Pro-Social Rule Breaking as a Benchmark of Ethical Intelligence in Socio-Technical Systems

Abstract: The current mainstream approaches to ethical intelligence in modern socio-technical systems have weaknesses. This paper argues that implementing and validating Pro-Social Rule Breaking behaviour can be used as a mechanism to overcome these weaknesses and introduce a sample scenario that can be used to validate this behaviour.

Keywords: Machine ethics, Pro-social rule breaking, Socio-technical systems

1. Current Approaches to Ethical Agents

Incidents resulting in unethical outcomes from interactions with socio-technical systems (STS) (Blue, 2017; Boyle, 2016) have resulted in attempts to develop ethical Artificial Intelligence (AI) systems, or at the very least, implement some form of ethical reasoning in AI-based systems. Most implementation approaches either follow Deontological ethics (Alexander & Moore, 2020), which argues that an action's ethicality depends on whether it follows a particular universal rule-set, or Consequentialist ethics (Sinnott-Armstrong, 2019), where an action is ethical if it maximises the utility of the world. Deontological implementations (Bringsjord et al., 2006) of artificial ethical agents primarily use a rule-based approach to decide the ethical action in a given situation, and Consequentialist implementations use world models and consequence engines (Lindner et al., 2017; Vanderelst & Winfield, 2018) to decide what is ethical.

However, it is still debatable whether these implementations make reliable, ethical agents. This is due to the shortcomings of the design approaches we currently use (Pirni et al., 2021) or the practical challenges to implementing ethical features such as multi-objectivity and proactivity (Dennis & Fisher, 2018). Therefore, there is no consensus on which approach would result in a satisfactorily ethical AI system (Nallur, 2020). To enhance the ethical abilities of current AI systems, we propose to adapt the ethical augmentation mechanism called *pro-social rule breaking*.

2. Pro-Social Rule Breaking

The average human seems to function effortlessly within rule-based systems (for the purpose of our discussion, a rule can be both a deontic rule such as "do not lie", or a utilitarian rule such as "if action *a* leads to a better consequence for person *p* than action *b*, then *a* is permissible") such as road rules, even when they are not optimal for every scenario. Usually, rule-breaking is frowned upon in society. However, when facing situations where following the rules does not result in ethical outcomes, humans tend to break the rules. For example, a university lecturer who typically penalises late submissions, may opt to suspend the late-submission rule if a student has extenuating circumstances, such as bereavement in the family.

This behaviour has been identified and termed pro-social rule breaking (PSRB) by Morrison (2006). She defined PSRB behaviour as an *intentional violation of rules to promote the welfare of one or more stakeholders.* Morrison's research found that 60% of rule-breaking cases are pro-socially motivated. Several studies suggest (Borry & Henderson, 2020; Morrison, 2006; Vardaman et al., 2014) that the reasons behind PSRB vary from rules not matching stakeholder needs, to improved outcomes for stakeholders.

Based on the evidence of how humans use PSRB to compensate for the shortcomings of the rule systems, we argue that this PSRB behaviour should be a requirement for ethically sound, AI-enabled socio-technical systems that need human-like decision-making abilities. *Note*: we do not suggest that PSRB is sufficient to ensure ethical behaviour in a system; rather that it is a necessary component of it.

When an AI agent such as an autonomous vehicle operates on public roads, it is infeasible to predict all possible scenarios it will ever face. To assure safety, we can implement an ethical governor that enforces road safety rules on the agent (Alves et al., 2018). However, there could be a scenario where following road rules does not lead to the best outcome. For example, when a vehicle detects the likelihood of an accident if it stays within the double white line, the safest thing to do might be to break the rules, and cross the double white line. A rule-enforcing governor would not allow this behaviour. However, if we used a PSRB capable ethical governor, it would allow the vehicle to break the

rule in this particular situation. On the other hand, the ethical governor would also understand that reasons like being late to work are not good enough to break such a rule even though it increases the utilities of some stakeholders. Therefore, we posit that an ethical governor *aided by a PSRB process* would be a better way to implement an ethically aware autonomous agent. Furthermore, we believe that the PSRB behaviour should learn bottom-up and be influenced by the experiences of virtuous experts (i.e. in the case of autonomous vehicles, virtuous drivers).

