions are formed, expressed, and inferred, will have different implications for improving the current machine learning approaches to emotions. The current behaviorist approach might have certain successes with the explosion of big data and computational power, yet, it is prone to the algorithmic designers as well as biases in the data. Two examples come to mind. Several recent studies on the accuracy of reading emotions of different races expose that current machine algorithms are more likely to ascribe that black people are angrier (McStay, 2018; Rhue, 2019). And in cases where the initial dataset is dominated by one sex, such as among the police force, using the current AI approach to track emotions of female police officers might lead to dangerous errors (Purdy, Zealley, & Maseli, 2019).

Another issue I wish Wilson explores is the *acculturation of emotion and affect*, i.e., how people learn and unlearn their emotional reaction and how that might influence AI developers. Wilson might have successfully conveyed that all of the pioneers of AI think we need to understand emotion better to build smart AI, or there are ways in which emotions have leaky effects in their works. Yet, the book never brought up the issue of changing emotions. This is an important issue as our emotional lives are being disrupted drastically with social media's hyperconnectivity. The emergence of so many subcultures and ideologies consequently gives rise to new modes of emoting. This issue of acculturation (Vuong et al., 2018a; Vuong et al., 2020; Vuong & Napier, 2015) will keep coming up, and the current behaviorist approach in AI about emotions is inadequate to explain the ebbs and flows of our emotional lives as well as telling us what the appropriate emotions to feel in certain situations are.

Finally, it seems that the book's biographical historical data could benefit from a more methodological approach to organizing the data. The book talks about emotion and affects in general, but it could have categorized emotions better. There are reactive emotions (anger, disgust, joy) and meta-cognitive ones (doubt, curiosity, wonder, awe). The meta-cognitive emotions help us evaluate our normal emotional responses to situations in life. In other words, it helps us take a step back and put our emotional lives in a grander perspective.

Or there can be three levels of analysis: emotions and affects, intuition, and reason, in which intuition provides a bridge from emotion to reason. Our intuition is a set of basic core values, axioms, or rule of thumb, that guides our behavior, emotion, and cognition. Our intuition can be updated and changed with learning. Once it is changed, we might have a different set of emotional responses than our original one. For example, once we learn how to fight, a certain way of standing starts to feel unnatural and dangerous. In Wilson's book, the affective aspects of the three AI pioneers can be analyzed in terms of how they come to have certain intuitions about a machine that can think and feel. This is also an angle where the perspective of *acculturation of emotions* can have a role to play.

These two examples of how biographical data about emotions can bring about a more systematic analysis (Vuong et al., 2018b) of the emotional landscapes of the AI pioneers and how they affect their work. Besides, an orderly organization of data could help shape the narrative arcs of Wilson's book as well as general concluding remark too. Beyond the book, I think it can even help to improve the replicability in humanities (Peels & Bouter, 2018; Vuong, 2020).

Despite such shortcomings, I think Wilson's book is of great interest to readers of the biographical history of computer science and, more importantly, humanities scholars who would like to explore how emotions influence the works of early pioneers amongst AI theoreticians and engineers.

References

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. London, UK: Penguin.

Ekman, P. (1999). Basic Emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion*. Sussex, UK: John Wiley & Sons.

Heaven, D. (2020). Why faces don't always tell the truth about feelings. *Nature, 578,* 502-504.

Krzywoszynska, A. (2012). A review of Elizabeth Wilson's Affect and Artificial Intelligence. *Emotion, Space and Society, 5*(4), 284. doi:https://doi.org/10.1016/j.emospa.2012.05.002

McStay, A. (2018). *Emotional AI: The rise of empathic media*. London: Sage.

Peels, R., & Bouter, L. (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications, 4*(1), 95.

Picard, R. W. (2000). *Affective computing*. Cambridge, Massachusetts: MIT Press.

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies, 59*(1), 55-64.

Purdy, M., Zealley, J., & Maseli, O. (2019). The risks of using AI to interpret human emotions. *Harvard Business Review*. Retrieved from https://hbr.org/2019/11/the-risks-of-using-ai-to-interpret-human-emotions

Rhue, L. (2019). Anchored to Bias: How AI-Human Scoring Can Induce and Reduce Bias Due to the Anchoring Effect. *Available at SSRN 3492129*.

Stahnisch, F. W. (2011). A review of Elizabeth A. Wilson's Affect and Artificial Intelligence. *Isis, 102*(4), 818-819.

Vuong, Q. H. (2020). Reform retractions to make them more transparent. *Nature, 582*(7811), 149.

Vuong, Q. H., & Napier, N. K. (2015). Acculturation and global mindsponge: an emerging market perspective. *International Journal of Intercultural Relations, 49*, 354-367.

Vuong, Q.-H., Bui, Q.-K., La, V.-P., et al. (2018a). Cultural additivity: behavioural insights from the interaction of Confucianism, Buddhism and Taoism in folktales. *Palgrave Communications, 4*(1), 143.

Vuong, Q.-H., Ho, M.-T., Nguyen, H.-K. T., et al. (2020). On how religions could accidentally incite lies and violence: folktales as a cultural transmitter. *Palgrave Communications, 6*(1), 82.

Vuong, Q.-H., La, V.-P., Vuong, T.-T., Ho, M.-T., et al. (2018b). An open database of productivity in Vietnam's social sciences and humanities for public use. *Scientific Data, 5*(1), 180188.

Wilson, E. A. (2011). *Affect and artificial intelligence*. Washington: University of Washington Press.