# Ethics By Agreement in Multi-Agent Software Systems

Vivek Nallur[a] and Rem Collier[b]

*School of Computer Science, University College Dublin, Ireland*
*{vivek.nallur, rem.collier}@ucd.ie*

Abstract:     Most attempts at inserting ethical behaviour into autonomous machines adopt the 'designer' approach, *i.e.,* the ethical principles/behaviour to be implemented are known in advance. Typical approaches include rule-based evaluation of moral choices, reinforcement-learning, and logic based approaches. All of these approaches assume a single moral agent interacting with a moral recipient. This paper argues that there will be more frequent cases where the moral responsibility for a situation will lie among multiple actors, and hence a designed approach will not suffice. We posit that an emergence-based approach offers a better alternative to designed approaches. Further we outline one possible mechanism by which such an emergent morality might be added into autonomous agents.

## 1 INTRODUCTION

The issue of AI and its effects on society is very topical, and many news articles have been devoted to alarmist concerns[1] as well as optimistic visions[2]. To this end, there have also been many proposals on how to create ethical AI (Bringsjord et al., 2006; Anderson and Anderson, 2015; Dennis et al., 2016; Conitzer et al., 2017; Char et al., 2018; Heidari et al., 2018). However, all of these approaches seem to assume that the AI will be a single machine/software/entity. This is a flawed assumption, both from a philosophical point of view (Floridi, 2013; Heersmink, 2016) as well as a software-engineering point of view. Increasingly, systems that we interact with, are an instantiation of multi-agent systems, whether these are human-machine hybrids (e.g., a human navigating using a GPS/smartphone) or smart machine-to-machine systems (e.g., package tracking, electronic payment processing, etc). Large complex systems, distributed over multiple physical locations, are also typically diverse in their goals, architectures (Song et al., 2015) and strategies (Lewis et al., 2014) to deal with changing environments. Therefore, it is highly likely from a computer science perspective, that the decisive AI behaviours will involve collective responsibility. That is, due to the distributed nature of storage and compu-

tation (`viz.` edge computing, IoT, etc.), the source of im/moral actions (and hence, responsibility) can be distributed as well. Ethics is a behavioural stance, and therefore highly dependent on the agents involved and the context of their interactions. In such cases, the current approach of assuming one locus of control will be inadequate to develop ethically interacting autonomous systems. This paper takes the position that a better approach to inducing and ensuring ethically acceptable behaviour is through repeated social interaction, even for purely software agents. Most approaches that have attempted to insert ethics into autonomous software/hardware have followed a template: choose a moral theory and use programming techniques such as logic-programming, constraint-satisfaction, or rule-based methods to implement the theory. This is conceptually easy for programmers and technologists, however it is very difficult to verify if the implementation is correct, or if it would generalize to other situations. A more serious concern is whether the moral theory chosen (Kant-ian or utilitarian or duty-based) is fit for purpose, to ensure ethical behaviour in an autonomous system (Tonkens, 2009). Yet another concern is that all of these approaches assume that there is only one autonomous machine that forms part of a system. However, it is easy to conceive of situations where there are multiple autonomous machines, interacting with humans. For instance, an elder-care facility can have autonomous robots from multiple companies interacting with multiple elderly people. More conventionally, autonomous cars from

[a] https://orcid.org/0000-0003-0447-4150
[b] https://orcid.org/0000-0003-0319-0797
[1]E.g., https://cnb.cx/2vQpXtM and https://bit.ly/2otrSjZ
[2]https://bit.ly/2Z4k2x9

different companies would need to share the road with each other, and with pedestrians. In such cases, would conventional assumptions of a single moral agent and single moral recipient suffice? The programmer/designer would have to foresee all possible *multi-agent* interactions and ensure that the implemented moral philosophy still achieves its goals.

The paper is structured as follows: we briefly describe the state-of-the-art with regard to technical approaches attempted for making autonomous systems behave ethically in Section 2. In Section 3, we consider the philosophical arguments why morality could potentially be distributed. In Section 4, we describe the foundations of our approach and one possible mechanism for realizing this approach. Since our mechanism is a work-in-progress, we describe no results. We argue, however, that this can potentially be used as a step in the software engineering process, before deploying the system. Finally, we conclude with a discussion of potential pitfalls of both, ethics-by-design and ethics-by-agreement.

