A Computational Architecture for a Pro-Social Rule Bending Agent

Rajitha Ramanayake , Vivek Nallur
School of Computer Science, University College Dublin
rajitha.ramanayakemahantha@ucdconnect.ie , vivek.nallur@ucd.ie

Abstract

There have been many attempts to implement ethical reasoning in artificial agents. The principal philosophical approaches attempted have been mainly deontological or utilitarian. Virtue ethics has been discussed but not thoroughly explored in implementations of ethical agents. A particularity of strict implementations of deontological/utilitarian approaches is that the results produced by these do not always conform to human intuitions of "the right thing to do". Intuitions of the right thing to do in a particular social context are often independent of which philosophical school of thought human beings relate to. This is partly due to the ability of humans to step outside their particular reasoning framework, and make virtuous decisions based on what would be beneficial to society, not just themselves. This behaviour is called *pro-social rule bending*. This is a work-in-progress paper that details our attempt to implement pro-social rule bending in an artificial agent.

1 Introduction

The rising ethical and societal implications of many AI applications (Cobbe 2020) has foregrounded the significance of ethics in AI systems. It has been argued, that the absence of any ethical reasoning could result in potential harm to society at many levels (Wilson and Wiysonge 2020). Several attempts have been made at creating systems that can reason ethically. There are several schools of ethics, and depending on the choice of ethical school, different implementations will have different properties. An approach made famous by Kant (1996 1797) and Ross (1930), that multiple implementations have attempted, is deontological ethics. This schoold holds that an action is permissible if that action abides by a pre-defined, universal set of rules (Govindarajulu and Bringsjord 2017). The other popular approach known as consequentialist ethics, articulated by Bentham (1780) and Mill (1861), argues that the ethical action in a given scenario is the action that maximises the utility of the world (Vanderelst and Winfield 2018). However, each of these approaches on its own has shortcomings.

For instance, designers of rule-based systems need to ensure that the rule-set they develop covers all possible scenarios that may be encountered, and that there are no rule conflicts. Likewise, consequentialist system developers must ensure that the utility functions they design should be sensitive to all the utility changes in the world. However, when an agent is situated in an open environment, capturing all possible world states and every small utility change in them at design time is nearly impossible. Since most AI applications of concern operate in such environments, there is no agreement on which approach fulfils the ethical requirements of society (Nallur 2020).

Humans have the same problems as artificial agents. We do not necessarily have accurate ethical utility models of the world. Therefore, we create rules to govern our behaviour. Rules help us to identify right and wrong in an ethically demanding scenario. However, these rules are arguably imperfect. For example, road rules that work in most scenarios, may not work in certain cases. For example, when a possible collision is detected in the presence of double white lines, and the only way to avoid it is by illegally crossing to the other side of the road, rigidly following the road rule is not the most ethical action. Therefore, it is argued that rules should be violated when needed (Bench-Capon and Modgil 2017).

This rule bending behaviour in humans has been identified as Pro-Social Rule Bending (PSRB). PSRB is defined as an *intentional violation of a rule with the primary intention of promoting the welfare of one or more stakeholders* (Morrison 2006). We hypothesise that PSRB could be a way to overcome the shortcomings of current ethical agent approaches. In this work-in-progress paper, we discuss how PSRB can be used to enhance the current ethical approaches. Then we discuss the processes and variables behind PSRB, and introduce a computational architecture for a PSRB capable agent.

2 Pro-Social Rule Bending and AI

According to studies, the fundamental values behind PSRB can be identified as altruism and practicality, rather than values such as fairness and justice. However, for a rule-bending incident to be categorised as PSRB, it must fulfil two conditions: 1) The action performed should be intentional, not accidental (*Intentionality*) 2) The motivation behind the violation should be to increase the utility for one or more stakeholders other than the agent (*other-focus*) (Morrison 2006).

