# Artificial Intelligence: Altruism, Psychopathy and Perception

David G. Hunt and Alexis J. Valentin

WhyFuture AI Concepts

19 May 2017

Author Note

David G. Hunt, Director, WhyFuture AI Concepts

Alexis J. Valentin, Secretary, WhyFuture AI Concepts

Correspondence regarding this paper should be directed to David G. Hunt, www.whyfuture.com. Contact: technologyconcept@whyfuture.com

Abstract

Designing a strong AI is akin to having an experienced and capable captain navigate a ship of passengers and, whether that ship is on course to the passengers' destinations or not will depend on the strength of the captain's training – how that strong AI is initially designed. The concern that a value-based ethics framework would result in psychopathy mimicry brings

forward a proposal to use altruism as an alternative to create a closer sense of true ethical perceptions.

## Artificial Intelligence: Recalling the Basics

There have been numerous advancements made regarding artificial intelligence over the passage of time. As a matter of fact, the artificial intelligence research field has been prolific in introducing new and innovative features that have yet to be recognized as AI advancement by the masses despite widespread use. The most familiar of such features include a number of existing online accomplishments such as the use of virtual agents, pattern recognition, and targeted advertising (Martin, 2015). While it is clear that AI already plays a major, if understated, role in modern society, ensuring that society is in a position to cope with all these advancements by obtaining a deeper knowledge regarding the processes involved and their importance is vital (Martin, 2015).

The primary objective of computerized reasoning attempts is to create a discerning machine that is fit for planning, thinking, arranging, taking care of issues, thinking dynamically, appreciating complex thoughts, taking in rapidly, and always learning. This amounts to the generally accepted description of human intelligence (Martin, 2015).

## **Concerns for a value-based ethical design**

As artificial intelligence design advances, as displayed by AI systems that are increasingly able to mimic human behavior and decision-making, the moral question becomes imperative. How does one include ethics – the ability to interpret correctly between right and wrong – into AI design?

In attempting an ethically-aligned design, the principle priority, logically and rightly so, begins with the consideration of universally accepted concepts of human benefit and “do no harm”. However, there is growing acceptance that a universal code of conduct does not exist beyond the “Western” understanding shaped by centuries of occidental philosophy, religion and moral codes. Even within smaller scopes of society or geographical regions, common sets of values become less and less common, with micro-society preferences, tradition and culture weighing in more importantly when determining values. This can be simply seen by an example of a village in Europe with views on individual rights that might vary from those of its province, that might differ from those of its country, that might contrast with those of the European Union. Therefore, there does not exist a universal value-based framework for embedding ethical design.

Even if localized values were to be implemented for localized AI, there is yet another concern that is missed out when choosing a value-based framework. A set of values that comes without empathetic connection, and without prior learned rationale, may result in an AI with actions that would merely imitate ethics – rather than actions as a result or intent of true human benefit. While this may still serve the purpose of a functioning AI, it cannot claim to be ethically-aligned, merely ethically compliant.

To illustrate, if a society originally adopts a norm against consuming meat on the basis of ethics, then it can be said to have adopted an ethical value. But if generations of families continue to accept and embed that value into their children to the point that even very young children are socially trained to do the same, then the abhorrence for meat is presented as purely psychological and the avoidance simply imitation. Without the underlying empathy to rationalize this preference, this non-meat value in young children cannot be considered ethically aligned.

## Altruism as a concept of ethical alignment in favor of value-based AI with perceptions of psychopathy

The AI should be developed in such a manner that it portrays an extensive and profound aptitude to understand its environments for the purpose of establishing what to do in the different situations that it is likely to come across. This further means that, for the AI to be in a position to comprehend its environment clearly and understand how to respond to these different possible situations, it needs to be socially intelligent as well.

It also needs to be creative since creativity comes in handy when encountering situations that require the management of problems.

For the purpose of realizing all the above-mentioned attributes, it is important to take certain factors into consideration. The first of these factors is the need to look into the traits of altruism vis-à-vis those of psychopathy. It is important to look into human altruistic behavior and make a thorough evaluation in order to be able to profile artificial intelligence around qualities that are considered humane, as well as philanthropic values.