## 2 RELATED WORK: ETHICS BY DESIGN

Corning has persuasively argued about the notion of information as a controlling aspect of purposive systems (Corning, 2001). Algorithmic systems can alter the complex web-of-systems in which they participate, through acts of withholding or spreading information. Since these systems are expected to be pervasive in human society, it is reasonable to expect that the cybernetic effects of information have an influence on human society, as well. In such a scenario, the ethical consequences of algorithmic action assume importance, especially if these algorithms are also autonomic in nature. To ensure that autonomic, intelligent systems behave in an ethically acceptable manner, there have been multiple attempts to instil morality/ethics into machines.

There are typically two approaches discussed with regard to embedding an artificial morality (Allen et al., 2005), the top-down approach and the bottom-up approach. The top-down approach depends on the existence of a generally accepted moral theory, such as Kant's categorical imperative or a consequentialist theory (such as utilitarianism). Regardless of which moral theory is preferred by the system designer, the computational intractability of gathering knowledge, processing alternative paths, and decision-making in real-time would result in top-down approaches being less-than-satisfactory. The bottom-up approach considers evolving artificial moral behaviour using techniques such as reinforcement learning (Abel et al., 2016). An important aspect common to both approaches, is that it is only the *engineering* of the system's ethical behaviour that is top-down or bottom-up (emergent). In all of the approaches, the actual desired ethical behaviour itself is human-supplied. This is to be expected, as system designers would want machines to express an ethical behaviour that is in line with human expectations.

One of the first in-depth implementation attempts to embed ethical control and reasoning system in the field of autonomous weapons was presented in (Arkin, 2008). The robots controlling the lethal weapons are assumed to have a reactive/hybrid architecture where a deliberative mechanism was introduced to modulate the response that the robot makes. The intention behind such an effort was to enable a robot to obey Laws of War and Rules of Engagement prescribed by international law. Robot control architecture typically use mappings between stimuli and possible responses to decide how to act. *Constraint-satisfaction* techniques were used to ensure that any plan of action chosen by the robot were always in consonance with moral laws.

(Dennis et al., 2016) used a formal verification technique called *model-checking* to ensure that any plan chosen by the autonomous machine would never result in a state that would be morally repugnant. The authors acknowledge that model-checking is a fairly compute-intensive technique, and may not be able to deal with dynamically changing contexts.

Another formal mechanism was implemented by (Bringsjord et al., 2006), where logic-based ethical governors were used to decide whether actions are permissible, obligatory, prohibited, etc. The governor attempts to arrive at a proof of whether a particular action is at least permissible in the ethical code that it has been loaded with. While this approach, along with the addition of AI-friendly deontic logic, allows for a good explanation of *why* a robot arrived at a particular conclusion, it is unclear how this methodology would function in the presence of contradictions.

(Anderson et al., 2019) have argued for explicit ethical reasoning in robots that are deployed in domains that involve ethical dilemmas. They argue that while it is unclear if humanity has a common system of ethics, in most domains, it is generally well-accepted what ethics robots *ought to have*. To this end, they create a case-supported principle-based methodology where robots generalize from cases that ethicists have already agreed upon and infer the correct behaviour for the specific cases they encounter. This is based on Ross' notion of *prima-facie duties* (Anderson and Anderson, 2007) where the dif-

ferent duties change in priority depending on the circumstances. The case-based reasoner extracts the ethically relevant features from its memory, generalizes principles and continually partitions all actions it could take into partially-ordered subsets. It then chooses the subset which is the most ethically preferred. While this seems reasonable, it is unclear whether such reasoning would work across multiple domains, *i.e.*, would every autonomous machine need re-training as soon as it moved across domains?

In contrast to the top-down approaches discussed, reinforcement learning (RL) attempts to learn the optimal policy for action by trying to maximize the long-term reward that the environment provides. The technique is useful when there are uncertainties in what the autonomous entity is able to observe, and how the world has changed in response to an action. In other words, even when the effect of an action is not immediately apparent, reinforcement learning methods can be used to learn what the best action to take is. (Abel et al., 2016) advocates the use of RL agents to solve POMDPs (Partially Observable Markov Decision Processes) to learn the optimal action to take in the presence of ethical dilemmas. While this has the advantage of not committing the system designer to any particular ethical theory, it still requires the system designer to design utility functions that can be used in the observation function of an agent. Also, the computational intractability of POMDPs that stretch into the future is a hindrance to agents that have to consider long-term consequences of their behaviour.

We refer to the approaches mentioned here as *ethics-by-design*, since the specific ethical obligations, constraints or value functions that the autonomous system must learn are already known by the system designer.

