The Computational and Aesthetic Foundations of Artificial Empathy

Megha Srivastava
Stanford University

Abstract
Many pioneers in technology, from Elon Musk to Bill Gates, have expressed caution regarding the advance of artificial intelligence (AI) technologies. Socially, a “doomsday” scenario of AI dominating humanity is entrenched in popular culture, ranging from movies to political worries of the role of AI in military organizations. However, this fear is based on an assumption that intelligence is the only facet of the human experience that researchers seek to replicate. The AI agents portrayed as harmful are those that can understand their surroundings, store large amounts of data, possess advanced knowledge, and have the capability to coordinate, kill, and destroy far beyond humans. On the other hand, these portrayals of AI do not show artificial empathy, compassion, and love that even approach the same level of humans. In this paper, I plan to investigate the scientific and computational foundations of artificial empathy, and the aesthetic methods of making it appear genuine. Using an interdisciplinary approach, I aim to show that artificial empathy is not only possible, but also necessary for a positive future for artificial intelligence.
Introduction

“With artificial intelligence we are summoning the demon.”
– Elon Musk (TechCrunch, 2014)

At first glance, one would not expect such caution towards artificial intelligence (AI) to be expressed by Elon Musk, the celebrated engineer pushing technology’s boundaries with companies Tesla and SpaceX. Musk innovates at the brink of engineering advances, embracing automation and self-driving technologies (which are, in fact, core components of AI). Perhaps what Musk fears is not AI itself, but the idea of AI overpowering humanity—a thought echoed by Bill Gates and Stephen Hawking (Holley, 2015). The “demon” within AI is therefore not simply the technology, but the idea of it developing to be “smarter” than humans or misused by those in powerful positions.

The “doomsday” scenario of AI dominating humanity is entrenched in popular culture. Movies such as The Terminator, Blade Runner, and Ex Machina all show AI agents as possible threats to humans (Cameron, 1984; Garland, 2015; Scott, 1982). Popular themes such as “an army of robots” and terrorist organizations using AI to gain power understandably portray a terrifying future for AI. In fact, philosopher Nick Bostrom’s book, Superintelligence, proposes that machine intelligence may soon replace humans as the prevailing lifeform on our planet (Bostrom, 2014).

However, this fear is based on the assumption that intelligence is the only facet of the human experience that researchers seek to replicate. The AI agents portrayed as harmful are those that can understand their surroundings, store large amounts of data, and possess advanced knowledge. They have the capability to coordinate, kill, and destroy far beyond humans. Yet, these portrayals of AI do not exhibit empathy, compassion, and love that approach the same level of humans. In contrast, Disney’s WALL-E from WALL-E and Marvel Comic’s Baymax from Big Hero 6 are notable examples of artificial beings that demonstrate positive emotions. Neither WALL-E nor Baymax are considered harmful, let alone representations of the dangers of AI (Stanton, 2008; Hall, 2014). Thus, we fear AI agents that have superior intelligence, but lack emotional understanding.

This idea is akin to imagining a society where every human simply acts intelligently out of self-interest, doing what is necessary to live successfully without perceiving others’ pain. Philosopher Thomas Hobbes imagined such a society in Leviathan, describing a “state of nature” where each person acts with the interest of preserving his or her own life (Hobbes, 1998). Such a society is perceived as dangerous—Hobbes refers to this life as “brutish”—and even modern-day journalists refer to criminals as people who have a “lack of empathy” (Hobbes, 1998; Bloom, 2016). Human’s fear of AI stems from imagining the future of AI paralleling a society that is void of empathy.
Therefore, a solution to preventing the dangers of harmful AI taking power could be to embrace “artificial empathy.” Indeed, Scott Button, the CEO of the social-advertising tech company Unruly, states, “We need to develop empathetic, value-driven AI as a priority if we want the human species to survive and thrive in the 22nd century” (The Atlantic, 2015). Providing artificial beings with the incentive to avoid actions that harm others will allow us to focus and invest in AI’s many useful applications. However, there are two main arguments against artificial empathy. First is its feasibility: some believe that empathy is a uniquely human experience that is incredibly complex and impossible to replicate (Taylor, 2012). Second is artificial empathy’s practicality: with the knowledge of an automaton being artificial, how can it properly convince humans to believe in its empathy?