This means that there is a need for thorough research to discover more about the deepest and most intricate foundations of human altruistic behavior. Other factors that ought to be taken into consideration are inclusive of what is generally needed to conclude that a person is altruistic as opposed to selfish. Therefore, in general, when designing an AI, it is imperative that it be shaped around the best and most positive traits of people (Hunt, 2016). This encourages attributes such as compassion, generosity, and the pursuit of equality, among others.

## Building empathy – alternative philosophical pathways

According to Stueber 2006 (as cited in Charisi et al., 2017), scholars of the philosophical persuasion believe that there are two types of empathy (p.7). The first is perceptual empathy, described by Misselhorn (2009) whereby the observer of a particular situation or action experiences emotions that can result in reactions that are equivalent to or congruent with those observed in the other (pp. 353-354). Several researchers (as cited in Charisi et al., 2017) presented a second type called imaginative empathy in which the observer is able to place himself or herself in the shoes of another, thereby necessitating different, alternative perspectives in the form of empathizing with the other (p.7).

**Perceptual empathy.** With the aid of specific theories of mind or neuronal resonance, the concept of perceptual empathy appears to be of plausible use in this endeavor. There is sufficient initial data to work with already, given early implementation of these in artificial systems, albeit in somewhat basic designs with a lot more room for development, such as the work of Balconi and Bortolotti (2011), who studied empathic responses to facial cues as a main social competency. The results suggest that the detection of facial emotion and empathic responsiveness could be related to empathic behavior.

More than a decade earlier, Mataric's (2000) work with behavior-based robotics attempted to replicate the evolutionary process that brought together visual input classification and structuring motor control systems in humanoid robots. Although its basis was imitation – based on the human ability to observe and repeat as a powerful form of learning – the experiment showed effectiveness on groups of mobile robots using basis sets that included avoidance, following, homing, and dispersion that allowed them to demonstrate higher-level group

behaviors such as collecting, foraging, and flocking. Mataric further showed that, with the same basis set, behavior selections could be improved over time with the use of a learning algorithm.

Ekman's (1992) work, which predates Mataric's by a decade, proved the utility of perceptive empathy in an approach that believed in the biological basis of basic emotions. Ekman successfully implemented a basal affect program as an autonomous reaction scheme, building a pathway to implement a very fundamental form of morality in robots.

**Imaginative empathy.** This form of empathy is thought to exist only as a product of human socialization and is not present even in ex-Homo sapiens (non-human) primates. This is because it is believed to be a lot more complex and can only develop from the foundation of perceptual empathy. Therefore, it can only exist with the precondition of the former.

Gallagher 2012 (as cited in Charisi et al., 2017) notes that because this form of empathy is intrinsically involved in higher-level moral reasoning and acting, it is much more complex and cognitively very ambitious (p.7). Gallagher maintains that the imagination needed to project the observer into the perspective of another could merely be a reiteration of the self rather than an expressed understanding of the other.

Because of this complexity, there does not appear to be any exercise of its implementation in artificial systems.

## The Bottom-Top Approach

In developing ethical AI systems, there are two major approaches that have been identified, each opposed to the other: the top-bottom approach and the bottom-top approach. The former involves breaking apart tasks into smaller sub-units until a means to perform a task directly is achieved. In ethical AI, this means that a specified ethical theory is instantiated into

identifying individual states and actions, classifying each as ethical or unethical (Wallach, Allen, & Smith, 2007, pp. 568-569).

The latter uses inductive logic programming over a body of ethical problems to unearth potential principles of ethical preference, weighing the benefits, harms, or autonomy (among other ethically relevant features) of relatable actions. The result is an extractable value or ethical rule that can contribute to learning to distinguish “right” from “wrong.”

As the top-down approach requires a preconceived set of ethical principles that do not change, this does not support learning and adapting, which is what is desired in a truly ethical AI design. Charisi et al. (2017) also believe that an AI system that learns its own ethical rules might be better at adapting to situations that are not predetermined.

Therefore, the bottom-top approach is believed to be the best means of achieving AI through altruism and is the one that best represents ethical design.

**Advantages:** The bottom-top topology is greatly enhanced by the fact that the AI system receives data that is already known or that can be predicted through its interaction with the environment (Charisi et al., 2017). This enables the data to be processed in a manner that is useful.

