Virtue Under Pressure: The Case for an Exemplarist Virtue Ethics Framework to Build Artificial Moral Agents for High-Pressure Tasks

Harry Collins
St. Edmunds College, University of Cambridge

This paper examines artificial moral agents (AMA) and seeks to justify their use to perform tasks involving high situational pressures that significantly impact human moral decision-making even when there is consensus on the correct decision. Moreover, these tasks can lead to moral injury for human decision-makers if their moral code has been violated, either by themselves or by external situational pressures. I argue that AMAs can potentially negate these concerns, particularly AMAs utilising exemplarist virtue ethics, a flexible approach to normative ethics, allowing agents to learn from experience to emulate the virtue of selected exemplars. To this end, I propose the outlines of an exemplarist framework for building AMAs using reinforcement learning from human feedback, where feedback is provided by moral exemplars in given tasks. Processes for selecting candidate tasks, identifying exemplars, and developing AMAs to emulate those exemplars are provided. Finally, potential objections are considered, both against the idea of exemplarist AMAs and their feasibility. I conclude that exemplarist AMAs for high-pressure tasks are promising candidates to perform high-pressure moral tasks, reducing moral injury for humans, although issues such as minimising cross-cultural disagreement on moral decisions and how well agents capture morally relevant features in their environment to emulate exemplars need further exploration through practical experimentation.

Introduction

As artificial intelligence systems become more advanced and require less human oversight, there are questions about whether such systems should be used for moral decision-making and over what kind of artificial moral agents (AMAs) should be developed (Wallach & Allen, 2008). Therefore, this paper aims to determine what kind of AMAs, if any, humanity should be striving to build. I will argue that AMAs are justified in tasks with high situational pressure that negatively impact human decision-making and can lead to moral injury (MI). Moreover, I will argue for an exemplarist virtue ethics approach to training AMAs to reproduce domain-specific moral exemplars' virtue.

Firstly, I will introduce AMAs and the kind I will consider. Next, I will overview virtue ethics, particularly the exemplarist approach. Then, I will address whether any AMAs should be used for moral decision-making, justifying them in high-pressure domains by showing that high situational pressure diminishes human virtuous behaviour, potentially also leading to MI. Next, I will suggest what kind of AMAs should be developed by presenting theoretical outlines for a virtue-based framework for training AMAs to emulate moral exemplars in high-pressure domains, showing a theoretical end-to-end implementation and addressing potential objections.

1. Defining AMAs

First, I will determine which AMA definition will be used. The term "artificial moral agent" was introduced by Wallach and Allen (2008, p. 4), stating that AMAs are robots that act independently from real-time human supervision and make moral decisions, meaning determining the right action based on ethical values.

However, determining to what degree a bot (software agent or physical robot) is a moral agent requires more specific definitions. For instance, a bot that alerts when you go over a speed limit and a vehicle-driving bot that must decide whose lives to prioritise in unavoidable crashes both have ethical impacts, yet the latter requires complex moral decision-making capabilities. Moor (2006, pp. 19-20) suggests four levels of moral agent:

1. Ethical impact agent: any machine that can have ethical consequences.
2. Implicit ethical agent: machines that reflect moral values without explicit representation.
3. Explicit ethical agent: machines that recognise and take morally relevant

information into account and can make explicitly moral decisions.
4. Full moral agents: machines with human-level moral agency.

Levels 1 and 2 lack the capacity for moral considerations. Level 3, however, is more advanced. Misselhorn (2022, p. 34) compares these AMAs to chess bots that recognise all chess-relevant information, evaluate it, and determine the best move. Explicit ethical agents recognise "general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations" (Moor, 2009, p. 12), meaning they can adapt to situations not explicitly accounted for, like how chess bots successfully evaluate novel positions. These AMAs can theoretically emulate moral decision-making without human moral decision-making features, like "consciousness, intentionality, and free will" (Misselhorn, 2022, p. 35) required for full moral agents. As there is much debate over defining consciousness and whether machines can achieve it (Searle, 1980; Himma, 2009), I will argue for explicit ethical agents which appear more practically achievable. Moreover, this avoids existential objections to AMAs as these agents do not require general superintelligence (Chalmers, 2010). They also lack emotions, so moral considerations towards them need not be considered. Next, I will overview virtue ethics as I will later argue for virtue-based AMAs.

