Contextual Morality for Human-Centered Machine Learning

Niels van Berkel
The University of Melbourne
Melbourne, Australia
n.vanberkel@student.unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

Benjamin Tag
Keio University
Yokohama, Japan
tagbenja@kmd.keio.ac.jp

Simo Hosio
University of Oulu
Oulu, Finland
simo.hosio@oulu.fi

ABSTRACT

Big data and the increased use of Artificial Intelligence (AI) and Machine Learning (ML) have opened many new opportunities for continuous decision-support by autonomous systems. While initial work has begun to explore how human morality can inform the decision-making of future AI's [4], these approaches consider human morality as a static concept. We note that human morality and decision-making is affected not only by cultures and personalities but is to a large degree affected by an individual's context. In order to align with the moral judgements of their users, future ML applications should adjust their decision-making accordingly based on user context. In this work, we discuss our critical take on the importance of contextual morality for AI and identify opportunities for future work.

KEYWORDS

Context; Machine Learning; AI; Artificial Intelligence; smartphone sensing; decision-making; recommendation systems; ethics.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
2019.

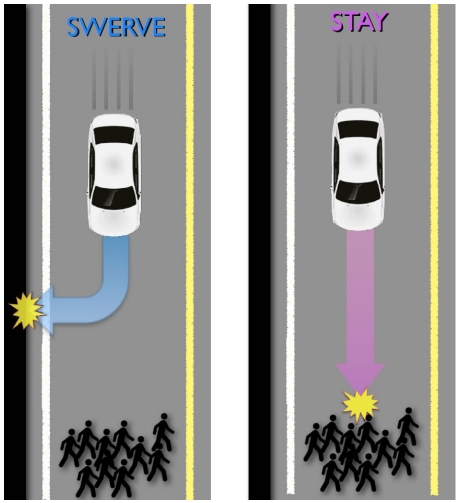

Figure 1: Two alternative scenarios for AI decision-making presented to participants, as included in [4].

INTRODUCTION

Morality describes the codes of conduct put forward by an individual, group, or society that distinguish right and wrong behaviour and decision-making [6]. Advances in Artificial Intelligence (AI), and Machine Learning (ML) in particular, have resulted in a world in which autonomous systems increasingly support or lead in everyday decision-making tasks. While these technological advances promise an increase in efficiency, decisions made by AI may not necessarily align with the moral judgement of the user or target of the AI. Recent advances have been made to ensure AI-powered decision-making aligns more closely with the core principles of human morality [4].

Work in Philosophy has pointed to the effect of context on our moral compass. Rather than a constant set of rules or preferences, our morality is affected by contextual changes such as time of day, language, and social presence [5, 11, 13]. It is therefore insufficient for user-facing algorithms, as found in *e.g.* mobile applications and content recommendation systems, to implement algorithms which follow culture-specific or general moral guidelines. Instead, in order to develop AI that seamlessly supports the user, Human-Computer Interaction (HCI) researchers and practitioners have to create a better understanding of the contextual implications on a user's morality. In addition to more morally intelligent agents, other potential applications include content suggestions (*e.g.*, news, entertainment) or pro-active interaction cues (*e.g.*, message replies) as based on the user's current context. This work explores this concept in more detail and presents opportunities for future work in this emerging domain.

Finally, a recent article in *Science* shows the frequent occurrence of moral events in daily life [8], advocating the study of morality 'in the wild' as opposed to artificial lab studies. Given the nature of our field, the HCI community can contribute to the further understanding of morality.

RELATED WORK

The study of morality has a rich history, with an extensive array of viewpoints developed since the ancient philosophers. Work by *inter alia* Aristotle focused strongly on 'virtue ethics', *i.e.* characteristics which make for a 'good' person [1]. Some of the virtues advocated by Aristotle include 'courage', 'truthfulness', and 'modesty'. The school of virtue ethics has been criticised for focusing on personal characteristics rather than the (moral) value of an individual's actions. This has hampered the ability to utilise virtue ethics as a base of *e.g.* legislation. We further note that the application of virtues in computer applications is inherently challenging; *i.e.*, how to create or define a courageous algorithm?