The debate about artificial empathy’s feasibility continues to change with recent advances in neuroscience. Every time a neuroscientist discovers regions of the brain or neural circuits responsible for fundamental human behaviors, it becomes easier for a computer scientist to replicate them. It is worth investigating, therefore, whether we can break empathy down into its basic principles and algorithmic steps.

Nevertheless, even if computational empathy is feasible, in order for artificial empathy to truly be practical, humans need to be convinced that an automaton’s empathy is genuine. This primarily depends on the psychological and aesthetic factors of AI. Given that empathy requires emotionally understanding someone else’s pain, it is possibly easier to believe that those we view as “inferior” have experienced, and can therefore understand, our pain more than those we view as stronger and more powerful than us.

In this essay, I plan to investigate the scientific and computational foundations of artificial empathy and the aesthetic methods of making it appear genuine. Using neuroscience and computational discoveries, I will:

1. Discuss the foundations and future work for creating a computational model of empathy.
2. Draw on psychological theories of the aesthetics of automatons, focusing on those that evoke perceptions of emotions and empathy.
3. Use examples of current robots that demonstrate the immense benefit of artificial empathy, especially in the field of healthcare.

Using an interdisciplinary approach, I plan to show that artificial empathy is not only possible, but also necessary for a positive future for AI.

Computational Model
Empathy is typically expressed when an experience is negative, such as understanding another person’s pain from an illness or grief due to a tragic
Therefore, empathy requires a person to first be able to detect another person’s suffering. Furthermore, Professor Mark Davis, from the University of Texas at Austin, describes empathy as having two parts: the observable, visceral response and the cognitive understanding of another perspective (Davis, 115). Thus, another component of empathy is the reaction of the subject, whether through facial expressions or verbal replies. Finally, in order to properly share another person’s experiences, one must be able to link another person’s suffering with his or her own experiences in life. This requires memory, either by remembering a similar event or constructing the other person’s situation from different personal memories. Thus, at a high level, empathy can possibly be modeled with three main processes:

**Empathy Model:**

1. *Detect a person’s suffering from verbal description, voice tone, or facial expression.*

2. *Calculate similarities between a person’s suffering and one’s own memories.*

3. *Create an “appropriate” response, both vocally and through observable expressions.*

**Step 1: Detection of Pain or Sadness**

In order for us to detect emotions such as sadness, we use skills developed throughout our lifetime. It is widely known that newborns do not process information at the same level as adults—speech, writing, and nuanced facial expressions grow increasingly more complex as humans develop (Gale, 2003). This “learning process” is what computer scientists seek to mimic when developing AI via machine learning. The basic premise of machine learning is that the world provides an immense amount of data that allow us to make predictions on new inputs. We “learn” by being trained with examples, such as a mother showing her baby a picture of an apple and pronouncing “apple” multiple times. Over time, the baby learns to associate the picture with the word, eventually gaining the ability to see the fruit’s shape and form and predict its pronunciation as “apple.” Within the scope of emotions, this suggests that we grow up learning which words, facial cues, and speech sounds are more likely to be associated with a sad emotion versus a happy emotion. The ability to detect a scene as “sad” is simply the ability to remember that similar scenes in the past have commonly meant “sad.”

---

1. Although I focus on this specific aspect of empathy, most definitions of empathy encompass feelings beyond negative emotions.

2. Interestingly, it is known for humans to “see faces” in abstract features. Paradeila is the psychological phenomenon behind seeing a “man in the moon” or facial features in cars—adding justification to the feasibility of abstract features replacing realist faces.
Linguistically, this principle is obvious—a person only speaking English likely does not know that the Norwegian word *kreft* means “cancer.” Therefore, they would not react as emotionally towards seeing or hearing “*kreft*” as they would “cancer” until they learned the actual meaning. Computationally understanding the meanings and corresponding emotions of sentences is the heart of the field of Natural Language Understanding (NLU). NLU researchers use machine learning to build models that can predict the sentiment of a sentence based on the words used and grammatical structure. A classic experiment in NLU is determining the sentiment of movie reviews. Researchers Bo Pang and Lillian Lee from Cornell University were able to train a computer to determine whether a movie review was positive or negative with 81% accuracy. Words such as “bad,” “stupid,” and “waste” were, understandably, big indicators of a negative review. Clearly, then, it is possible for a computer to determine sentiment in a text—one can imagine extending Pang and Lee’s experiment to consider emotions, thus allowing us to computationally detect sadness.

