Social Consequence Engines (Abstract)

Julian Padget, Marina De Vos, Charlie Ann Page
Department of Computer Science, University of Bath, United Kingdom
{j.a.padget,m.d.vos,c.a.page}@bath.ac.uk

An autonomous agent can broadly be characterised as an entity that decides what to do using: (i) a model of the world in which acts built from the percepts it receives, (ii) its currently active goals, and (iii) the actions available to it. We use the term agent to refer to either a software agent or to an embodied agent, while collectives of such agents might control an intangible, complex business process, a tangible entity like an autonomous vehicle, or a mixed environment such as a warehouse picking and stocking system.

We sketch an outline solution to the question of how to add normative reasoning to the "consequence engine" (CE) action selection process (Blum et al. 2018). The CE evaluates the consequences of actions over a short horizon in see how action choices might fit with potential actions of other agents, against a model built from the physical world percepts received so far. Our aim is to augment that model with normative percepts (permission, prohibition and obligation) to construct a *social* consequence engine, building on the norm representation and reasoning work of Padget et al. (2016), King et al. (2017), Shams et al. (2017) and Padget et al. (2018). In this way, action may be predicated not only on physical observations, but also on the social interpretation(s) of those observations, which may in turn lead to more appropriate choices that better meet human expectations.

What should an agent do next? Action or plan selection – depending on the agent architecture – in the majority of interesting problem domains, is inevitably predicated on incomplete, uncertain information due to the twin problems of transduction, and representation and reasoning in the context of a continuous, dynamic, non-episodic and non-deterministic environment. Equally inevitably, a chosen action/plan will often then be the wrong thing to do. This leads to the notion of committing an agent (Cohen et al. 1990) to the ends – the state of affairs to achieve – and the means – the actions to bring ends about, where flexibility comes from replanning, triggered by changes in circumstances.

Typically, plan selection, and thereby the actions constituting the plan, is a function of the agent's goals, some representation of "now" and the actions/plans available to the agent. This process does not take account of: (i) consequences of an action or sequence of actions, (ii) norms – what "ought" an agent to do, as against what maximizes utility through goal satisfaction – that govern an agent and those with which it interacts and (iii) normative interpretation(s) of an agent's taken or intended actions. We summarise how each of these is addressed in selected literature, then how to build on these ideas to deal with the problem identified at the outset:

1. Blum et al. (2018) propose the consequence engine as an internal evaluative mechanism to help a robot select an action that keeps the danger ratio (sic) low. This engine runs a model built from the robot's percepts, at a frequency higher than its action capability, so that it can make short-term predictions about the state of its perceived world (i.e. possible worlds) including the actions of others, and select actions that ought to bring about world states that maximize safety. Their experiments report how one robot can prevent another from taking a path that puts it at risk, because the first considers the possible worlds in a short time horizon, sees the "bad" outcome and makes an intervention to change the path of the second. An experiment with three robots shows how repeated consequence evaluation leads to vacillation, as an action to save the second is overridden by an action to save the third and vice versa, with the outcome that neither is saved approximately half the time.
2. Shams et al. (2017) show how to consider norms in action selection, while pursuing goals whose achievement may conflict with those norms. The planning mechanism handles multiple goals and norms in the presence of durative actions that proceed concurrently. Plans are ranked by the utility gains from goals and losses from norm violations, leading to a set of optimal plans that maximise overall utility, any of which can be chosen by the agent to follow. What is relevant here is that: (i) planning accounts both for brute (or domain) and normative (or institutional) facts (Searle 1995), and (ii) actions are not necessarily atomic, but may last for finite periods of time and may be concurrent (although an agent may only initiate one action at once).

3. Padget et al. (2018) describe a resource-oriented architecture (ROA) pattern for the creation of deontic sensors that observe agents' actions and deliver normative interpretations of them. The pattern is instantiated with the InstAL (Institutional Action Language) (Padget et al. 2016), so that an agent may know what it is permitted, prohibited or obliged to do, or request "what-if" normative interpretation(s) of a sequence of actions. The benefit of the deontic sensor pattern is that it makes normative reasoning available as a RESTful service, decoupled from the agent platform itself.

