Planning for Ethical Agent-Agent Interaction

Jesse David Dinneen
Victoria University of Wellington, New Zealand
jesse.dinneen@vuw.ac.nz

ABSTRACT
In this position paper for the 2019 CSCW workshop *Good Systems: Ethical AI for CSCW* I propose one tool and one idea for navigating the complex ethical problem space that results from the interaction of human and/or AI agents in shared, hopefully cooperative, computing environments.

KEYWORDS
AI ethics; human-computer interaction; ontic trust

AGENT-AGENT INTERACTION
The introduction of non-human agents (e.g., AI-powered virtual assistants like Apple's Siri), which are increasingly indistinguishable from human agents, to everyday computing raises many questions about what AI, our interactions with it, and AI-AI interactions (all collectively: *agent-agent interactions*) could and should be like, especially as AI grow in capacity to be moral agents akin to humans (Floridi & Sanders, 2004). To guide implementations and expectations of moral AI requires considering the many ways agents' actions can be undesirable, but presently most news media and even scholarly discourses focus narrowly on poor performance, transgressions of the law, and negative outcomes for particular individuals irrespective of their socio-economic group or status (Stahl *et al.*, 2016). Thus, while the moral-problem space of agent-agent interaction is arguably more vast than that of human-human interaction, we designers and critics have so far used fewer tools to analyse it.

Towards addressing narrow thinking in AI ethics, and to aid anticipatory (rather than reactionary) policy, I would like to consider the use of a *multi-moral matrix* for assessing (in design or *post hoc*) particular AI cases along multiple moral frameworks. Figure 1 shows an example matrix with agent actions in the leftmost column and each other column showing possible transgressions according to various moral frameworks (summarised to the point of caricature), including (left to right): legalism, consequentialism, virtue ethics and deontology, social justice, and social contractualism. Cells at the intersections of actions and frameworks may reflect only that there are possible transgressions or may contain more detail. Agent actions may be relatively easy to populate, e.g., by identifying actions in user stories during development.

Action	**Illegal**	**Bad outcome**	**Ill-intentioned, deceptive, neglectfdul**	**Socially unjust**	**Violates a social contract**	**...**
1			X		X	
2				X		
...						

Figure 1. Partial example of a multi-moral matrix for analysing ethical issues in agent-agent interaction.

The example matrix includes only a few moral frameworks and is meant to be neither exhaustive nor prescriptive; any use of such a matrix requires customisation according to the expectations of, e.g., the relevant sectors and cultures.

Finally, I suggest to adopt or adapt into thinking about AI ethics an idea that appears thematically appropriate and also promising for addressing cultural moral differences like those just mentioned (Hongladarom, 2008): the concept of *ontic trust* (Floridi, 2009). Put briefly, ontic trust is a responsibility to care for the intrinsically valuable information objects that populate our world/*infosphere;* shorter still, causing entropy in a shared information environment is unethical. Such an idea shows us a less obvious way agent-agent interactions can be bad (i.e., by causing entropy – it could thus go in the above matrix), but arguably also implies rights for non-human agents (i.e., AI). Surprisingly, little has been said about ontic trust in AI ethics discourse (and to my knowledge, nothing has been said in the context of CSCW). It may therefore be useful to raise the idea at the workshop and discuss questions like:

- Is an imperative to care for information objects equivalent to an imperative to *prevent* entropy?
- What do such imperatives imply entail for privacy, data logging, the right to be forgotten, and CSCW community values? (Bruckman *et al.*, 2017)
- Can ontic trust aid universal design by, for example, mediating disparate cultural views about ethical AI?

CSCW 2019 | Austin TX, USA | 10 November, 2019

REFERENCES

Bruckman, A. S., Fiesler, C., Hancock, J., & Munteanu, C. (2017). CSCW research ethics town hall: Working towards community norms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 113-115). ACM.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349-379.

Floridi, L. (2009). Global information ethics: The importance of being environmentally earnest. In *Human Computer Interaction: Concepts, Methodologies, Tools, and Applications* (pp. 2450-2461). IGI Global.

Hongladarom, S. (2008). Floridi and Spinoza on global information ethics. *Ethics and Information Technology, 10*(2-3), 175-187.

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The ethics of computing: A survey of the computing-oriented literature. *ACM Computing Surveys (CSUR), 48*(4), 55.