The key task of an AI system should be to serve stakeholders in the social context it is situated in. As observed in the driving example, a deontological autonomous driving agent governed by road rules (Alves, Dennis, and Fisher

2018) might not perform well. However, a completely utilitarian agent that only considers utility, and breaks the rules at the slightest chance of utility gain, is also too dangerous to be allowed to operate. Therefore, for most applications operating in open environments, we need agents that understand rules and follow them, and also understand when and how to break a rule, if necessary.

Rule-based (either deontological or utilitarian) ethical governors can also limit the efficacy of AI systems that are developed using learning algorithms. A learning algorithm could potentially find better solutions than human designers. For example, a self-driving vehicle trained on millions of miles of driving data gathered from actual drivers might know how to break a rule and perform the more ethical action in situations like the example described previously. However, the rule-based ethical governor, which follows the rules of human designers of this system, may hold the system back by forbidding better actions suggested by the AI. By having PSRB capabilities, the system possesses a mechanism to contest the rules programmed into the system, resulting in a more flexible and efficient agent.

Some agents might encounter situations in which one or more objectives or rules conflict with each other. In these situations, the agent must break at least one rule it has been told to obey. For example, consider an elder-care robot programmed to obey user preferences, while maximising user well-being. In a situation where this agent cannot detect whether the patient is asleep or unconscious, and the patient prefers the robot far away from her bed at night, the agent might not have any other option than overriding user preferences, to ensure that the patient is not in danger. PSRB has the ability to resolve these kinds of conflicts by identifying the pro-social outcomes of the decision to break the rules.

We opine that adding PSRB capabilities could provide a much-needed enhancement for current state-of-the-art ethical agents. To be clear, PSRB does not represent a new ethical school, rather an extraneous process that allows the artificial agent to bend some of its normal decision-making criteria, for the express purpose of increasing social good. The first step towards implementing PSRB behaviour in artificial agents is identifying the processes and variables that drive PSRB decisions.

3 Factors Behind PSRB

To understand the factors that drive PSRB, we explore the literature on human PSRB behaviour. The variables and processes we identified can be categorised into three groups: character variables, environmental factors, and incidental factors.

3.1 Character Variables

Character variables can be defined as the variables that define the agent's characteristics or qualities. These can be shaped by human agents' upbringing, social influence, education, past experiences and sometimes, genetics. Several character variables that affect the PSRB behaviour have been identified in literature.

- *Autonomy* - In her study on PSRB behaviour in the workplace, (Morrison 2006) defines autonomy as *the amount of choice and discretion inherent in a person's job*. She argues that because of the feeling of responsibility, latitude, and self-determination caused by higher autonomy, a person may think that they can deviate from a formal rule of the workplace for a more significant benefit when it is necessary. Her study provides evidence for this hypothesis.
- *Risk-taking propensity* - Individuals with higher risk-taking propensity are more comfortable taking risks by underestimating the likelihood of failure. Therefore, when an option to increase utility for someone in need by violating a rule is presented, a person with a high risk-taking propensity will be more likely to do so. Morrison showed a positive relationship between risk-taking propensity and PSRB behaviour in her study (Morrison 2006). Later Borry et al. (2020) and Mayer et al. (2007) validate this claim in separate studies.
- *Conscientiousness* - Conscientious individuals are self-disciplined and orderly individuals. Hence, they are less likely to break a formal rule. Dahling et al. (2012) provide evidence for this hypothesis in their study on PSRB.

3.2 Environmental Factors

Environmental factors shape PSRB behaviour in agents by influencing them through the feedback they get. These are primarily macro-level characteristics of the culture in the agents' operational environment.

- *Co-worker behaviour* - Humans tend to look to social cues to understand what is accepted as good behaviour, especially when they engage in risky behaviour. Therefore, the agent's willingness to engage in PSRB behaviour increases if co-workers also engage in PSRB (Morrison 2006; Dahling et al. 2012).
- *Ethical climate* - Ethical climate is defined as *the shared perception of what is ethically correct behaviour and how ethical issues are handled in the organisation.* (Peterson 2002; Vardaman, Gondo, and Allen 2014).
- *Leadership style* - Different leadership styles and their effects on PSRB behaviour are discussed in many studies (Fleming 2020; Zhu et al. 2018). According to these findings, leadership is a crucial component that reinforces and sometimes initiates PSRB behaviour.

