A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings

Collective Intelligence
Volume 2:2: 1–14

DOI: 10.1177/26339137231162025
journals.sagepub.com/home/col
Sage

Eugene Vinitsky
UC Berkeley, Berkeley, CA, USA

Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets and Joel Z Leibo
Deepmind, London, UK

Abstract
Society is characterized by the presence of a variety of social norms: collective patterns of sanctioning that can prevent miscoordination and free-riding. Inspired by this, we aim to construct learning dynamics where potentially beneficial social norms can emerge. Since social norms are underpinned by sanctioning, we introduce a training regime where agents can access all sanctioning events but learning is otherwise decentralized. This setting is technologically interesting because sanctioning events may be the only available public signal in decentralized multi-agent systems where reward or policy-sharing is infeasible or undesirable. To achieve collective action in this setting, we construct an agent architecture containing a classifier module that categorizes observed behaviors as approved or disapproved, and a motivation to punish in accord with the group. We show that social norms emerge in multi-agent systems containing this agent and investigate the conditions under which this helps them achieve socially beneficial outcomes.

Keywords
Multi-agent systems, social norms, reinforcement learning