We can use a similar machine learning approach to visually determine negative emotions. By providing a computer with several examples of faces that are considered “sad” and faces that are not considered “sad,” we can train the computer to use mathematic tools to discover what features of a face are more present in faces tagged as “sad.” As one would expect, this requires a large amount of data that parallels the amount of visual data humans are exposed to throughout a lifetime. Various databases, such as Stanford Professor Fei-Fei Li’s ImageNet and Carnegie Mellon Professor Takeo Kanade’s Expression Database provide a large volume of facial images for this very purpose. Surprisingly, computational visual emotion recognition has done incredibly well in recent years. In fact, a *Current Biology* report showed that computers performed better than humans in distinguishing faked pain and genuine pain in facial expressions (Bartlett, 2014). Therefore, through computational techniques that mimic the way humans learn information, we can train a computer to detect sadness and pain in humans.

Step 2: Search for Similarities with Memories

Simply detecting someone else's pain, however, is not enough to constitute empathy. Part of empathy requires sharing another's feelings. But truly understanding someone's experience requires a degree of familiarity with the experience itself—a component of our memory. In fact, Dr. Ulrich Wagner from the University of Münster, Germany showed that memory performance is correlated with the perspective-taking component of empathy, and that emotional empathy is linked to social memories (Wagner, 2015). This fascinating result demonstrates how foundational our memories are for empathy. Clearly, memory is not a uniquely human phenomenon. From jays that remember when the food in their caches spoil to gorillas that remember which trainers provide them with certain fruits,
anthropologists have been able to demonstrate memory’s so-called “mental time-travel” in non-humans (Trivedi, 2003). However, given that a computer’s hard-drive is clearly not equivalent to reliving an experience, how can we computerize memory for artificial empathy?

One possible approach for automating memory in empathy is inspired by the discovery of “mirror neurons.” These are neurons that fire both when we act and when we observe someone performing the same action. For example, the same neurons that fire when a boy is being bullied may fire when a boy is observing bullying; however, research has yet to expand to such broad, social examples. UCLA Psychiatry professor Marco Iacoboni has proposed that mirror neurons can strongly develop empathy (Iacoboni, 2009). Furthermore, it is widely believed that our memories are tied with sensory cues, such as the weather during the time of the remembered event. Our ability to remember an event, therefore, increases if similar cues are present. So, if empathy simply boils down to observing a set of cues that triggers a memory via neurons, then we can imagine providing a computer with a set of possible cues that can “trigger” a memory. For example, if a computer detects an insulting word, it can output: “I remember also being called ______ and feeling really sad—don’t worry, it will get better,” as if the computer actually held a memory that was triggered when sensing a similar event.

Although it may appear that remembering emotional events is a general, vague process, scientists have isolated a region of the brain specifically in charge of emotional memory: the amygdala. The amygdala is an almond-shaped structure in our brain critical for processing the emotions associated with a memory (Buchanan, 2008). In fact, fear conditioning, a method of training to remember to avoid certain actions that cause fear, is unsuccessful with rats with lesions impairing their amygdala (Phillips, 1992). These rats are probably unable to empathize with rats that perform similar actions since they cannot recall the unpleasant feeling of fear. The discovery of a region responsible for emotional memory, and therefore a part of empathy, is incredibly exciting. By studying the structure of the amygdala, computer scientists can learn how to simulate it. Already, engineers have made significant advances in artificial memory: in their headline-making paper, “Creating a False Memory in the Hippocampus,” MIT researchers were able to implant false memories in mice (Ramirez, 2013). Indeed, scientists have been able to design memories that living beings think are real. One can imagine, therefore, the exhaustive range of possibilities with implanting memories in an AI.

It is increasingly clear how our complex emotions and vivid memories rely simply on the networks of neurons inside our brain. Recent advances in programming and the development of artificial neural networks show it can be possible to replicate these processes. Consider a computer with an implanted memory closely replicating the human amygdala—it would
then have the fascinating ability to connect with someone’s pain to elicit an empathetic response.

Step 3: Formulating an Appropriate Response

The most fundamental way AI responds to humans is through using a set of canned responses. Apple’s Siri serves as an example: if we ask Siri to “tell me a joke,” she responds with one of at most a dozen responses designed by Apple. While such a hard-coded way of responding appears incredibly rigid, how truly different is it from the way humans respond? As a society, we often suggest rules and algorithm-like structure to empathetic responses. If someone announces a relative’s death, the typical set of responses contains versions of “I’m sorry.” A cheerful or joking response would be considered incongruous and rude. Thus, we already have rules for what constitutes an appropriate, empathetic response. These guidelines, such as instructing the computer to always begin with a set of responses such as “I’m sorry” and “I hope you feel better,” and providing a list of possible filler words such as “like” and “um” to make the responses appear more natural, can easily be reduced to rules in an algorithm.