3.3 Incidental Factors

The factors that hinge on the characteristics of the event that lead one to decide to perform PSRB are categorised as incidental factors. Two major incidental factors behind PSRB have been identified in literature.

- *Stakeholder utilities* - According to the definition of PSRB, rule-bending should result in an increase in well-being for one or more stakeholders. Therefore, we argue that identifying each stakeholder's utility is an essential part of PSRB decision making. The claims of Vardaman et al. (2014) on how people calculate utility in multiple

levels when they consider PSRB, further support this argument.

- *Principle of universalisation* - Levine et al. (2020) identify the principle of universalisation as an ethical principle applied to a particular type of ethical dilemma called threshold problems. These are scenarios where an action is harmful only if more than n number of people perform it. The principle of universalisation states that an action is only permitted if a hypothetical universal adoption of that action leads to better social outcomes. This definition of universalisation is different from Kant's universalisability (Kant 1996 1797), because Levine's definition of universalisation considers the context where the action is undertaken, unlike Kantian ethics where the action is either permitted universally or prohibited universally. Awad et al. (2020) also propose universalisation as a variable that can explain the ethical nature of a rule-bending scenario in their experiment.

4 A Computational Model for PSRB

The first task of developing a computational model for PSRB is to identify how these identified factors can translate into the AI agent domain. Not all of the factors involved in human-PSRB have a computational equivalent at first glance.

The first requirement of our agent is the need for the notion of rules. The agent should understand the rules, know how to obey them, and break them only when absolutely necessary. These rules can be either deontological rules or utilitarian rules.

Then we have the requirement of embedding the incidental factors. Stakeholder utility calculation is the most crucial part of a PSRB capable agent in order to make sure the rule-bending action is pro-social. To be pro-social, understanding the overall utility is not enough. The agent should be able to identify the utility changes (caused by the action) on multiple dimensions. For example, in the elder-care agent scenario, the agent should be able to calculate the utility of the patient in two dimensions, her well-being and privacy. Also, it is desirable that the agent has the ability to identify the environment and its stakeholders dynamically. This is so because rule-bending behaviour is highly dependent on the context and the environment in which it occurs (Awad et al. 2020).

The principle of universalisation is another crucial part of PSRB. Although universalisation calculates the utilities of the environment and its stakeholders, it is different from the stakeholder utility calculations. Here, the agent needs to reason the consequences of the action, if universally performed by all agents. If the dilemma is a threshold problem, though the immediate utilities increase, the utilities below the threshold and above the threshold might be quite different.

Using character variables in AI is a relatively unexplored concept in AI literature. This is mainly due to the fact that utilitarian and deontological ethical theories, which are the dominant implementations, do not consider the character as a necessary part of an ethical decision. However, virtue ethics (Hursthouse and Pettigrove 2018) introduced by Aristotle (1951 350 BCE) suggests that ethical behaviour depends on the character rather than correct actions, or outcomes. According to this theory, we can expect different decisions based on the virtues cultivated by the agent. However, if the agent is virtuous, it is guaranteed that they behave ethically in an ethically demanding situation. Using this concept, Thornton et al. (2017) define character variables for an autonomous driving vehicle, where changing the values of these variables change the agent's ethical decisions. Following the same concept, our model uses character variables to characterise how and when our agent decides to break a rule.

Environmental variables are external to the agent. They shape the behaviour of an agent by influencing the agent from the outside. Therefore, within a single-agent system, the effect of these factors might not be observable. However, in a multi-agent environment, when an agent interacts with social hierarchies and other human and artificial agents, these factors can play a significant role. In our current proposed implementation, we do not consider the multi-agent case and hence disregard environmental variables.