2. Virtue Ethics

2.1. Overview

This section will explain virtue ethics and exemplarist approaches. Virtue ethics is a normative ethical theory focused on cultivating strong moral dispositions (virtues) like honesty and helpfulness to determine the right kind of person to be (Hursthouse & Pettigrove, 2023). Virtues must be learnt through repeated practice, so one honest act does not make someone honest, nor does a single lie make someone dishonest. Sometimes lying may be appropriate, like in the "Murderer at the door" thought experiment (Varden, 2010). Someone runs into your house to hide. A murderer then appears, asking if the would-be victim is inside. Although honesty is generally virtuous, lying would save the victim's life. To determine this, virtuous agents require phronesis, "the "practical wisdom" [...] learned by acting in social situations and gives those agents that possess this quality the ability to make new or novel judgments" (Sullins 2021, p. 136). Learning from experience is critical to habituating virtuous behaviour. I will now overview exemplarist virtue ethics as a learning approach, as I will later argue for its applicability to AMAs.

2.2.Exemplarist Virtue Ethics

Zagzebski's (2013) exemplarist approach suggests that, to become virtuous, moral exemplars (agents with admirable moral qualities, also called phronimos) must be identified and learnt from by observing their actions. Right and wrong moral actions are determined as follows: "a wrong act = an act that the phronimos characteristically would not do, and he would feel guilty if he did = an act such that it is not the case that he might do it = an act that expresses a vice = an act that is against a requirement of virtue (the virtuous self)". Rather than defining a set of virtues, exemplars embody virtue. Exemplars can be identified through admiration, defined as "attraction that carries the impetus to imitate" (Zagzebski, 2013, p. 201). After identifying potential exemplars, people should critically reflect on whether they are suitable to learn from. Further details will be given when applying this to AMA development. However, assuming moral exemplars can be identified, why not teach other humans to emulate them rather than AMAs? I will now seek to justify AMAs based on situationist critiques of virtue and MI.

3. If Any: How situationism and Moral Injury Justify AMAs in High-Pressure Domains

3.1. Situationism

Several debates surround whether AMAs should exist, with Formosa and Ryan (2021, p. 9) suggesting that debates should be more specific towards morally appropriate or inappropriate use cases. Hence, I will focus on a specific use case, showing how situational pressures and MI impact human moral decision-making, thus justifying AMAs in high-pressure domains/tasks. Firstly, I will discuss situationism, the argument that "variance in human behaviour is typically a function of the situation [...] rather than any traits of character" (Upton, 2009, p. 104), claims decisions are based more on situational pressures than universal virtues. Situational pressures are contextual factors that can influence behaviour, like the Milgram experiment (Milgram, 1963), where participants were instructed to electrically shock someone in another room if they got a word-pair recall question wrong, with shock intensity increasing each time. Participants were frequently urged to continue by an authority figure. The other person screamed in pain but was not actually shocked, unbeknownst to the participants. Over two-thirds of participants continued shocking after the other person feigned unconsciousness. Situationists argue that the authority figure's pressure explains why most participants inflicted deadly shock levels, as the large random sample negates the possibility that all participants were cruel. This study has been criticised for being unrealistic (Orne & Holland, 1968). However, participants believed their situation was real, giving genuine responses. For virtue ethicists, one cruel behaviour does not make someone cruel, although sufficient situational pressure clearly diminishes most participants' virtuous behaviour. Virtue ethicists may also argue that virtues cannot be simplified into behaviours (Kupperman, 2001). Whilst virtue is more about character than individual actions, virtuous agents should use phronesis to make good decisions, particularly when there is a clear consensus on the morally correct decision, as in situationist experiments.