## 3  DISTRIBUTED MORALITY

Multi-Agent systems exist all around us already. From human-machine hybrids, such as humans navigating using a GPS/smartphone and smart factories, to machine-to-machine systems such as fully automated warehouses, package tracking systems, fraud detection systems and electronic stock trading, these MAS are ubiquitous in our economies and daily lives. Each individual agent in these MAS only has access to partial information, and depends on the correct functioning of other agents to complete its task. Even if each individual agent, using approaches previously mentioned, acts morally, there is no guarantee that the *distributed morality* that emerges from the MAS will be acceptable. Distributed morality refers to, quoting

from (Floridi and Sanders, 2004),

> "the macroscopic and growing phenomenon of global moral actions and non-individual responsibilities, resulting from the invisible hand of systemic interactions among multi-agent systems (comprising several agents, not all necessarily human) at a local level."

To simplify, each individual agent's action may/may not be morally significant, but the combined effect of the MAS may have moral implications. In such situations, it is difficult to use current ethical theories to assign moral responsibility to any single agent. We need to define a mechanism that understands *distributed morality* and is able to reason about the duties and responsibilities of agents involved in the distributed system.

This paper does not claim to create a philosophical theory that deals with distributed morality. Rather, it takes the position that the technological approach to implementing any such theory would need to be grounded in social interactions. That is, unlike the approaches mentioned in Section 2, the beliefs, desires and goals of *multiple* agents are relevant to ensuring ethically acceptable outcomes. Given that agents in a MAS are not necessarily designed/implemented/controlled by a single entity, the actions performed by these agents must be constrained by some set of permissions, obligations and prohibitions to ensure the afore-mentioned ethically acceptable outcomes. We shall refer to this set of permissions, obligations and prohibitions as the *code-of-conduct*.

## 4  ETHICS BY AGREEMENT

Now we describe our approach, its philosophical divergence as well as the consequent divergence in implementation. This implementation is a work-in-progress and hence no results are described. Almost all implementations of machine ethics have had a single set of ethics (Anderson and Anderson, 2015) that the robot (or artificial moral agent) attempts to learn, and moral effects of the said action are known to all agents. However, in real-life, this is unrealistic. There are many sub-groups of human society with differing sets of ethics, and principles of action.

**Philosophical Roots:** Our approach has its roots in Humean notion of morality wherein human beings possess both reason as well as passion. According to Hume, we gain awareness of moral good and evil by experiencing the pleasure of approval and the uneasiness of disapproval when we contemplate a character

trait or action from an imaginatively sensitive and unbiased point of view (Cohon, 2018). In this account of human nature, our conception of the ethically correct thing to do, emerges from our interactions with fellow human beings, rather than any demonstrative or probabilistic reasoning. This view implies the ability to have, and act upon, long-term thinking. Evidence in human behaviour of long-term thinking such as the denial of instant self-gratification, investing time and energy in agriculture, saving for old age, etc. all point to the ability and willingness to imagine or conceive of a future, and then take steps to achieve/avoid that future. If we accept Hume's theory that human ethics are socially constructed due to the presence of long-term thinking, then we must also accept the evidence that these social constructs vary from culture to culture due to differences in the long-term socio-economic conditions. Taking the Humean notion of socially constructed ethics to its logical conclusion therefore establishes a reason for the emergence of a non-homogeneous set of ethics among autonomous machines. That is, if diverse ethical standards among human societies could have emerged through a process of implicit bargaining via repeated interactions across several socio-economic conditions, then the emergence of ethical standards among machine societies due to repeated interactions in heterogeneous domains is also likely to exhibit diversity. The difference between the emergence in human societies and machine societies would likely be the presence of explicit proposal, bargaining and agreement mechanisms. The agreement over a standard of behaviour or strategy in a group must sustain over multiple generations for it to be recognized as an ethical standard of that group. This requirement ensures that no standard that is strategically unviable would survive as an ethical standard. The process of reaching ethics by agreement has the advantage over other schools of ethics in that it can be described very simply using evolutionary techniques set in a social domain. From a scientific point of view, the simplicity of the mechanism is very appealing since it can be simulated and tested under various conditions. This emphasis on the process of ethics-formation (as opposed to simply picking a school or a set of ethics) is due to two factors. One, in human history we have not been able to converge on a single unified set of ethics that everyone agrees with. There are no indications that this will change soon. Hence, it is more advantageous to focus on the process of ethics formation, since we can then accommodate the need for differentiated sets of ethics in different domains. Two, even in the same culture, the notions of ethically acceptable behaviour have changed over time. This implies that any ethics implemented

for machines, even while interacting with the same users, could need to change over time. Again, picking a process of ethics-formation allows us to create ethically acceptable behaviour for machines that adapts along with its users, while also being amenable to analysis and prediction.