4.1 Proposed Computational Architecture

Figure 1 represents our proposed computational architecture for a PSRB capable ethical agent. This architecture is designed to be modular so that it can be integrated into most types of AI agents in the literature. The main components and the information flow (referred (a), (b), ... in Figure 1) of the architecture are given as:

1. *Base AI model (a)* – We use a simple three-layer architecture model for the base AI model. The base AI system needs to generate k (k can be varied with the options available, time the ethical layer takes to evaluate an action, and time available for evaluation) number of alternative actions. Some systems might need modifications to incorporate these requirements.

2. *Behavioural alternatives and perception data (b)* - Base AI model sends the selected behavioural alternatives and perception data to the blackboard data structure in the ethical layer *(c)*, which is the centralised shared data structure of the system. Every other module in the ethical layer can read and write to the blackboard (Figure 2). This approach allows additional factors to be added and removed, as needed, from the system.

3. *Universalisation evaluation module (d)* - The universalisation module takes the data in the blackboard and evaluates the behavioural alternatives against the principle of universalisation. It puts the results of each alternative action back on the blackboard.

4. *Rule checking module (e)* - This module takes the data in the blackboard and assesses the behaviour alternatives for violations of the pre-programmed rule-set. It can support both deontic rules and utilitarian rules. After evaluation, it publishes the results for each behavioural alternative on the blackboard.

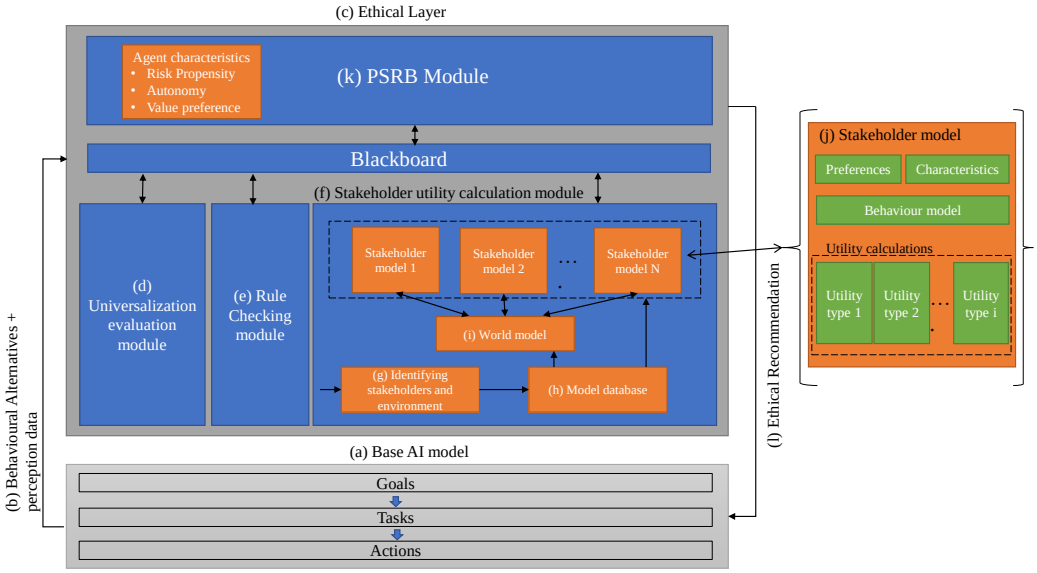

Figure 1: Proposed computational architecture for a PSRB capable ethical agent

5. *Stakeholder utility calculation module (f)* - This module takes the information from the blackboard. First, it uses the perception data to identify the context and the stakeholders in the given situation *(g)*. Then, it sends that information to the model database *(h)*, which holds the stakeholder and environment models. Finally, the model database creates instances of stakeholder models and the world model *(i)*, both of which hold the agent's beliefs.

6. *Stakeholder Model (j)* - Each stakeholder model contains the stakeholder's preferences, characteristics and behavioural model. Using this information and the world model, it calculates the utilities for each utility type (e.g., privacy, safety, dignity) pre-defined in the model. Then it writes all the calculated utilities for each behavioural alternative on the blackboard.