Another example showing how time pressures can impact helping behaviours is the Good Samaritan experiment (Darley & Batson, 1973). Here, theology students travelled to a building to discuss the Good Samaritan parable under differing amounts of time pressure. One-third had to rush (high-pressure), another third were due to be just on time (medium-pressure), and another third had excess time (low-pressure). While travelling, participants encountered someone who needed help. Over 63% of low-pressure students helped versus sub-10% of high-pressure students, showing that as time pressure increased, helping decreased. Many situationist experiments similarly demonstrate virtuous behaviour decreasing as situational pressure increases (Alzola, 2008). Hence, some situationists argue that virtues have minimal behavioural impact versus situational pressure (Doris, 1998). Although some virtue ethicists dismiss this as oversimplifying virtue as behaviour in a single situation (Kupperman, 2001), high pressure clearly negatively impacts behaviour approximated as virtuous. However, this does not refute exemplarist virtue ethics, as a minority of participants exhibit exemplary moral behaviour despite situational pressures, like the 10% of high-pressure students who helped. Therefore, AMAs that learn from these exemplars would morally outperform most humans, justifying them. Furthermore, de Bruin *et al.* (2023) provide empirical evidence suggesting virtuous behaviours are more stable than situationists claim, demonstrating that the "ability to withstand the pressure and act virtuously is particularly present in mid-range situations" (470), so virtuous behaviour only diminishes under very high pressure for most humans. Hence, AMAs are justified where high situational pressure is expected because AMAs emulating exemplars displaying virtue despite high pressures would prioritise helping distressed individuals over timeliness or refuse to electrically shock someone when pressured to. Next, I will show how MI links to situationism, reinforcing this justification.

3.2. Moral Injury (MI)

Moral injury is "the strong cognitive and emotional response that can occur following events that violate a person's moral or ethical code" (Williamson *et al.*, 2021, p. 453), with these events either perpetrated or observed by that person. For instance, healthcare staff may feel unable to offer appropriate care if given inadequate supplies or when managing overly high workloads. MI can involve long-term feelings of shame, altered beliefs and self-destructive coping mechanisms. Coimbra *et al.* (2024) also associate it with an increased risk of suicidal ideation, burnout, depression, anxiety and Post-Traumatic Stress Disorder. Therefore, if someone suffers from MI, they may struggle to perform to previous standards, resulting in negative outcomes for them and anyone impacted by such lowered standards. For instance, 8-out-of-10 UK National Health Service (NHS) doctors suffered from MI during the COVID-19 pandemic due to increased workloads and pressure (Rimmer 2021, p. 1). The NHS also suffered record departures post-pandemic, citing stress and work-life balance issues (Savage, 2022; Best, 2021, p. 2). The remaining healthcare workers then face even more pressure due to understaffing. As shown by situationists, higher pressures diminish most people's moral decision-making, meaning higher chances of MI if a poor decision violates someone's morals. This creates a vicious cycle of increasing pressure, leading to a greater risk of MI and vice-versa. Therefore, AMAs are further justified in tasks where high situational pressure can be expected, as they can mitigate both the drop-off in humans' virtuous behaviour as situational pressure increases and MI risk. Other factors could also reduce situational pressures and MI risk, such as hiring more staff to reduce workloads in healthcare. However, high situational pressure can still arise, as shown by the pandemic. As high situational pressure is key to MI, both moral decision-makers and those affected by moral decisions could benefit from AMAs to further reduce situational pressures, meaning humans will be less likely to violate their moral codes.

4. What Kind: Establishing an Exemplarist Virtue Framework for AMAs

Having demonstrated that AMAs can be beneficial in domains with high situational pressure, I will argue for what kind of humanity should strive to build. As established, the focus will be on explicit ethical agents. Whilst lacking the human capacities necessary to be responsible for their actions, they can still perform morally desirable actions without human oversight (Anderson & Anderson, 2007, p. 19), either by entirely taking over tasks or by advising humans. Firstly, I will outline existing approaches to building AMAs and their issues before outlining an alternative exemplarist framework that addresses them. Then, I will defend this framework from potential criticisms like the frame problem, responsibility gaps, and relativism.

4.1.Current Approaches

AMAs are generally built in two ways: top-down and bottom-up (Allen & Wallach, 2009, p. 106). Top-down approaches impose an ethical theory onto a bot and are popular as they can be applied to existing systems. Deontological approaches add moral rules to follow, such as not to lie. Although seemingly intuitive, it disregards outcomes. For example, a self-driving car may be unable to avoid a low-speed crash, either with an elderly person ahead or with a young, athletic person to the side. The rule may be not to steer towards pedestrians, meaning the car would hit the elderly person. However, were it to steer into the young person, they would likely not be seriously injured, whereas the collision could kill the more fragile elderly person. Deontological approaches ignore this, meaning the AMA cannot learn from the outcome. Real-world examples of poor deontological outcomes include Google's Gemini image generator, where images of white people or images with historically accurate diversity levels could not be generated even when explicitly requested (Raghavan, 2024). The rules were implemented to enhance diversity yet had detrimental, offensive outcomes like portraying Nazis as black men, highlighting issues with situational inflexibility.