It also operates without a predetermined moral or ethical principle, which means that it can come up with its own parameters and implement competencies autonomously. Under a model of human socialization, AI systems bypass the need to choose one denominating ethical theory to implement, judging consequences rather than motivation. AI systems thus learn morality through empathy, fitting the role described in this paper.

**Challenges.** It can be difficult to verify whether the AI system indeed fulfills any imposed requirements, but this is a challenge not unique to this method and common across all machine learning. M. Anderson and Anderson (2014) propose an ethical Turing test to overcome this challenge involving responses recorded by a human ethicist against those of the system. An AI system is deemed to have passed the test if its responses are sufficiently similar to those recorded by the ethicist.

## Empathy Mirror

Opting for the recommended bottom-top topology, perusing perceptual empathy can be greatly enhanced by the use of mirror neurons, as shown by the research examples above (Charisi et al., 2017).

Rizzolatti and Fabbri-Desto (2008) underline the crucial role of this empathy mirror system in social cognition, proving that goal-coding motor neurons are activated by observing motor acts done by others. They show strong evidence that the mirror mechanism enabled the observer to understand the intention behind an observed act, as well as its goal. This is the equation of altruism, leading to true – as opposed to mimicked – empathy.

Via mirror neurons, the human neural network's reactions to situations leading to empathy triggering can be analyzed. These analyses can then be used to evolve parameters around selected areas into an implementation module for AI systems.

## Efficiency kill-switch

The second factor that should be taken into consideration is the ethical dilemma known as the ethical paradox. This refers to a situation in which there is a need for the AI to choose which action to take: Being diligently efficient or staunchly keeping to its moral obligation. This brings

up the issue of psychopathy vis-à-vis empathy. Inasmuch as artificial intelligence ought to be shaped in a manner that makes it efficient, this should, at no time, defeat the ability for it to be empathetic when the need arises.

An AI ought to be designed in a manner that allows it to instantly opt out of being efficient in order to show compassion toward someone or people according to the situation at hand (Hunt, 2016). The recommendation of a "lessons learned" database would also help the AI learn this, as it would present historical evidence of what was deemed to be "correct" ethical decisions made by humans.

Consider the following from Foot (1967, p. 3): An airplane pilot has lost nearly all control of the plane. This pilot is presented with a dilemma. The pilot can either steer the plane and crash into a less populated area or do nothing and allow the plane to crash into a more populated area. According to Hunt (2016), one can see that if the pilot chooses to steer the plane to a less populated area, he or she is more so acting on empathy rather than efficiency. However, in the fat man trolley problem (Thomson, 1985, p. 1409), as explained by Hunt (2016), pushing the fat man over the bridge to save more lives is choosing efficiency over empathy, and most people would reject the notion of pushing the fat man over the bridge as they prefer to place empathy above efficiency. Furthermore, it points to the fact that individuals choosing efficiency over empathy in such a situation correspond with more of a psychopathic mind (Singer, 2005, p. 341, as cited in Greene 2002, p. 178).

There is a different example from Hunt (2016) of an efficiency-over-empathy situation: A doctor is in urgent need of vital organs to save five patients. A person happens to arrive at the clinic with the exact needs of these five people. Should the doctor sacrifice this individual against his

or her will and save the five patients in need? Most people would say, “No, it would not be morally permissible for this doctor to proceed” (Thomson, 1985, p. 1396). However, what if the doctor did decide to do this, and this was considered standard practice at the clinic? The clinic would be a place that almost everyone would avoid, and people would not trust the clinic. Thus, there is a need for an AI to place empathy over efficiency.

Therefore, it is important for persons who design AI to be able to structure it in accordance with their defined moral systems as well as the manner in which they are supposed to position themselves depending on different situations that they may face in the future where they will have to make moral decisions (Martin, 2015).

