

Building blocks for artificial moral agents

Vincent Wiegel¹

¹ Faculty of Policy, Technology and Management, Delft University of Technology
v.wiegel@tbm.tudelft.nl

Abstract

Constructing artificial moral agents serves at least two purposes: one, better understanding of moral reasoning, and, two, increasing our trust and confidence in creating autonomous agents acting on our behalf. When constructing agents we must identify what kind of agent we are trying to implement acknowledging still severe restrictions in both our understanding of moral reasoning and technological capabilities. In constructing agents, as part of the SophoLab project, we adopt a set of requirements and design principles. As building blocks the belief-desire-intention model is used along with the deontic-epistemic-action logic framework as modeling and specification languages. Implementation in executable software is done using the JACK agent language. Attempts at constructing agents and actual experiments provide some first encouraging results but also point at the large hurdles still to be taken.

Drivers for artificial moral agents research

There are at least two perspectives on the combination of artificial agents and ethics that make it a very interesting combination. One, attempting to create artificial agents with moral reasoning capabilities challenges our understanding of morality and moral reasoning to its utmost. It requires attention to the general structure and architecture, as well as to the little details. As the computer is a relentless component that will not function until completely instructed it pushes us moral philosophers to explicate everything we know. It also challenges our formal apparatus to express our knowledge such that is formal enough for software engineers to be understood. Two, with the advancement of technological development artifacts play increasingly important roles in our lives. They do not only contain information about us, they start to act on our behalves. With the increasing autonomy comes an increased need to ensure that their behavior is in line with what we expect from them. We want to ensure that no abuse is made when opportunity offers, no harm is done to others, etc. Besides this negative, constraining aspect there is also a positive aspect. If we want to develop machines that can act to help us, for example in circumstances where we cannot come because it is too dangerous, they will need all the reasoning power we can give them including moral reasoning.

When attempting to construct artificial moral agents we need to define, at least roughly, what it is that we are aiming at. