Kelly Quinn

BIOETH 505

May 2021

**Artificial Moral Agents and Social Dilemma: Limitations of Machine Morality**

*Introduction*

In the popular and brilliant science fiction work of Cixin Liu, *The Dark Forest*, the protagonist must contend with the stark realization that there is limited space in the universe, that extraterrestrial life is abundant, and therefore resource conflict is inevitable. Coupled with a communication breakdown, conceptualized as ‘a chain of suspicion,’ due to differences and distance that render cooperation impossible, the characters find themselves in a social dilemma on a cosmic scale. Scaled down in much less fantastical fashion, humans cope with social dilemmas of varied proportions daily. They apply to a broad range of problems between as little as two agents, small groups, or entire nations (Van Lange et al., 2013). Many of the contentious social dilemmas that humanity deals with involve conflict between self-interest and the good of the collective. This is seen in commercial overfishing, deforestation, pollution, and illegal activity such as drug and human trafficking. As mankind contends with solving moral dilemmas, they are increasingly relying on artificial intelligence as decision support tools. The concept of an artificial moral agent (AMA) is proposed by Wallach & Allen (2009) and requires a study of ethical theory to hypothesize on the limitations of AI to handle moral dilemma. We will explore this idea in-depth in this paper, consider classical moral dilemmas, and modern advancements on classic matrix game social dilemmas that provide insight to the nature of social cooperation.

*Moral Machines*

We already rely on artificial intelligence (AI) to work with some autonomy. As this reliance continues to develop the question of what we wish to program into our artificial intelligence remains up for discussion. However, it will be imperative that ethical considerations are on the forefront of our engineering initiatives. Semi-autonomous machines are already working among us that require a function of moral decision making (Wallach & Allen, 2009). With the advent and implementation of self-driving cars, fundamental philosophical morality cases, such as trolley cases, may be even more difficult to answer because inaction or ‘gut reaction’ are not possible. The goal of an artificial moral agent in such a scenario would be to act in a way that minimizes potential to do harm and to work against neglect of duty. We expect the same of our human moral agents, but it is reasonable that we will tolerate mistakes or poor judgment less in regard to artificial intelligence agents. I argue this for a simple reason: that they have been designed by our hand.

Wallach & Allen (2009) simplify the concept of building a moral machine down to the parameters of 1.) programming the right set of constraints and 2.) the right algorithm for conflict resolution. They use the term ‘bounded morality,’ wherein the artificially intelligent system would function in an acceptable manner as long as it encountered situations that fell within the constraints set down by the engineer (Wallach & Allen, 2009). Predicting moral dilemma is no small feat. As mentioned above, the lack of human intuition in machines necessitates the programming of instruction to follow. Where there may be a moral distinction between action and inaction, the engineer in charge of designing a moral machine would not be able to just make the choice for inaction over good action. This leaves a designer with two options. One, the programmer makes a set of predictions about what kind of scenarios the AMA may function in and program a preset of actions or the programmer could design a more open-ended system in which the AMA has the ability to collect information, compare it to similar data, and respond with a best fit (Wallach & Allen, 2009). An interesting consideration for the latter scenario is the type of solution an AMA may come up with. With is superior ability to process huge amounts of data Text

Description automatically generatedDiagram

Description automatically generatedquickly the result may surprise us.

In their book, Wallach & Allen (2009) offer a framework for the development of AMAs. A simple cartesian plane with an axis for autonomy and another for moral sensitivity help illustrate developments and future iterations.

Low scores on both dimensions are categorized as ‘operational morality’ where any moral significance is preprogrammed by the engineer. A high score in either of the two axes partnered with a low score in the other is categorized as ‘functional morality’ where the AI may respond to moral dilemma with less programmed constraint. The final category, ‘full moral agency,’ where an AI may operate with both high autonomy and high ethical sensitivity remains in the realm of science fiction and hypothesis.

Source: Wallach & Allen, 2009

To attain this level of moral agency many scientists believe that an AMA would need to have a semblance of consciousness and yet, more scientists believe that a machine will never be capable of attaining either consciousness or a deep understanding of human emotion and relationship dynamics (Wallach & Allen, 2009). Despite these identified limitations, humans are increasingly relying on machines to assist in moral decision making. Friedman and Kahn (1992) have expressed concern that an overreliance on decision support tools may supersede critical thinking, where human agents relinquish control of moral decision making to a machine.

*Decision-Making in Moral Machines*

Luciano Floridi and J. W. Sanders (2004) identified a few features that would comprise an AMA in decision making capacity. These are interactivity, autonomy, and adaptability. Interactivity is the ability to respond to stimulus, autonomy is the ability to act when stimulus is absent or indirect, and adaptability is the ability for the machine to learn based on experience with changing stimuli. Similar to other frameworks for agency, the degree to which the AMA has freedom within these domains reflects the complexity of the AI. As cognitive complexity increases so too does an awareness of conflict between agents or situations . These conflicts or differences in either perspective or orientation may produce different outcomes.

At a fundamental level, programming of AI should have conditional logic to handle interactivity and respond to environmental cues. AI is a problem-solving agent presented with a set of parameters and constraints and programmed to produce output. This is not unlike the problem-solving nature of the very engineers that design AI (Wallach & Allen, 2009). Current AI is not capable of navigating the varied situations that the average human contends with daily. Expecting it to handle this complexity is not a practical approach. However, steps continue to be made in the direction toward moral sensitivity as technology advances (Wallach & Allen, 2009).

*Social Dilemmas*

This paper will now pivot to discuss social dilemmas and how the aforementioned restrictions on machine morality may impact their ability to handle dilemma. Furthermore, by looking at machine morality through the lens of social dilemma we can imagine what insights AI can provide to us in approaching classic problems. A social dilemma is a conflict, typically between two or more agents, where there is a divergence between self-interest and collective interest (Van Lange et al., 2013). There are two important concepts related to social dilemmas that will be considered: cooperation and defection. The inputs that go into decisions regarding cooperation and defection are not always simple. Cooperation may mean contributing to a collective interest or it may mean not taking from a collective interest. Despite the knowledge that cooperation will benefit the collective (inclusive of the individual agent), individual actors still struggle with choosing to operate in self-interest (Van Lange et al., 2013). Dawes in his 1980 article in the *Annual Review of Psychology*, proposed two considerations. That a.) each decision maker has a dominating strategy toward non-cooperation, or the option that produces for them the highest possible outcome, and b.) that if every agent chooses the dominating strategy, the collective will be worse off than if they had all cooperated (Dawes, 1980).

What may be missing from Dawes’ initial synopsis is the time dimension. There is the concept of temporality, particularly of consequences (Van Lange et al., 2013). There are often short-term outcomes to consider as well as long-term outcomes. Should an agent work for their short-term benefit if the outcome for the collective is a long-term detriment? Should an agent sacrifice a short-term benefit or even possibly suffer a short-term detriment with the understanding that the benefit for the collective will be long-term? It is important to consider the limitations that temporality imposes on conceptualizing coping with social dilemmas (Van Lange & Joireman, 2008). While maintaining the temporal dimension in this evolving concept of social dilemma we will expand the simplistic definition of a social dilemma provided above. Caroline Whitbeck, a professor at Case Western Reserve University, expands the concept of social dilemmas to include not only two opposing principles, but as “problems in which there are multiple ethical constraints which may or may not turn out to be satisfiable simultaneously” (Wallach & Allen, 2009). To study these multiple constraints, scientists and engineers have turned to interdependence theory to test outcome measures.

*Matrix Game Social Dilemmas*

Finding its foundation in game theory, or choosing from a rational set of options in a multi-agent scenario (Mishra, 2021), interdependence theory assumes that interactions are the combined sum of SABI – Structure, Partners A and B, and Interaction Dynamics (Van Lange et al., 2013). The structure is the environment and constraints of the game, partners A and B are the opposing agents or principles, and interaction dynamics are the rules of engagement. The outcome possibilities are often referred to as a matrix, in which the agents transform outcomes through behavior where one’s choices impact the choices of the other agent bidirectionally (Mishra, 2021). Given matrices are simplistic where the self-interested preferences of action are driven by one agent’s needs or skills versus an effective matrix where a decision-maker must consider macro, temporal, and collective concerns (Van Lange et al., 2013). This distinction opens the door for theorizing on the effect of altruism and reciprocity on bias toward cooperation (Van Lange et al., 2007).

The theory of general sum matrix games are employed to study the application of interdependence theory and help study emergent cooperation, its promotion, and potential to stabilize (Van Lange et al., 2013). We recall the concept of cooperation and defection to discuss a classic matrix game social dilemma (MGSD), the Prisoner’s Dilemma. The Prisoner’s Dilemma structurally has two prisoners being interrogated in separate rooms. Neither knows what the other prisoner will confess to and thus must make their choice based on assumptions. Each prisoner is offered a reduced sentence if they provide testimony against their partner (defect). It is set up in a way that the most logical choice is for each partner to testify against their partner, despite the fact that if both partners remain silent (cooperate) there is the highest potential for collective payoff (Wallach & Allen, 2009). A matrix game may be identified as a social dilemma by satisfying one of four inequalities with the following components (Leibo et al., 2017):

|  |  |
| --- | --- |
| **Components** | **Inequalities** |
| **R** (Reward for mutual cooperation)  **P** (Punishment for mutual defection)  **S** (Sucker outcome where one partner cooperates and another defects)  **T** (Temptation outcome to defect on a cooperating partner) | **R > P**  *Mutual cooperation is preferred to mutual defection* |
| **R > S**  *Mutual cooperation is preferred to being exploited by a defector* |
| **2R > T + S**  *This ensures that mutual cooperation is preferred to an equal probability of unilateral cooperation and defection* |
| **Greed**: **T > R**  *Exploiting a cooperator is preferred over mutual cooperation*  **Fear** : **P > S**  *Mutual defection is preferred over being exploited* |

MGSDs offer insight into the role of altruism in social cooperation. Reciprocal altruism (Trivers, 1971) also referred to as *direct reciprocity*, is the idea that people respond quite strongly to the behavior of others. In fact, a dated but still interesting article found that people behave with reciprocal cooperation to the degree of which they expect the other agent to behave cooperatively (Van Lange, 1999). The second concept of *indirect reciprocity* refers to reputational cooperation, where an agent may base behavior decisions on the reputation of the other agent as cooperative. This may be learned from a history of repeated interaction and may be dependent on numerous factors such as one’s identity, personal history, and heuristics such as *The Golden Rule* (Weber et al., 2004). Social desirability also biases agents toward wanting to have a reputation as cooperative. Interestingly, this concept may explain why indirect reciprocity holds true between strangers, so long as there is a known and expected reputation of cooperation (Van Lange et al., 2013).

MGSDs, similar to Dawes’(1980) concept of classic social dilemma, have likewise been guilty of not taking into account real-world considerations (Leibo et al., 2017). This includes, without being limited to, ignoring temporality where consequences of action for both the self and the collective may be either immediate or delayed (Van Lange et al., 2013); decisions to cooperate and defect may happen near simultaneously where agents respond to a change in behavior of their opposition; and decisions often must be made when an agent has only incomplete information about the state of the environment, their opposition’s behavior, or their motives (Leibo et al., 2017). Google DeepMind ran a series of experiments in MGSDs to contend with these very concerns (Leibo et al., 2017).

*DeepMind and Sequential Social Dilemmas*

As technological advances in AI increase the capacity of the machines themselves, it also clears the way for increased capacity for handling complex problems with entirely new perspectives. Classic social dilemma problems are often discussed as a basic, binary set of choices where an agent chooses to either cooperate or defect (Leibo et al., 2017). In real-world scenarios, as we have discussed, this simplistic modelling is impractical. Choices to cooperate or defect are often complex, dynamic, interrelated, interdependent, and subjective. To contend with the simplicity of MGSDs, DeepMind introduced the concept of *Sequential Social Dilemmas* (SSDs), where the mixed-incentive structure holds true, but repeated play and complex learning strategies for the agents themselves are incorporated to better replicate real-world outcomes (Leibo et al., 2017). By restructuring the MGSDs as SSDs, DeepMind was able to study the factors that influence agent cooperation in two experiments called *Gathering* and *Wolfpack*.

There were two domains aside from the rules of engagement that changed in the games. Consequent behavior was studied as environmental changes were introduced as well as manipulation of the agent’s ability to learn and employ complex strategies (Leibo et al., 2017). DeepMind observed that while both scenarios satisfied a Prisoner’s Dilemma, the ability to learn complex strategies produced opposite results for Gathering and Wolfpack in regard to cooperation (Leibo et al., 2017). This will be discussed in more detail in the following paragraphs.

*Gathering and Wolfpack*

The goal of Gathering is for each agent to collect apples. Upon collection, the agent receives a point of 1 and the apple is temporarily removed from gameplay. Agents have the ability to shoot a beam directly across to ‘tag’ the opposition, temporarily disabling them. There is no reward for tagging, it simply paralyzes the opposition facilitating the agent in collecting more apples and preventing the opposition from doing so. The choice to tag is the defect action in this scenario versus the choice to refrain from tagging as the cooperative action. Manipulations in the environment included changing the respawn rate for apples (scarcity) and respawn rate for paralysis (conflict-cost) on aggressiveness (beam use rate). In sequential, repeated game play as the AI learned strategy, DeepMind found that when learning lead to change in aggression, it was almost always to increase. Unsurprisingly, in scenarios where there was low-abundance or high conflict-cost aggression was quite high and in scenarios where abundance was high or conflict-cost low, aggression was comparatively lower. DeepMind concluded that Gathering satisfied a Prisoner Dilemma, where for inequality 4, greed was motivated by removing a rival and collecting apples and fear was motivated by the chance that an agent may be tagged by a defecting agent (Leibo et al., 2017).