Alternatively, consequentialist approaches evaluate which action brings the best outcome. Winfield (2014) shows an implementation where a robot can predict all possible outcomes of actions taken in its environment. Its goal is to move to a point whilst avoiding a hole, but if a bot representing a human (H-bot) is likely to fall into the hole, it must prevent this by colliding with them. When a single H-bot headed towards the hole, the bot could predict outcomes and prevent them from falling in. The bot was successful 33-out-of-33 times. However, when a second H-bot was added, the optimal outcome became more difficult to compute. The experiment was again repeated 33 times. In 3 cases, both H-bots were saved; in 16, one was saved, but in 14 cases, neither were saved. Winfield explains that if the bot detects the second H-bot slightly after the first then the consequences being calculated completely change, with the bot "dithering" whilst considering new information. This highlights an issue: outcomes are not always certain, and new factors can drastically alter them, so outcomes alone cannot be relied upon. Related to this is the frame problem, where "potentially every new piece of information may have an impact on the whole cognitive system of an agent" (Misselhorn, 2022, p. 40), meaning it is difficult for top-down AMAs to discern what will change after an action. Therefore, top-down AMAs explicitly following moral rules or targeting specific consequences face fundamental challenges.

Bottom-up approaches do not explicitly employ ethical theories; instead, ethics are developed by learning from experience (Allen & Wallach, 2009, p. 107). This can be done via machine learning (ML) algorithms, giving machines "the ability to learn without explicitly being programmed" (Samuel, 1959). Modern ML models are very good at finding patterns in data and decisions to complete tasks at human levels or better, such as playing video games and language generation (Jordan & Mitchell, 2015). The frame problem is less significant as bottom-up agents learn from all data given to them, using prior learning experience to determine what data is morally relevant. However, without explicit ethical theories, explaining and controlling bottom-up agents' decisions is difficult due to the complexity of ML algorithms, making them impractical.

4.2. Towards an Exemplarist Virtue Framework for AMAs

Having shown issues with top-down and bottom-up approaches, I will now argue for a hybrid, virtue-based approach to AMAs that alleviates these issues. Allen and Wallach (2009, p. 107) suggest that virtue ethics represents a promising hybrid between both approaches, as virtues can be approximately represented top-down by behaviours, whilst moral character is cultivated by learning from experience akin to a bottom-up ML approach. Decisions are made using artificial phronesis where the AMA applies its learnt knowledge to new situations. Therefore, frame problem concerns are reduced as the AMA self-determines what information is morally relevant for emulation. Also, concerns over the inadequacy of general rules or potential outcomes alone are reduced, as virtue ethics enables agents to use their own judgment based on experience in any given scenario. However, few implementations have been attempted, largely due to questions over how to represent virtue (Vishwanath *et al.*, 2022, p. 666). Brewer (2009) argues that virtue is un-codifiable and can only address overall moral character. However, to cultivate virtue in an AMA, virtue must be learnt from virtuous acts, necessitating the approximation of virtue into measurable behaviours. Whilst some nuance is lost, if the AMA can learn to emulate virtuous behaviours, this is sufficient. The question of which virtues should be represented is also challenging, as virtues may be interpreted differently in different settings, and there is no universally agreed set of virtues.