How might an agent make a better decision? We propose drawing together the ideas outlined above to enable the construction of a social consequence engine that combines the consideration of brute and normative facts in its possible worlds construction and evaluation process:

1. The consequence engine (Blum et al. 2018) demonstrates how robots can take account of the consequences of their actions (or inaction). By exposing the action selection process to deontic as well as physical sensors, decision-making can be informed by what the agent is permitted, prohibited or obliged (this last is typically a consequence of an earlier action) to do, as well as what is directly observable. Considering the three robot scenario, once an action/plan is selected the controller might then generate a prohibition to undo the chosen action/plan, thereby committing the robot to a choice unless a more beneficial course of action can be identified, at which point the prohibition can be violated.

2. The normative planning process (Shams et al. 2017) finds plans that fulfill the identified goals, however long that plan may take and however long that planning process may take, whereas Blum et al. (2018) demands answers in short, finite time frames. But once these plans are obtained, the output from the deontic sensor can be used to select which plans are still viable and preferred. These checks involve short time frames so these updates will take little time, making them suitable for reasoning within the CE. When no more plans are available, re-planning is needed. Equally pro-active re-planning can take place in the background through a service.

3. The deontic sensor (Padget et al. 2018) is conceived to operate in a loosely-coupled, distributed environment, where network communication between client and service is the norm. In contrast, tight-coupling with minimal overheads are essential for both the consequence and the social consequence engine. In practice, the abstraction can be maintained while delivering the services locally, which should mitigate the situation, while keeping an explicit, external representation of norms allows for independent verification and validation, even the potential for certification, as well as straightforward modification or replacement as governance requirements or jurisdiction change.

Our aim with the above is to suggest how the notion of the physical consequence engine might effectively be complemented by a *social* consequence engine, that provides a normative view of the possible worlds, and hence the capacity to base decision-making on the combination of brute and normative outcomes of action selection. In this way, a robot may receive advice that is influenced by the social and regulatory context in which it operates, as well as the physical aspects, and thus be able better to meet human and legal requirements for the behaviour of "intelligent" software-controlled artefacts.

References

Blum, Christian, Alan F. T. Winfield, and Verena V. Hafner (2018). "Simulation-Based Internal Models for Safer Robots". In: *Frontiers in Robotics and AI* 4, p. 74. DOI: 10.3389/frobt.2017.00074.

Cohen, Philip R. and Hector J. Levesque (1990). "Intention is Choice with Commitment". In: *Journal of Artificial Intelligence* 42, pp. 213–261.

King, Thomas C., Marina De Vos, Virginia Dignum, Catholijn M. Jonker, Tingting Li, Julian Padget, and M. Birna van Riemsdijk (2017). "Automated multi-level governance compliance checking". In: *Autonomous Agents and Multi-Agent Systems*, pp. 1–61. DOI: 10.1007/s10458-017-9363-y.

Padget, Julian, Marina De Vos, and Charlie Ann Page (2018). "Deontic Sensors". In: *Proceedings of IJCAI 2018*. Ed. by Jérôme Lang. In press. Preprint available from: https://researchportal.bath.ac.uk/en/persons/julian-padget/publications/.

Padget, Julian, Emad ElDeen Elakehal, Tingting Li, and Marina De Vos (2016). "InstAL: An Institutional Action Language". In: *Social Coordination Frameworks for Social Technical Systems*. Springer, pp. 101–124. DOI: 10.1007/978-3-319-33570-4_6.

Searle, John (1995). *The Construction of Social Reality*. Allen Lane, The Penguin Press.

Shams, Zohreh, Marina De Vos, Julian Padget, and Wamberto W. Vasconcelos (2017). "Practical reasoning with norms for autonomous software agents". In: *Engineering Applications of Artificial Intelligence* 65, pp. 388–399. DOI: https://doi.org/10.1016/j.engappai.2017.07.021.