The Wolfpack gameplay was comprised of two agents (hunters) chasing a third agent (prey). When the prey is caught, rewards are directly proportional to the number of wolves in the capture radius. For instance, a lone wolf capture may garner one point, where a team capture (both wolves within the capture radius) may garner two points. The learning strategy for cooperation varied and was learned over sequential gameplay. Depending on the environmental manipulation – such as a change in group capture bonus – different strategies emerged. One strategy resulted in the wolves first finding one another and then sought the prey in tandem and the second where a lone wolf hunted the prey and then waited for a team member to enter the capture radius before attacking. In these scenarios, a lone wolf capture would be a defection and the learned strategies cooperation (Leibo et al., 2017).

In both experiments, environmental manipulation altered the behavior of the agents. Unexpected to the researchers at DeepMind, however, a difference between the games did emerge. While both experiments satisfied Prisoner’s Dilemma (though Wolfpack occasionally satisfied other MGSDs such as Chicken and Stag Hunt, *see Figure 6* below), the team at DeepMind observed that in situations where agents had a higher capacity to implement strategies the outcomes were opposite between games. In Gathering, the more complex the learned strategy the more aggressive the AI became regardless of apple scarcity. Conversely, the more complex the learned strategy in Wolfpack, the more agents chose to cooperate rather than defect. This leaves us with some insight into intelligent behavior, but also the puzzling new question of why, as a situation changes coupled with an agents’ higher capacity for complex strategy implementation, does cooperation vary?Chart, scatter chart

Description automatically generated

Source: Leibo et al., 2017

*Implications & Conclusion*

This first begs the question, *why would humans wish the create artificially moral machines*? Followed closely by *what may be the consequences*? Perhaps what has most directly separated homo sapiens from its close relatives is the use of tools (Wallach & Allen, 2009). As mankind becomes increasingly dependent on its tools, we may be at a crossroads of autonomy and control. Focus on the future of technology and boundless advancement seem to exclude the inevitable changes on mankind as it creates deepened intimacy and entanglement with its machines (Turkle, 2002; Wallach & Allen, 2009). Where technology currently stands ethical and moral complexity in machines extends as far as the values that have been programmed into them by engineers and designers given contemporary capability and limitation. In a field that tends to oscillate between unbounded optimism and unfounded ‘doomsdaying,’ there is perhaps a middle ground where the increased capacity of intelligent machines offer new perspectives to old problems and offer better solutions to dilemmas than their human counterparts.

So, the final question is posed: *Should humans want moral artificial intelligence making important decisions?* Philosophers continue to warn against overreliance on technology. We are faced with the conundrum that emergent technologies may be more easily modified than those that are entrenched, but a deep understanding of impact cannot be evaluated until there is wide-spread adoption (Wallach & Allen, 2009). Perhaps, as with all unknowns, the best bet is to proceed with caution.

**Resources**

Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology*, *31*(1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Friedman, B., & Kahn, P. (1992). Human Agency and Responsible Computing: Implications for Computer System Design. Journal of Systems and Software 17, 7–14.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). *Multi-agent Reinforcement Learning in Sequential Social Dilemmas*.

Mishra, P. (2021). *Game Theory in Artificial Intelligence—Programming Class [2021]*. Study.Com. https://study.com/academy/lesson/game-theory-in-artificial-intelligence.html

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Turkle, S. (2002). *Whither Psychoanalysis in a Computer Culture? « Kurzweil*. https://www.kurzweilai.net/whither-psychoanalysis-in-a-computer-culture

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. https://doi.org/10.1037/0022-3514.77.2.337

Van Lange, P. A. M., De Cremer, D., Van Dijk, E., & Van Vugt, M. (2007). Self-interest and beyond: Basic principles of social interaction. In *Social psychology: Handbook of basic principles* (2nd ed., pp. 540–561). The Guilford Press.

Van Lange, P. A. M., & Joireman, J. A. (2008). How we can promote behavior that serves all of us in the future. Social Issues and Policy Review, 2, 127– 157. doi: 10.1111/j.1751-2409.2008.00013.x. *Journal of Personality and Social Psychology*, 10–1037.

Van Lange, P. A. M., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125–141. https://doi.org/10.1016/j.obhdp.2012.11.003

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. In *Moral Machines*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780195374049.001.0001/acprof-9780195374049-chapter-1

Weber, J. M., Kopelman, S., & Messick, D. M. (2004). A Conceptual Review of Decision Making in Social Dilemmas: Applying a Logic of Appropriateness. *Personality and Social Psychology Review*, *8*(3), 281–307. https://doi.org/10.1207/s15327957pspr0803\_4