# 5 EMERGENT ETHICS USING AGREEMENT IN GAMES

A frequent question regarding human society is *why* did fairness emerge in human society? According to anthropologists, the answer is simple (Binmore, 2006). Ensuring fair amount of food-sharing in hunter-gatherer societies allowed humans to stave off the threat of starvation. The more interesting question is *how* did fairness arise? In multiple accounts of pure hunter-gatherer societies, equitability in food-sharing has been observed (Boehm, 2009). For such a social contract to be established as an evolutionarily stable strategy it must be both, efficient (in outcomes for everyone), as well as deviation must be easily punishable. According to the folk theorem of repeated games, reciprocal altruism is a stable strategy if the players know they are going to interact together in the future, and their behaviour can be monitored without too much effort(Trivers, 1971). Thus, the presence of repeated games ensures that notions of a social contract arise spontaneously and persist across generations (Binmore, 2014). The stability of the social contract, it has been argued in (Gauthier, 1986), is a rational outcome of agents mutually agreeing to behave in a certain manner. Thus a code-of-conduct can be a rational outcome for computational agents, using repeated games with feedback loops. Therefore, we propose to situate the modified evolutionary process in a game-theoretic framework. Game theory provides us with the ability to reduce real-world cooperative and competitive problems into a stylized mathematical model, called games. Examples of such games include the Minority Game (Challet and Zhang, 1997), the Iterated Prisoners Dilemma (Binmore, 2006), the Public Goods Game (Isaac et al., 1994), etc. Most game-theoretic literature does not focus on strategies or behaviour across games, however in the real world we constantly switch domains and games, and cooperate or compete depending on context. The switching of games is critical to our experimentation since our ethical standards do not seem to demonstrate any domain-specificity. Hence, any ethical standard that we induce in an artificial agent must also be able to survive across games. Game-playing simulators such as the
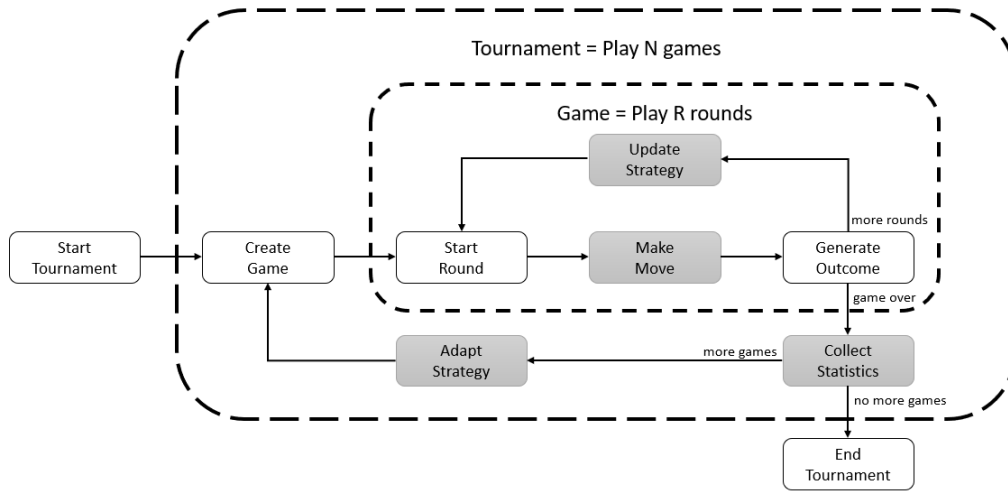
Figure 1: A tournament of diverse games

General Video Game AI Framework (Pérez-Liébana et al., 2016), Arena Framework (Lawlor et al., 2018) or *Stanford General Game Playing Framework*[3] allow for a reconfigurable game environment that allows co-evolution of players that play multiple games in an iterated fashion. The games can vary across multiple dimensions such as number of players (two-player, multi-player), moves (sequential, simultaneous), payoff (zero-sum, non-zero-sum), duration (repeated, one-shot), etc. The evolutionary process of emergent ethics, from an implementation perspective, requires three properties that have previously not been considered as a part of evolutionary processes. We propose the following additions to game playing, to enable the emergence of a code-of-conduct:

1. **A consequence simulation engine for each individual**: There is increasing evidence that human brains tend to simulate future states as way of predicting consequences of their actions(Lake et al., 2016)

2. **An ability for an individual to recognize another individual across multiple interactions**: Players are able to recognize each other across games, thus creating a notion of persistent identity, which can then be used for creating group identity.