To avoid these issues, I propose that to emulate exemplary human performance in high-pressure situations, an exemplarist framework for training AMAs is appropriate due to its practicality, and I will now show a theoretical outline for such a framework. Rather than explicitly modelling individual virtues, an exemplarist AMA would learn to emulate virtuous behaviour shown by moral exemplars in the target domain/role. The top-down element is that the AMA's ethics are based on moral exemplars exhibiting virtuous behaviours, whilst the bottom-up element is the learning process cultivating these behaviours. This would require human feedback during training, which modern techniques like Reinforcement Learning through Human Feedback (RLHF) make possible. RLHF enables artificial agents to act and learn directly from human feedback as to how desirable an action is (Christiano *et al.*, 2017). With this, an AMA could learn both to perform a general task and to emulate exemplary decision-making in high-pressure situations. Furthermore, the learning process avoids top-down approaches' situational inflexibility and means the AMA can learn from decisions resulting in poor outcomes. Moreover, as ML innovation continues, both phronesis and the data given to the AMA can become more complex and nuanced. For instance, large language models have shown emergent reasoning capabilities (Wang *et al.*, 2024), and Chella *et al.* (2020) show how giving AMAs continuous inner dialogue to explain their reasoning can improve artificial phronesis, also adding further explainability to AMAs, making it easier to understand and challenge their decisions. To avoid unwanted biases, bias auditing tools like Aequitas (Saleiro *et al.*, 2018) could be implemented. The AMAs can also be benchmarked and evaluated against the exemplars' to ensure similarity. Therefore, this framework is promising for implementing AMAs to make moral decisions in high-pressure domains as capabilities are measurable and can potentially increase in the future. Such AMAs are theoretically capable of performing tasks and making moral decisions when required without having to explicitly determine whether the situation is morally challenging or not, as they always seek to emulate their exemplars. Depending on the use case, they could either advise humans or take over tasks. For instance, driving AMAs would drive a car to get to a location, yet would also emulate exemplars' virtue in moral scenarios like unavoidable collisions by determining the exemplar's most likely action. Having outlined how exemplarist AMAs could be implemented, a process is required to identify suitable high-pressure domains where moral decision-making is required, and AMA effectiveness is heavily dependent on selecting high-quality moral exemplars, so I will now elaborate on these elements.

4.3. Selecting Suitable Domains and Exemplars

To identify domains/tasks where high situational pressures diminish moral decision-making, situationist-style analysis can help analyse whether this occurs by setting tasks where situational pressures increase until the task's highest pressure levels are tested. As shown previously, NHS workers reported high MI levels, largely due to situational pressures impacting decision-making, thus highlighting this domain's potential suitability. Driving is another domain where high situational pressure like traffic levels, weather and time-to-react impact decision-making (Soares *et al.*, 2021) and where people make moral decisions that differ depending on situational pressure (Johnson *et al.*, 2023). Therefore, researchers should first seek domains where MI or high situational pressure have been reported. Then, interviews could be conducted with individuals operating in that domain to identify whether there are tasks that involve high-pressure moral decision-making, cause MI, and have measurable virtuous behaviours. If so, situationist-style experiments can be established to measure whether moral decision-making ability diminishes at high pressure. Having shown how to identify potentially suitable domains, I will now illustrate how exemplars could be identified.

Zagzebski (2013) argues that moral exemplars can be identified through admiration, and verifying that they are worthy of imitation. For full, generally intelligent moral agents, locating universal exemplars is difficult. However, for specific domains/tasks, situationist-style experiments can identify exemplary behaviour under pressure. For healthcare workers, this may involve asking staff which of their peers they admire and then verifying whether they are exemplary by, for example, analysing patient satisfaction surveys during periods of very high demand. For driving, Johnson *et al.* (2023, p. 6) suggest that prosocial, cooperative drivers generally align with virtuous traits like benevolence and end up in accidents far less frequently than other drivers. They also found that most drivers' decision to self-sacrifice or self-preserve in unavoidable accident scenarios changed depending on time-to-react, testing this by asking what they would do in a survey and then in real-time simulations. 76.8% of participants self-sacrificed in the survey, but only 22.8% of participants self-sacrificed in both survey and simulation. This highlights that participants admired self-sacrifice, so exemplars can be identified by displaying admirable behaviour under pressure like those 22.8% of participants. Therefore, to identify exemplars, interviews should be conducted with those performing the selected task, asking which peers they admire morally. This could be used in conjunction with situationist-style experiments. Once identified, exemplars' suitability must be verified via audit to ensure they lack unwanted biases, e.g. underestimating black hospital patients' needs (Obermeyer *et al.*, 2019). Having outlined a theoretical framework for virtuous AMA, I will demonstrate a theoretical end-to-end AMA implementation.