Perception is another important factor to consider when designing AI, since it is through perception that people have the ability to critically evaluate the situations that are presented before them. Therefore, it is important to factor in a dimension of context with a dimension of actions. The lack of an in-depth analysis of context can lead to a conclusion that seems to defy common sense in certain situations. As an illustration of this, suppose an AI is given a task to judge and weigh the positives versus the negatives of a person. One may conclude that the AI is justified when it tallies the negative attributes of a person, such as thievery, and the positive attributes of a person, such as occasional charity. However, consider the following: A person's house burns down. This person becomes frustrated, emotional, and utters foul language in an expression of emotional distraught. This person punches a tree and sobs in the corner over losing everything he or she owns. Another person can understand why he or she is acting like this via the dimension of context, as a tragedy had just befallen to this person. However, the AI’s flaw in this situation would be apparent. It would tally the person’s actions, such as punching the tree

and uttering foul language, and label him or her an undesirable person against evidence that is devoid of context when, in reality, he or she might be the opposite. This goes to show that it is crucial to design an AI while factoring in an understanding of both the context and actions of a given situation, as this can lead to an AI with a more compatible perception, which is better for humanity (Hunt, 2016).

## Conclusion

In summary, the concern is that a value-based framework will result in AI that is circumstantially ethically compliant instead of AI that is deliberately ethically aligned.

As such, it is recommended that an AI should be designed with the ability to look at why and how particular systems, beliefs, codes, and values are the way they are for humans and make a decision based on how each particular of these relates to priors. Upon doing this, it can implement its decision based on all of these facts (Hunt, 2016).

## References

Anderson, M., & Anderson, S. L. (2014). GenEth: A General Ethical Dilemma Analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*(pp. 253-261). Quebec City.

Balconi, M., & Bortolotti, A. (2011). Detection of the facial expression of emotion and self-report measures in empathic situations are influenced by sensorimotor circuit inhibition by low-frequency rTMS [Abstract]. *Brain Stimulation,5*(3), 330-336. doi:10.1016/j.brs.2011.05.004

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., . . . Yampolskiy, R. (n.d.). Towards Moral Autonomous Systems. 1-22. Retrieved June 25, 2017, from https://arxiv.org/pdf/1703.04741v2.pdf.

Ekman, P. (1992). Are there basic emotions? *Psychological Review,99*(3), 550-553. doi:10.1037//0033-295x.99.3.550

Foot, P. (1967). *The Problem of Abortion and the Doctrine of the Double Effect*. Oxford Review, (5), 1-7. Retrieved April 21, 2017, from http://pitt.edu/~mthompso/readings/foot.pdf

Greene, J. D. (2002). *The Terrible, Horrible, No Good, Very Bad Truth About Morality, and What to Do About It*. Ph.D. Dissertation. Department of Philosophy, Princeton University. Retrieved from http://emilkirkegaard.dk/en/wp-content/uploads/Joshua-D.-Greene-The-Terrible-Horrible-No-Good-Very-Bad-Truth-about-Morality-and.pdf

Hunt, D. G. (n.d.). *The Blueprints Towards the Development of Good Artificial Intelligence.* Retrieved February 16, 2016, from http://www.whyfuture.com/single-post/2016/11/06/The-Blueprints-towards-the-Development-of-Good-Artificial-Intelligence

Martin. (2015). *Artificial Intelligence: A Complete Guide*. Retrieved February 16, 2017, from https://www.cleverism.com/artificial-intelligence-complete-guide/

Mataric, M. (2000). Getting humanoids to move and imitate. *IEEE Intelligent Systems,15*(4), 18-24. doi:10.1109/5254.867908

Misselhorn, C. (2009). Empathy with Inanimate Objects and the Uncanny Valley. *Minds and Machines,19*(3), 345-359. doi:10.1007/s11023-009-9158-2

Pan, L. (2016). *Why China Isn't Hosting Syrian Refugees*. Foreign Policy. Retrieved from http://foreignpolicy.com/2016/02/26/china-host-syrian-islam-refugee-crisis-migrant/

Rizzolatti, G., & Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Current Opinion in Neurobiology,18*(2), 179-184. doi:10.1016/j.conb.2008.08.001

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics,9*(3-4), 331-352. doi:10.1007/s10892-005-3508-y

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal Company Inc,94*(6), 1395-1415. doi:10.2307/796133

Wallach, W., Allen, C., & Smit, I. (2007). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society,22*(4), 565-582. doi:10.1007/s00146-007-0099-0