Applying a Utilitarian perspective has thus far been more effective. Utilitarianism states that ethical decision making maximises value (*i.e.*, utility) [14]. As such, an algorithm can be constructed that – as long as the utility of potential outcomes is defined – can calculate the morally most optimal decision. Bonnefon et al. [4] utilise this perspective in studying the perception on the moral implications of

autonomous vehicles (AVs). AV algorithms will eventually have to make decisions which negatively affect either the passenger(s) or other road users (*i.e.*, avoid collision with pedestrians by steering off the road). Participants were asked to make a moral judgement on this decision (Figure 1). Participants generally approved of maximising the utilitarian value of human life. Results from these evaluations can offer detailed insights for the implementation of both algorithms and policy.

However, this example fails to consider the significant differences in morality between cultures. For example, in a study comparing the responses of British and Chinese participants concerning the trolley problem, results indicated that Chinese participants were less likely to sacrifice one person to save five others [7]. As such, we argue that AI applications should consider the context in which it operates to function successfully. A commonly used framework to distinguish differences between cultures is Hofstede's cultural dimensions theory (see sidebar) [9].

Recent advances in Philosophy reveal that a person's moral judgement is variable. For example, Leavitt et al. [12] revealed that ones assumed occupational identity (*e.g.*, manager vs engineer, soldier vs medic) substantially influence moral judgement. Similarly, Reynolds et al. [16] show how contextual cues can shape moral behaviour in a business setting. These examples show that both the long-term and short-term context affect the moral judgement of individuals.

Hofstede's 6D Model. In this model, a total of six dimensions are used to categorise cultures. These dimensions represent independent preferences that explain differences between the moral norms of respective countries [9].

Dimensions:

- Power distance index (PDI)
- Individualism vs. collectivism
- Uncertainty avoidance (UAI)
- Masculinity vs. femininity
- Long-term orientation vs. short-term orientation
- Indulgence vs. restraint

Examples: The scores reported on each dimension are relative, and – as all inhabitants of a country are unique – are only meaningful in comparison to the scores of other countries. For example, the United States has a high level of Individualism (91) as compared to Portugal (27), which features a more collectivist society. Japan has a high level of Masculinity (95) in which competition between groups is encouraged. Compared to Japan, The Netherlands features a low level of Masculinity (14), with values such as compromise and solidarity valued more prominently.

CONTEXT AND HUMAN-CENTERED MACHINE LEARNING

Although the interplay of context and interaction has long been studied in HCI, only recent work has begun to quantify the effects of (digital) context on moral perception and subsequent decision-making. For example, Barque-Duran et al. [2] found that participants are more utilitarian when using smartphones instead of desktop computers in high conflict moral dilemmas. Hosio et al. [10] find substantial differences in participant's perceived willingness to help another participant out, by donating non-monetary resources under different contexts. Using 'in the wild' deployments, researchers can collect both participant's moral judgements (*e.g.*, mobile questionnaires) and contextual data [3]. This allows for the repeated collection of participant responses across time and context and is a critical first step towards building an understanding of the moral implications of context (see Figure 2).

Following the quantification of context on expectations towards morality, application designers can embed this knowledge in their applications. We envision future application designers establish a predefined bandwidth in which the morally-conscious agent operates as informed by combined invariable (*e.g.*, country, culture) and variable (*e.g.*, social condition, smartphone usage) data streams. A key element of this work is the reduction of biases which are present in *e.g.* publicly available 'Social Data' [15].