4.4. Theoretical Implementation

Having presented the framework, I will show a simple theoretical implementation based on Winfield's (2014) consequentialist bot experiment. The initial set-up is the same, where the moral agent's task is to travel to a point whilst avoiding a hole, and if they notice that a human might fall into the hole, they should display helping behaviour by colliding with them to prevent the fall. Firstly, interviews would be conducted to establish the task, whether high situational pressures can occur, whether there is a risk of MI, and whether there are clearly measurable virtuous behaviours. Here, the task of moving to a point is simple, with the potential for high situational pressure with humans heading towards a hole in the ground. Failing to save someone due to the pressure of the situation could lead to MI, and there is a clearly virtuous behaviour of not hesitating and saving as many humans as possible. Next, situationist tests would be devised to determine whether there is sufficient situational pressure to diminish most humans' moral decision-making skills. For instance, participants are faced with saving one human, then two for heightened situational pressure. Assuming these results are the same as Winfield's results for the consequentialist bot, all participants save the human in the low-pressure scenario. In the high-pressure scenario, 14-out-of-33 participants save no humans, 16-out-of-33 rescue one, and 3-out-of-33 rescue both. Here, virtuous behaviour diminishes with increased pressure, so the task is suitable for AMAs. Exemplars can also be identified as the 3 participants who saved both humans. They may be verified by analysing whether they consistently display these behaviours when performing similar tasks. Next, potential exemplars are audited for unwanted biases. Then, an environment is created with as much information as possible for an AMA to be trained via RLHF to perform the task and similar tasks, such as different routes with different numbers of humans. Chosen exemplars give feedback as to whether the decisions made by the AMA align with what they would do. Once trained, tested and measured against exemplars, the AMA should be able to independently perform the task whilst efficiently processing moral dilemmas that may occur, like which humans to prioritise if not all can be saved or whether multiple humans can be saved, without dithering. It would demonstrate its exemplars' virtuous behaviours without individual virtues being explicitly programmed, and the lack of hard coding means it may more easily adapt to new situations than Winfield's consequentialist bot. Having illustrated a theoretical implementation, I will defend this framework against potential objections that have yet to be considered.

4.5. Objections to the Framework

A key objection regards cultural disagreement over who exemplars are, as different cultures may admire different behaviours (Kotsonis, 2020, p. 228). This is highlighted by Awad *et al.*'s (2018) global survey of responses to moral dilemmas for driving. Whilst some moral preferences were global, many differed culturally, such as the propensity to spare those obeying traffic laws versus jaywalkers. However, whilst cultural preferences may vary, the core approach of developing strong moral character remains, and this framework's goal is not to solve all moral dilemmas universally but to emulate virtuous exemplars whose moral decision-making ability in specific domains withstands high situational pressures. Indeed, Macintyre (1981) argues that virtues must be interpreted by the community using them, and Zagzebski (2013) states that "identification of exemplars is revisable" (p. 200), so exemplars can differ by culture. However, this can raise objections regarding moral relativism, meaning if morals are relative to cultural attitudes, there is no objective morality. Basing moral judgements on exemplars within cultures and domains can seemingly support relativism (Kotsonis, 2020, p. 229). However, this framework does not claim that exemplar's actions are always correct, but that they exhibit virtuous behaviours in specific tasks/domains. For instance, whilst cultural preferences for whose treatment healthcare workers should prioritise may differ, exemplars should still be generally virtuous, e.g. kind and helpful, without their decision-making ability diminishing under pressure, like not neglecting patients despite high stress. Such a virtuous nature is universal, although cultures may interpret specific virtues differently. Macintyre (1981) suggests that reflecting on virtue enables the changing of morals for societies, so universal moral truths can be gradually built towards. Although this does not fully refute relativist objections, this framework's purpose is only to match human morality, not to exceed it, and whilst this significantly challenges the feasibility of cross-cultural AMAs, localised solutions or AMAs designed for specific tasks where there is cross-cultural consensus are still possible.