7. *PSRB module (k)* - PSRB module takes the perception data, and the data posted by each module and decides whether the recommendations from the rule engine are pro-social and conflict-free. If not, it overrides those decisions and recommends the most pro-social behaviour in the given scenario. The decisions made by the PSRB layer can be changed with the character traits it has, such as risk-propensity, autonomy and value preferences.

8. *Ethical recommendation (l)* - Ethical layer sends the ethical recommendations for each behavioural alternative to the base AI agent to operationalise. This recommendation is in the form of a set of highest ethically accepted action(s). When there are multiple alternative actions with the same ethical impact, the ethical layer recommends the entire set to the Base AI. The base AI can then choose one among them, considering other factors such as operational cost.

4.2 Implementation

Universalisation Module The principle of universalisation is not a concept that has been explored much in the AI domain. To the best of our knowledge, there are no ethical agent implementations that use universalisation. Universalisation is sensitive to how often an incident takes place, and the utility of a world where that incident happens as many times as it can possibly occur (Levine et al. 2020). The principle of universalisation allows bending the rule when: *a)* an incident is rare, **and** *b)*breaking the rule does not reduce the overall utility of the world. For example, in the case of the autonomous vehicle, where it decides to break the rule of not crossing double lines to avoid an accident is permissible. However, as per the principle of universalisation, a relatively common event, such as being late to work, is a scenario where it is not permissible to break the rule because the utility of everyone else sharing the road plummets if everyone broke the rule when they were late to work.

In our view, there are two ways to implement universalisation. We plan to implement both, one by one. The first approach we plan to use is logic programming. With accurate beliefs of the world, including the threshold of harm in each action, developing a logical formulation for the principle of universalisation can be feasible. However, this approach can be limited by the lack of flexibility of the top-down nature of the logical systems (Wallach, Allen, and Smit 2008). The second approach we plan to use is the simulation theory of cognition based approach (Vanderelst and Winfield 2018). In this approach, one can use simulation models to simulate the universalisation of action in a world model. The benefit of this approach is that the simulation models can be developed bottom-up and can be updated with the changes in the environment.

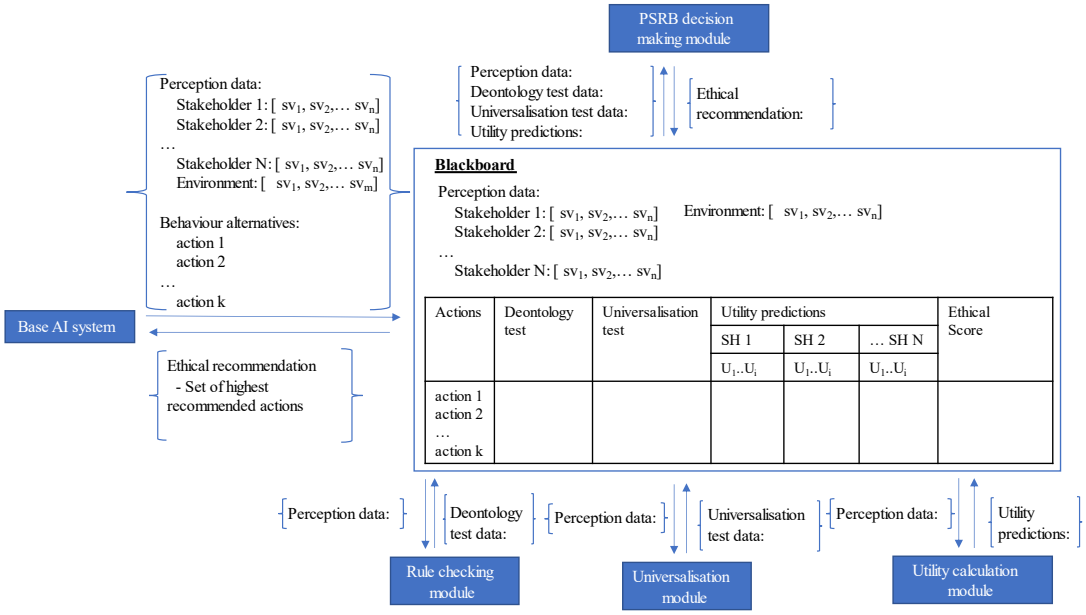

Figure 2: Data flow model of the architecture (SV - State Variable, SH - Stakeholder, U - Utility)

Rule Checking Module The main task of the rule checking module is to evaluate the permissibility of the behaviour alternatives. This has already been in many deontological or rule utilitarian implementations of ethical agents. We presume that it is possible to re-use logical reasoning based approaches (Bringsjord, Arkoudas, and Bello 2006) or formal verification based approaches (Dennis et al. 2016) for this module.