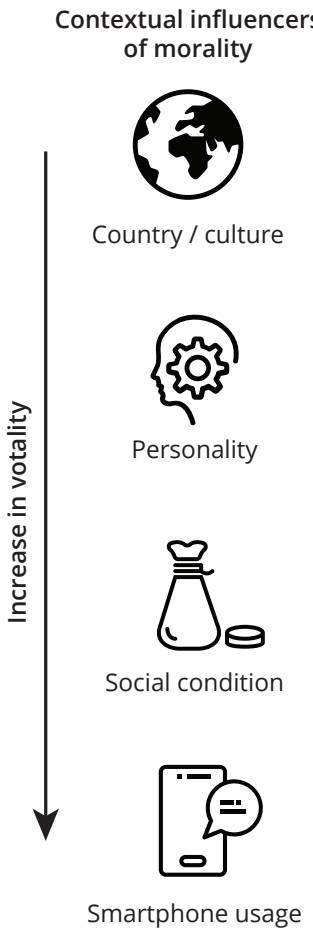

Figure 2: Examples of entities which affect our morality, ordered by their votality.

CONCLUSION

A challenge for future work in HCI and ML is to account for differences in the end user's contextual morality. To meet this challenge, we identify two goals. First, researchers must obtain an understanding of the effect of different contexts on end-user morality. This understanding is currently lacking but can be obtained through combined human- and sensor-data collection. Second, future ML applications should be able to sense user context, identify the corresponding moral implications within predefined guidelines, and apply the required changes to the application. Our work identifies future research for the Human-Centered Machine Learning community.

REFERENCES

[1] Aristotle. 2000. *Aristotle: Nicomachean Ethics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802058
[2] Albert Barque-Duran, Emmanuel M Pothos, James A Hampton, and James M Yearsley. 2017. Contemporary morality: Moral judgments in digital contexts. *Computers in Human Behavior* 75 (2017), 184–193.
[3] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6, Article 93 (2017), 40 pages. https://doi.org/10.1145/3123988
[4] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576. https://doi.org/10.1126/science.aaf2654
[5] Janet Geipel, Constantinos Hadjichristidis, and Luca Surian. 2015. How foreign language shapes moral judgment. *Journal of Experimental Social Psychology* 59 (2015), 8 – 17. https://doi.org/10.1016/j.jesp.2015.02.001
[6] Bernard Gert and Joshua Gert. 2017. The Definition of Morality. In *The Stanford Encyclopedia of Philosophy* (fall 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[7] Natalie Gold, Andrew Colman, and Briony Pulford. 2015. Cultural Differences in Responses to Real-Life and Hypothetical Trolley Problems. *Judgment and Decision Making* 9, 1 (2015), 65–76.
[8] Wilhelm Hofmann, Daniel C. Wisneski, Mark J. Brandt, and Linda J. Skitka. 2014. Morality in everyday life. *Science* 345, 6202 (2014), 1340–1343. https://doi.org/10.1126/science.1251560
[9] Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.
[10] Simo Hosio, Denzil Ferreira, Jorge Goncalves, Niels van Berkel, Chu Luo, Muzamil Ahmed, Huber Flores, and Vassilis Kostakos. 2016. Monetary Assessment of Battery Life on Smartphones. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1869–1880. https://doi.org/10.1145/2858036.2858285
[11] Maryam Kouchaki and Isaac H. Smith. 2014. The Morning Morality Effect: The Influence of Time of Day on Unethical Behavior. *Psychological Science* 25, 1 (2014), 95–102. https://doi.org/10.1177/0956797613498099
[12] Keith Leavitt, Scott J Reynolds, Christopher M Barnes, Pauline Schilpzand, and Sean T Hannah. 2012. Different Hats, Different Obligations: Plural Occupational Identities and Situated Moral Judgments. *Academy of Management Journal* 55, 6 (2012), 1316–1333.
[13] Stanley Milgram. 1963. Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology* 67, 4 (1963), 371–378.
[14] John Stuart Mill. 1863. *Utilitarianism*. London: Parker, Son and Bourn.
[15] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016). https://doi.org/10.2139/ssrn.2886526
[16] S. J. Reynolds, K. Leavitt, and K. A. DeCelles. 2010. Automatic ethics: the effects of implicit assumptions and contextual cues on moral behavior. *Journal of Applied Psychology* 95, 4 (Jul 2010), 752–760. https://doi.org/10.1037/a0019411