3. **A bargaining process that allows each individual to propose a behavioural rule in a context, which can then be accepted or rejected via a negotiation protocol**: This can be used to propose and agree on mutual behaviours, given a certain context. Once the mutual behaviour persists

across repeated games, it becomes a part of the code-of-conduct which is resistant to being violated.

Apart from these properties, we propose to utilise the standard mechanisms employed in simulating an evolutionary process: an objective function to recognize relative fitness, the notion of reproduction/survival of fitter individuals, and the ability to mutate/change behaviour in an attempt to increase fitness.

Figure 1 shows the abstract framework for a tournament, where agents play multiple heterogeneous games with other agents that do not necessarily share the same strategies. For example, a tournament may consist of 100 rounds of the Minority Game followed by 100 rounds of the Iterated Prisoners Dilemma. Further, the participants in this game may be broken down such that 30% are using a random strategy, 40% are using a Tit-for-Tat strategy, and 30% are using a random strategy. The consequence simulation engine is used as a part of the `Update Strategy` process. Agents can use reinforcement-learning (Dosovitskiy and Koltun, 2016), clonal-plasticity (Nallur et al., 2016), or any other learning/adaptation method to adjust to the game that it is playing.

The ability to recognize each other across games, allows for the bargaining process to take place. The negotiated code-of-conduct is encapsulated by the `Adapt Strategy` process which affects agents in the next game. An agent will not deviate from the code-of-conduct that it has agreed on. The outcomes of both, the in-game learning (`Update Strategy`) and out-of-game learning (`Adapt Strategy`), are affected by the individual capabilities and negotiating strength of each agent. This creates an evolution-

---

[3]http://ggp.stanford.edu/notes/overview.html

ary pressure (`Collect Statistics`), where only the code-of-conduct that is beneficial to agent survives across games. *Note:* It is not necessary that the code-of-conduct be the same for all agents in the framework. In fact, we actively expect a diversity of codes to emerge from multiple interactions among agents. The code-of-conduct that persists across a tournament,*i.e.,* can be found among a majority of agents, form the ethics of that agent society.

A foreseeable consequence of such an arrangement is that the society of agents might agree on a code-of-conduct, that is evolutionarily stable, but not palatable to human beings. A possible mechanism of preventing such codes-of-conduct would be through the use of grammatical evolution techniques (Nicolau, 2017), where the grammar can be used to specify illegal genotype constructions.

# 6  ETHICS BY DESIGN VS. ETHICS BY AGREEMENT

Should we embed machines with ethics that we know to be good (ethics-by-design), or should we repose our faith in a *ethics-making* method that leads to ethics emerging by agreement in a society of machines? At first glance, the former appears to be superior to the latter. A consequence of ethics-by-agreement is that we do not know *apriori* what code-of-conduct, the agents will agree on. This is an unpleasant or uncertain outcome that software engineers and system designers would like to avoid. However, ethics-by-design also suffers from problems. The need for explainability makes the use of many kinds of learning processes (e.g., Deep Learning or Reinforcement Learning) problematic because of fundamental problems with these methods in tracing any decision to any specific input or rule. While the use of rules, policies and logics can be used to get around the problem of explainability, we are subsequently confronted by the issue of the system-designers bias in creating the rules, policies and logics, which is difficult to resolve. A more intractable issue is the need for dynamically adjusting to contexts, especially those that have not been foreseen by the designer. As soon as the machine is able to transcend its governing rules and policies, the explainability also suffers. In this scenario, the Humean school of ethics by agreement offers a better consistency between human constructs of ethics and machine-based constructs. If a suitable implementation mechanism [say, using causal networks (Halpern, 2016)] was able to offer explanatory power to the ethical standards reached, then ethics-by-agreement might prove to be a more robust way of implementing machine ethics.

# 7  CONCLUSIONS

There are several unexamined assumptions in this paper, not least of which is, to paraphrase Floridi (Floridi, 2011) — Can and should artificial agents have ethics? While the philosophical aspect of that conundrum is still open for debate, this paper takes the position that from a computer science point of view, the mechanisms of adding ethical behaviour must be investigated. The paper also takes the position, that ethics-by-agreement is an interesting mechanism to use for adding ethical behaviour. It is difficult to conclusively state that, in all cases, ethics-by-agreement is better/worse than ethics-by-design. In domains, where ethicists all agree on the correct behaviour, and where the patterns of interactions are well-known, ethics-by-design might be an acceptable mechanism. However, in domains where there are multiple agents, and there are multiple patterns of interaction, ethics-by-agreement is a promising mechanism for ensuring that the emergent distributed morality is acceptable to human society.