There may also be objections regarding moral deskilling. Vallor (2015) suggests that offloading tasks to AMAs can result in losing the moral skills required for the task. This would be a major issue if AMAs were to take over too many responsibilities from humans. However, explicit ethical AMAs cannot be responsible for their actions, so humans must critically evaluate them constantly to ensure that they are performing similarly to exemplars, and they require consistent human feedback. Therefore, although some tasks may be passed to AMAs, exemplars will still need to teach them, and these AMAs should only be used where most humans' moral decision-making is already poor.

A related concern is responsibility for AMAs' mistakes. As this framework does not involve full moral agency, responsibility should fall jointly between all parties developing the AMA. However, Sparrow (2007) demonstrates the possibility of responsibility gaps occurring when an AMA is not designed to break an ethical code but does so unforeseeably without human oversight. Therefore, nobody appears responsible for the AMA's action. This should be combated by ensuring that domains/tasks are narrow enough that most general moral scenarios can be addressed in the AMA's training. Then, whether the risks of an AMA failing are worth the potential benefits must be carefully evaluated.

Another objection may be that ML can perpetuate unwanted biases held by exemplars, such as racism or sexism (Fazelpour & Danks 2021). High-profile examples include the aforementioned Gemini case, so experts in domains besides the exemplars', like critical race scholars, feminist theorists and philosophers, should be involved in the selection and training process to ensure potential biases are found and eradicated before deployment. Ultimately, this is not a reason to avoid this approach, but it shows that great care should be taken to avoid perpetuating biases in these systems.

Finally, I will address possible objections to the proposed ML approach. One objection may be that ML algorithms cannot guarantee outputs (Kläs & Vollmer, 2018), so they will not always make decisions in line with their exemplars. Whilst true, certain decisions can be guaranteed by hardcoding deontological rules that override the ML output to comply with certain laws or regulations, such as never deactivating a life support system. Also, situationism shows that human decisions cannot be guaranteed under high pressure, so AMAs emulating exemplars would be more consistent in this regard. This is ultimately virtue ethics' goal, to promote strong moral character, not necessarily to always make the correct moral decision, and AMAs can learn from their mistakes via RLHF to constantly improve. A final practical objection may be that training environments cannot offer all the relevant moral information needed for moral decision-making. However, exemplars could suggest important moral features to capture for given tasks, and practical experimentation is required to determine how specific a task must be and how much information is required for an AMA to accurately emulate exemplars in that task.

5. Conclusion

I have argued that AMAs can be justified by highlighting a specific area where they can be beneficial whilst avoiding existential and feasibility concerns, demonstrating how humans' virtuous behaviours diminish under high situational pressures, potentially leading to MI, therefore justifying AMAs that can match exemplary human performance under high pressure. Additionally, I showed the suitability of an exemplarist, a virtue-based framework for building AMAs to perform moral tasks where high situational pressure impacts human performance and presented a theoretical implementation. Future work could build on and practically test this framework and experiment with training approaches, such as asking exemplars to imagine they are machines when giving training feedback because human and machine morals may not always align. For example, in Winfield's (2014) experiment, if a human were the moral agent preventing others from falling in the hole, self-preservation may also be a factor. However, for this level of AMA, there is no self to preserve, enabling different potential actions like jumping into the hole to reduce the falling distance. Therefore, practical experimentation is required to further develop AMAs, but overall, this paper presents a clear justification and an outline of a theoretical framework for practically applying exemplarist virtue ethics to AMAs.

References

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms or abdication of human responsibility. *Robot ethics: The ethical and social implications of robotics*, pp. 55-68.

Alzola, M. (2008). Character and environment: The status of virtues in organisations. *Journal of Business Ethics,* 78, pp. 343-357.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine,* 28(4), pp. 15-15.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature,* 563(7729), pp. 59-64.

Best, J. (2021). Undermined and undervalued: how the pandemic exacerbated moral injury and burnout in the NHS. *BMJ,* 374(1858).

Brewer, T. (2009). *The retrieval of ethics.* Oxford University Press, USA.

Chalmers, D. J. (2016). The singularity: A philosophical analysis. In *Science fiction and philosophy: From time travel to superintelligence,* pp. 171-224.

Chella, A., Pipitone, A., Morin, A., & Racy, F. (2020). Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7, 16.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems,* 30.

