ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.p02

TOWARD HUMAN-LEVEL MORAL COGNITION IN A COMPUTATIONAL COGNITIVE ARCHITECTURE

PAUL BELLO

Head, Interactive Systems Section,
U.S. Naval Research Laboratory, USA

The ostensible target for much of the work in machine ethics is to develop action-selection routines for intelligent agents that flexibly incorporate norms as soft constraints so as to guide their behavior. For now, let us call this the ``Context of Deliberation" (CD). CD can be contrasted with the ``Context of Judgment" (CJ), where an agent is deciding if and how blame should be apportioned in a situation S which elicits norms $\boldsymbol{N}_S$, given interactions between agents $\boldsymbol{A}_1 \ldots \boldsymbol{N}_J$. Building a system capable of judgment in CJ is just as important as building a system that can flexibly decide and act in CD. More clearly, part of flexibly choosing how to act with respect to norms may involve how such actions will be evaluated by others in CJ. Because my lab is interested primarily in human-machine interaction, our efforts will consist in getting a system to reason about how a human observer might apportion blame in various scenarios. Such judgments can then be put to use in choosing if, when, and how to act. Such a seemingly simple thing that most of us do every day involves the coordination of a dizzying array of capacities that range from perception up through higher-order cognition. The critical conclusion to draw is that machine ethics is not just a matter of formalism, or even of normative ethics, but demands an approach grounded in cognitive architecture. In this talk, I present first steps at building a cognitive architecture capable of simultaneously operating in CD and CJ, using judgments generated in the latter to inform action-selection in the former: all while engaging in ongoing moment-by-moment perception and action.