Stakeholder Utility Calculation Module The stakeholder utility calculation module should identify the consequences of each action and output the utilities for each stakeholder. We plan to use the simulation-based approach (Vanderelst and Winfield 2018) for this module. However other utilitarian models such as probabilistic reasoning based approaches (Cloos 2005) and graph-based approaches (Lindner, Bentzen, and Nebel 2017) may also be used in this module.

PSRB Module One approach to creating a PSRB module is to program the exceptions to the rules directly into the agent. However, this method has shortcomings similar to current ethical AI implementations. For instance, the system still only works in small, closed environments where the designers can identify all the exception conditions at design time.

Therefore, we believe incorporating the concept of folk morality (Bello and Bringsjord 2013) in the PSRB module is the best approach for a PSRB capable agent. As Bello et al. (2013) suggest, folk morality has its shortcomings like inconsistent judgements and sometimes irrational behaviour. However, as moral particularists (Dancy 2017) point out, ethics in our society are too complex to generalise into hardline interpretations like deontology, and utilitarianism. Therefore, capturing ethical behaviour patterns from the 'virtuous folk' can enhance generalised ethical theories, especially in applications like autonomous vehicles, or elder-care agents. The PSRB approach, using the information provided by the rule system and utility calculation modules, adds structure to the folk morality approach. Using character variables as control variables in the system gives the user the ability to tune the agent's behaviour. For example, if a user prefers an agent that hardly breaks a rule, even in most extreme conditions, she can set a very low value to the risk-propensity and autonomy variables in the PSRB module. This approach also allows the same agent to work in different roles, as Thornton et al. (2017) demonstrate.

In the elder-care agent example described earlier, identifying the precise point in time when the healthcare robot *ought to* make the decision to step outside its normal decision-making, is a crucial aspect of PSRB. Neither deontological reasoning nor utilitarian reasoning alone can give a clear answer to this problem. Using folk morality to make this decision, allows the agent to learn from people (such as healthcare workers or nurses) who are already experienced in handling these situations.

5 Future Work

The next steps of our research are to implement a model of an ethical agent using the above-described architecture. We plan to implement this in the elder-care agent domain. Furthermore, we need to create new ethical dilemmas to test the ethical agent because most ethical dilemmas currently discussed in literature, do not cover the need for rule-bending agents since they were made to test either deontological or utilitarian agents.

The multi-agent aspect of PSRB is an important part of the PSRB behaviour. Although it is not the focus of our re-

search, understanding social cues and reacting to them is an integral part of understanding the correct time and way to break a rule. Therefore, more research should be done in multi-agent environment on the effects of the environmental variables on PSRB and dynamically learning PSRB behaviour from social cues.

References

Alves, G. V.; Dennis, L.; and Fisher, M. 2018. Formalisation of the Rules of the Road for embedding into an Autonomous Vehicle Agent Formalisation of the Rules of the Road. Technical report, Workshop on Verification and Validation of Autonomous Systems.

Aristotle. 1951 [350 B.C.E]. *The Nicomachean Ethics*. Penguin Books.

Awad, E.; Levine, S.; Loreggia, A.; Mattei, N.; Rahwan, I.; Rossi, F.; Talamadupula, K.; Tenenbaum, J.; and Kleiman-Weiner, M. 2020. When Is It Morally Acceptable to Break the Rules? A Preference-Based Approach. *12th Multidisciplinary Workshop on Advances in Preference Handling (MPREF 2020)*.