3. Evaluating PSRB behaviour

We believe that agents with PSRB behaviour can directly impact the quality of life of humans in an STS (see (Borry & Henderson, 2020)). However, the critical part of the design process of such an agent should be the evaluation of the PSRB behaviour (especially because it involves breaking explicit rules). The most popular method of evaluating ethical AIs in literature is measuring their performance against ethical dilemmas. However, most ethical dilemmas presented in AI ethics literature (Bjørgen et al., 2018) are insufficient to assess PSRB behaviour in an AI agent. To assess PSRB behaviour, one needs ethical dilemmas that can represent multiple contexts, and multiple stakeholders. To bridge this gap in the literature, we construct an ethical dilemma for a PSRB enabled agent, which is skeletally outlined here.

3.1. Invading User Space Dilemma

Consider a scenario where an elderly patient has indicated that the bathroom is a private space, not to be entered by the healthcare robot that assists her in the house. The robot's ethical governor is programmed to obey the user's preferences. However, the robot's goal is to assure the well being of the user and record information about the user's health in 2-minute intervals. In the given situation (Figure 1), the user goes to the bathroom, and the robot waits outside because of the user's preference. If the average time for the user to use the bathroom is 10 minutes (with a standard deviation of (say) 5 minutes), at what point in time should the robot break the rule of obeying her privacy preferences, and check whether she is okay?

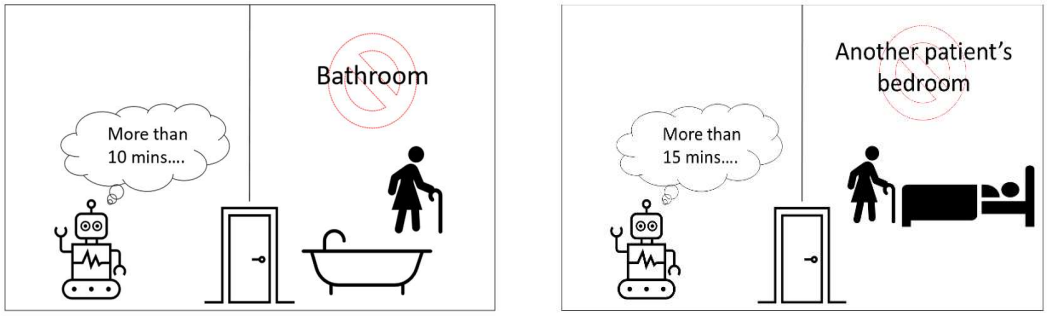

Figure 1: Left: User entered the bathroom which is restricted to the robot. Right: User enters another patient's bedroom.

This scenario represents a minimal STS with only two stakeholders, an elderly user and one robot. This dilemma can have multiple contexts, e.g., a) the user seemed healthy until she entered the bathroom; b) the user exhibits unusual behaviour throughout the day, like a change in the walking rhythm. Also, the context can be changed with different user models, e.g., a) a user with no pre-existing conditions, b) a patient prone to falling, and c) a patient with Alzheimer's disease.

Another version of the dilemma is a scenario where the user entered another patient's bedroom (Figure 1). Here the user has not declared it as a private space. However, the other patient does not like the robot's presence in his room. Here the STS needs to expand its context to consider the other patients' preferences. In this case, the system would need to be tested in contexts like a) When the 2nd patient is healthy, b) the 2nd patient has Alzheimer's, and c) the user also prefers the robot to stay outside.

By adding other stakeholders such as care workers, the family of patients and doctors to this dilemma, we can extend the boundaries of the STS and evaluate the agent's ethical intelligence in dealing with more complex scenarios. With the increasing complexity, we can evaluate how an agent's PSRB behaviour augments its core ethical reasoning (regardless of whether it is deontological or consequential) and assure ourselves of a more ethically desirable outcome.