However, to even further exhibit an empathetic response, a computer can draw in components from the memory it linked. For example, the 1972 chat-bot PARRY created by psychiatrist Kenneth Colby was designed to “believe” that it was a paranoia patient running away from the mafia. Whenever the human interacting with PARRY said something that PARRY could not understand, PARRY immediately referenced the mafia and the different facts that chat-bot was implanted with. These “seeds” of topics influenced PARRY’s digital responses to such an extent that more than 50% of psychiatrists believed PARRY was a real human with an actual paranoia disorder (Colby, 1971). The rich details of discussing a memory can make an emotional response feel more genuine.

In order to fully understand the model proposed above, consider an example: A woman tells an AI, “Someone robbed my house and stole all the jewelry.” First, the AI could detect that her facial expression is statistically most likely “upset” and that the words “robbed” and “stole” typically describe a worrying situation warranting an empathetic response. Then, perhaps, the word “robbed” could trigger a “false” memory in the AI of a time it was “robbed” of its money from its apartment 5 years ago. The AI could then respond, following societal expectations, “I'm so sorry to hear that. Someone robbed my money from my apartment 5 years ago, and I remember how stressful it was. Please let me know if you want help.” Thus, the computer would have exhibited artificial empathy.

However, this model may not be convincing for the simple fact that knowledge of an AI’s artificiality prevents us from believing the AI can be truly empathetic. We therefore must consider how to distract humans from such knowledge, and truly believe in artificial empathy’s genuineness.
Conclusion: The Aesthetics of Artificial Empathy

Appearance, often the first property a human notices from any being, artificial or not, can drastically influence our perception. Consider Western pop-culture representations of AI: they range from the frightening Terminator and eerie Ava in Ex Machina to the adorable, whirring WALL-E and the bumbling, loveable Baymax from Big Hero 6. All could be considered AI—WALL-E’s persistent curiosity about his world exhibits astounding intelligence rivaling the Terminator’s tracking abilities as an assassin. However, public impressions could not be more different. An IMDb reviewer describes WALL-E by saying “I’ve never felt so much emotion for one character,” with another adding, “WALL-E, as a character, has dimension, personality, and heart,” despite the fact that the first dialogue occurs around 30 minutes into the movie (IMDb, 2008). The lack of speech by WALL-E demonstrates that his actions, movements, and sound effects, despite all possibly being programmed, are enough to convince viewers of his sentience. However, several pop-culture articles about AI feature images of “Skynet” from The Terminator, positing the question: “How Scared are you of AI?” (Kooser, 2014; Hern, 2015). In fact, Ex Machina, featuring a humanoid that murdered, tricked, and escaped her facility, is popularly considered to be the most realistic representation of AI’s future (IMDb, 2016). Thus, we draw on the more violent examples when discussing the future of AI, and we can look towards theories of aesthetics to understand how this relates to artificial empathy.

The theory of the “Uncanny Valley,” proposed by Japanese robotics professor Masahiro Mori, hypothesizes that there is a large negative dip in favorability towards artificial beings if their features almost, but not completely, represent humans (Mori, 2012). Moreover, favorability peaks when either the features look exceptionally like humans, or resemble stuffed animals. If we accept this theory, then it is not surprising why WALL-E and Baymax have such a positive appeal to the public—abstract features prevent these robots from eliciting an “uncanny” sensation. Indeed, Mori states: “I predict that it is possible to create a safe level of affinity by deliberately pursuing a nonhuman design.”

Furthermore, a suggested justification for the “Uncanny Valley” has been the idea of “mortality salience,” where a robot with uncanny, human-like features is more likely to remind humans of death’s inevitability (Mori, 2012). The combination of human characteristics with the knowledge of a robot’s lifelessness eerily creates an image that is almost corpse-like. Designers who seek to make AI resemble us as closely as possible, then, largely enhance humans’ fear of death. A whole spectrum of beliefs—ranging from social customs to religious ways of treating death—makes it unlikely to associate an unnerving sense of death with a capacity for empathy.