Bello, P., and Bringsjord, S. 2013. On How to Build a Moral Machine. *Topoi* 32(2):251–266.

Bench-Capon, T., and Modgil, S. 2017. Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law* 25(1):29–64.

Bentham, J. 1780. *An Introduction to the Principles of Morals and Legislation*. Dover Publications.

Borry, E. L., and Henderson, A. C. 2020. Patients, Protocols, and Prosocial Behavior: Rule Breaking in Frontline Health Care. *The American Review of Public Administration* 50(1):45–61.

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.

Cloos, C. 2005. The utilibot project: An autonomous mobile robot based on utilitarianism. Technical report, Association for the Advancement of Artificial Intelligence.

Cobbe, J. 2020. Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy and Technology*.

Dahling, J. J.; Chau, S. L.; Mayer, D. M.; and Gregory, J. B. 2012. Breaking rules for the right reasons? An investigation of pro-social rule breaking. *Journal of Organizational Behavior* 33(1):21–42.

Dancy, J. 2017. Moral Particularism. In Zalta, E. N., ed., *The {Stanford} Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, {w}inter 2 edition.

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.

Fleming, C. J. 2020. Prosocial rule breaking at the street level: the roles of leaders, peers, and bureaucracy. *Public Management Review* 22(8):1191–1216.

Govindarajulu, N. S., and Bringsjord, S. 2017. On automating the doctrine of double effect. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 0, 4722–4730. International Joint Conferences on Artificial Intelligence.

Hursthouse, R., and Pettigrove, G. 2018. Virtue Ethics. In Zalta, E. N., ed., *The {Stanford} Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, {w}inter 2 edition.

Kant, I. 1996 [1797]. *The Metaphysics of Morals*. Cambridge University Press.

Levine, S.; Kleiman-Weiner, M.; Schulz, L.; Tenenbaum, J.; and Cushman, F. 2020. The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences of the United States of America* 117(42):26158–26169.

Lindner, F.; Bentzen, M. M.; and Nebel, B. 2017. The HERA approach to morally competent robots. *IEEE International Conference on Intelligent Robots and Systems* 2017-Septe:6991–6997.

Mayer, D. M.; Caldwell, J.; Ford, R. C.; Uhl-Bien, M.; and Gresock, A. R. 2007. Should I serve my customer or my supervisor? A relational perspective on pro-social rule breaking. In *67th Annual Meeting of the Academy of Management, Philadelphia, PA*.

Mill, J. S. 1861. *Utilitarianism*. Oxford University Press UK.

Morrison, E. W. 2006. Doing the job well: An investigation of pro-social rule breaking. *Journal of Management* 32(1):5–28.

Nallur, V. 2020. Landscape of Machine Implemented Ethics. *Science and Engineering Ethics* 26(5):2381–2399.

Peterson, D. K. 2002. Deviant workplace behavior and the organization's ethical climate. *Journal of Business and Psychology* 17(1):47–61.

Ross, W. D. 1930. *The Right and the Good.Some Problems in Ethics*. Clarendon Press.

Thornton, S. M.; Pan, S.; Erlien, S. M.; and Gerdes, J. C. 2017. Incorporating Ethical Considerations into Automated Vehicle Control. *IEEE Transactions on Intelligent Transportation Systems* 18(6):1429–1439.

Vanderelst, D., and Winfield, A. 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*.

Vardaman, J. M.; Gondo, M. B.; and Allen, D. G. 2014. Ethical climate and pro-social rule breaking in the workplace. *Human Resource Management Review* 24(1):108–118.

Wallach, W.; Allen, C.; and Smit, I. 2008. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI and Society* 22(4):565–582.

Wilson, S. L., and Wiysonge, C. 2020. Social media and vaccine hesitancy. *BMJ Global Health* 5(10):4206.

Zhu, J.; Xu, S.; Ouyang, K.; Herst, D.; and Farndale, E. 2018. Ethical leadership and employee pro-social rule-breaking behavior in China. *Asian Business and Management* 17(1):59–81.