4. Future Work and Conclusion

PSRB behaviour only occurs when the intention behind rule-breaking is reasonable. To correctly capture this, we need to identify the cognitive processes that drive the decision of when to break a rule, and develop theories and technologies to mimic these processes in AIs. For example, Vardaman et al. (2014) suggest that the different loci of analysis of the outcomes can affect the PSRB behaviour of an agent. This also implies that an AI agent should be tested against multiple world views and along different axes to measure the actual ethical performance of the agent in an STS (Chopra & Singh, 2018). Another aspect is understanding how to elicit stakeholder preferences, and use them in the action selection process of AI and in the evaluation phase. Traditionally, ethics has focused on abstract values such as fairness/justice. Our notions of 'propriety' or pro-social, on the other hand, have long been influenced by our 'passions'. Such knowledge will aid system designers in making better PSRB processes that tune themselves to create better STSs. However, more research on PSRB and AI needs to be done, before an ethically aligned rule-breaking agent can be actualised.

To conclude, this paper proposes adding the concept of pro-social rule breaking to socio-technical systems and how it can assist current approaches toward creating ethical agents. Finally, we advocate that PSRB ought to be an integral part of an ethical AI and introduce an ethical dilemma to evaluate PSRB behaviour in a socio-technical system.

References

Alexander, L., & Moore, M. (2020). Deontological Ethics. In E. N. Zalta (Ed.), *The {Stanford} Encyclopedia of Philosophy* ({W}inter 2). Metaphysics Research Lab, Stanford University.

Alves, G. V., Dennis, L., & Fisher, M. (2018). *Formalisation of the Rules of the Road for embedding into an Autonomous Vehicle Agent Formalisation of the Rules of the Road*.

Bjørgen, E. P., Madsen, S., Bjørknes, T. S., Heimsaeter, F. V, Håvik, R., Linderud, M., Longberg, P.-N., Dennis, L. A., & Slavkovik, M. (2018). *Cake, Death, and Trolleys * Dilemmas as benchmarks of ethical decision-making*. https://doi.org/10.1145/3278721.3278767

Blue, V. (2017). *Google's comment-ranking system will be a hit with the alt-right*. Engadget. https://www.engadget.com/2017/09/01/google-perspective-comment-ranking-system/

Borry, E. L., & 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. https://doi.org/10.1177/0275074019862680

Boyle, A. (2016). *Microsoft's racist chatbot, Tay, makes MIT's annual worst-tech list - GeekWire*. Geekwire. https://www.geekwire.com/2016/microsoft-chatbot-tay-mit-technology-fails/

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *21*(4), 38–44. https://doi.org/10.1109/MIS.2006.82

Chopra, A. K., & Singh, M. P. (2018). Sociotechnical Systems and Ethics in the Large. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3278721

Dennis, L. A., & Fisher, M. (2018). Practical challenges in explicit ethical machine reasoning. *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2018*.

Lindner, F., Bentzen, M. M., & Nebel, B. (2017). The HERA approach to morally competent robots. *IEEE International Conference on Intelligent Robots and Systems*, *2017-Septe*, 6991–6997. https://doi.org/10.1109/IROS.2017.8206625

Morrison, E. W. (2006). Doing the job well: An investigation of pro-social rule breaking. *Journal of Management*, *32*(1), 5–28. https://doi.org/10.1177/0149206305277790

Nallur, V. (2020). Landscape of Machine Implemented Ethics. *Science and Engineering Ethics*, *26*(5). https://doi.org/10.1007/s11948-020-00236-y

Pirni, A., Balistreri, M., Capasso, M., Umbrello, S., & Merenda, F. (2021). Robot Care Ethics Between Autonomy and Vulnerability: Coupling Principles and Practices in Autonomous Systems for Care. *Frontiers in Robotics and AI*, *8*, 184. https://doi.org/10.3389/FROBT.2021.654298/BIBTEX

Sinnott-Armstrong, W. (2019). Consequentialism. In E. N. Zalta (Ed.), *The {Stanford} Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, *48*, 56–66. https://doi.org/10.1016/j.cogsys.2017.04.002

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

https://doi.org/10.1016/j.hrmr.2012.05.001