Coimbra, B. M., Zylberstajn, C., van Zuiden, M., Hoeboer, C. M., Mello, A. F., Mello, M. F., & Olff, M. (2024). Moral injury and mental health among health-care workers during the COVID-19 pandemic: meta-analysis. *European Journal of Psychotraumatology*, 15(1), 2299659.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27(1), pp. 100.

de Bruin, B., Zaal, R., & Jeurissen, R. (2023). Pitting virtue ethics against situationism: An empirical argument for virtue. *Ethical Theory and Moral Practice,* 26(3), pp. 463-479.

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous*, 32(4), pp. 504-530.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760.

Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *AI & society*, 36(3), pp. 839-851.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11, pp. 19-29.

Hursthouse, R., & Pettigrove, G. (2023). Virtue Ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Fall 2023).* Metaphysics Research Lab, Stanford University. https://plato-stanford-edu.ezp.lib.cam.ac.uk/archives/fall2023/entries/ethics-virtue/

Johnson, K. A., Berman, S., Pavlic, T. P., Ulhas, S. S., Elkins, J. K., & Ravichander, A. (2023). Virtuous Vehicles: Identifying the Values Profiles of Human Drivers as a Basis for Programming Virtuous Decision-Making in Self-driving Cars.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp. 255-260.

Kläs, M., & Vollmer, A. M. (2018). Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In *Computer Safety, Reliability, and Security: SAFECOMP 2018* Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37 (pp. 431-438). Springer International Publishing.

Kotsonis, A. (2020). On the limitations of moral exemplarism: Socio-cultural values and gender. *Ethical Theory and Moral Practice*, 23(1), pp. 223-235.

Kupperman, J. J. (2001). The indispensability of character. *Philosophy*, 76(2), pp. 239-250.

MacIntyre, A. (2013). *After virtue.* A&C Black.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology,* 67(4), pp. 371.

Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible*

Artificial Intelligence: Interdisciplinary Perspectives. Cambridge: Cambridge University Press, pp. 31–49.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems,* 21(4), pp. 18-21.

Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now,* 72, pp. 12-14.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), pp. 447-453.

Orne, M. T., & Holland, C. H. (1968). On the ecological validity of laboratory deceptions. *International Journal of Psychiatry,* 6(4), pp. 282-293.

Raghavan, P. (2024, February 23). Gemini image generation got it wrong. we'll do better. *Google*. https://blog.google/products/gemini/gemini-image-generation-issue/

Rimmer, A. (2021). Covid-19: Eight in 10 doctors have experienced moral distress during pandemic, BMA survey finds.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint* arXiv:1811.05577.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), pp. 210-229.

Savage, M. (2022, February 26). Stressed NHS staff in England quit at record 400 a week, fuelling fears over care quality. *The Guardian*. https://www.theguardian.com/society/2022/feb/26/stressed-nhs-staff-quit-at-record-rate-of-400-a-week-fuelling-fears-over-care-quality

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences,* 3(3), 417-424.

Soares, S., Lobo, A., Ferreira, S., Cunha, L., & Couto, A. (2021). Takeover performance evaluation using driving simulation: a systematic review and meta-analysis. *European Transport Research Review,* 13, pp. 1-18.

Sparrow, Robert (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), pp. 62–77.

Upton, C. L. (2009). Virtue ethics and moral psychology: The situationism debate. *The Journal of Ethics*, 13(2), pp. 103-115.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28, pp. 107-124.

Vishwanath, A., Bøhn, E. D., Granmo, O. C., Maree, C., & Omlin, C. (2023). Towards artificial virtuous agents: games, dilemmas and machine learning. *AI and Ethics,* 3(3), pp. 663-672.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., ... & Yang, H. (2024). Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint* arXiv:2401.06805.

Williamson, V., Murphy, D., Phelps, A., Forbes, D., & Greenberg, N. (2021). Moral injury: the effect on mental health and implications for treatment. *The Lancet Psychiatry*, 8(6), pp. 453-455.

Winfield, A. F., Blum, C., & Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection. In *Advances in Autonomous Robotics Systems: 15th Annual Conference,* TAROS 2014, Birmingham, UK, September 1-3, 2014. Proceedings 15 (pp. 85-96). Springer International Publishing.

Zagzebski, L. (2013). Moral exemplars in theory and practice. *Theory and Research in Education,* 11(2), pp. 193-206.