# REFERENCES

Abel, D., MacGlashan, J., and Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In Bonet, B., Koenig, S., Kuipers, B., Nourbakhsh, I. R., Russell, S. J., Vardi, M. Y., and Walsh, T., editors, *AAAI Workshop: AI, Ethics, and Society*, volume WS-16-02 of *AAAI Workshops*. AAAI Press. 978-1-57735-759-9.

Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3):149–155.

Anderson, M. and Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–26. Association for the Advancement of Artificial Intelligence Winter 2007.

Anderson, M. and Anderson, S. L. (2015). Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot: An International Journal*, 42(4):324–331.

Anderson, M., Anderson, S. L., and Berenz, V. (2019). A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm. *Proceedings of the IEEE*, 107(3):526–540.

Arkin, R. C. (2008). Governing lethal behavior. In *Proceedings of the 3rd international conference on Human robot interaction - HRI*. ACM Press.

Binmore, K. (2006). The origins of fair play. Report, Papers on economics and evolution.

Binmore, K. (2014). Bargaining and fairness. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10785–10788.

Boehm, C. (2009). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.

Bringsjord, S., Arkoudas, K., and Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44.

Challet, D. and Zhang, Y.-C. (1997). Emergence of co-operation and organization in an evolutionary game. *Physica A: Statistical Mechanics and its Applications*, 246(3-4):407–418.

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care — addressing ethical challenges. *New England Journal of Medicine*, 378(11):981–983.

Cohon, R. (2018). Hume's moral philosophy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., and Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *AAAI*, pages 4831–4835.

Corning, P. A. (2001). Control information the missing element in norbert wieners cybernetic paradigm? *Kybernetes*, 30(9/10):1272–1288.

Dennis, L., Fisher, M., Slavkovik, M., and Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14.

Dosovitskiy, A. and Koltun, V. (2016). Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*.

Floridi, L. (2011). On the morality of artificial agents. In Anderson, M. and Anderson, S. L., editors, *Machine Ethics*, pages 184–212. Cambridge University Press.

Floridi, L. (2013). Distributed morality. In *The Ethics of Information*, pages 261–276. Oxford University Press.

Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, 14(3):349–379.

Gauthier, D. (1986). *Morals by agreement*. Oxford University Press on Demand.

Halpern, J. Y. (2016). *Actual Causality*. MIT Press.

Heersmink, R. (2016). Distributed cognition and distributed morality: Agency, artifacts and systems. *Science and Engineering Ethics*, 23(2):431–448.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2018). A moral framework for understanding of fair ml through economic models of equality of opportunity. *arXiv preprint arXiv:1809.03400*.

Isaac, R. M., Walker, J. M., and Williams, A. W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of public Economics*, 54(1):1–36.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2016). Building machines that learn and think like people.

Lawlor, F., Collier, R., and Nallur, V. (2018). Towards a programmable framework for agent game playing. *Adaptive Learning Agents Workshop at AAMAS 2018*.

Lewis, P. R., Goldingay, H., and Nallur, V. (2014). It's good to be different: Diversity, heterogeneity, and dynamics in collective systems. In *Self-Adaptive and Self-Organizing Systems Workshops (SASOW), 2014 IEEE Eighth International Conference on*, pages 84–89. IEEE.

Nallur, V., Cardozo, N., and Clarke, S. (2016). Clonal plasticity: a method for decentralized adaptation in multi-agent systems. In *Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, pages 122–128. ACM.

Nicolau, M. (2017). Understanding grammatical evolution: initialisation. *Genetic Programming and Evolvable Machines*, 18(4):467–507.

Pérez-Liébana, D., Samothrakis, S., Togelius, J., Schaul, T., and Lucas, S. M. (2016). Analyzing the robustness of general video game playing agents. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE.

Song, H., Elgammal, A., Nallur, V., Chauvel, F., Fleurey, F., and Clarke, S. (2015). On architectural diversity of dynamic adaptive systems. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, volume 2, pages 595–598. IEEE.

Tonkens, R. (2009). A challenge for machine ethics. *Minds & Machines*, 19(3):421 – 438.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, 46(1):35–57.