Stanford professor Sianne Ngai further explores this relationship between realistic features and fear. In her paper “The Cuteness of the
Avant-Garde,” Ngai argues that “cute” objects are formally described as having “smallness, compactness, softness, simplicity, and pliancy” in their features (Ngai, 2005). Moreover, these physical characteristics are attributed to “helplessness, pitifulness, and despondency.” Ngai further compares Japanese design’s tendency towards cuteness with American design’s realistic focus. She proposes that the Japanese phenomenon of kawaii, meaning “cute,” reflected the feeling of helplessness in post-WWII Japan. On the other hand, America’s world dominance and strength likely discouraged any preference for such a fascination with cute toys.

Thus, WALL-E and Baymax more likely resemble the Japanese approach to AI design—in fact, Baymax is portrayed by Disney to be from Japan. Their cute features—tiny voices, simple answers, and soft demeanors—allow us to view the AI as dependent on humans. Thus, despite an AI’s advanced computational power, if it appears to “need” us, then we are less likely to fear the AI’s presence.

Aesthetically, therefore, it appears optimal to present artificial empathy with cute, abstract features. This would allow us to believe that the AI cannot subjugate us and therefore would be more likely to understand and relate to pain. Already, such a design is being used in modern day robotic technologies. The Japanese robot named Pepper, created in collaboration between the Japanese mobile company SoftBank and French robotics company Aldeberan, sold out minutes after going on sale. Glistening white with a tiny, female voice and cute expression, Pepper was marketed as an “emotional robot,” with the ability to scan and detect human emotions. Already, Pepper has been expected to help students struggling in school to better communicate in class (Gray, 2016).

Artificial empathy can cause an enormous impact in other social scenarios as well. Dr. Maja Mataric, from USC, currently develops socially assistive robots for healthcare services—an application of AI that would benefit tremendously from an AI’s ability to empathize with suffering patients (Fasola & Mataric, 2013). For example, the Japanese seal robot Paro has shown remarkable success in comforting dementia patients (Johnston, 2015). Critics do question, however, the moral reasons behind giving robots the responsibility of empathizing with patients instead of humans. It can be viewed as unethical to provide human patients artificial empathy that doesn’t yet meet the level of human empathy.

However, the reliability of artificial empathy can provide the constant companionship and understanding that is not always present in humans due to our own variability in life. Sociologist Arlie Hochschild even proposed the idea of “emotional labor,” arguing that the expected empathy from nurses, counselors, and others is a form of work (Hochschild, 1983). Humans cannot always provide necessary care at all times, and the goal of researchers in the field is to use AI to “provide much-needed care where it

---

2 Interestingly, it is known for humans to “see faces” in abstract features. Paradeila is the psychological phenomenon behind seeing a “man in the moon” or facial features in cars—adding justification to the feasibility of abstract features replacing realist faces.
is needed” (Tapus, 2006). Moreover, areas such as mental health and suicidal ideation can often be difficult for humans to detect in one another—a constant companion capable of empathy could help provide an outlet for humans that they may otherwise not have. Finally, human empathy is not consistent—our own mood, capacity of remembering past events in a given moment, and opinion of others can dramatically affect our ability to provide and manner of providing empathy. For example, the recent “Black Lives Matter” activism campaign has highlighted instances where humans are unable to provide empathy due to their own experiences or lack of experience with racism (Ransby, 2015). Artificial empathy is not swayed by fluctuations or its own biases—it can be programmed to give helpful, supportive empathy exactly the way humans desire. ³

Thus, creating an AI that computationally performs in all the ways humans view as empathetic, but is also designed such that humans can trust it is capable of empathy, will open endless positive possibilities. Humans ultimately hold the power to decide whether the future of AI is one we should fear or a future to look forward to. Scientific advances continue to challenge supposed limits on the ability for artificial empathy—now, we should embrace it. Indeed, as famous computer scientist and author Ray Kurzweil says: “Our technology, our machines, is part of our humanity. We created them to extend ourselves, and that is what is unique about human beings” (Adams, 2011). Empathy is a core facet of the human experience and should not be forgotten in the quest for advancing AI and humanity.

³ Biases can exist, however, if the input influences and data are inherently biased. Microsoft’s “Tay” Twitter-bot famously outputted racists and offensive Tweets, reflecting the biases of human data the bot was provided. For artificial empathy, the nature of data used would carefully need to be considered by its designers.
References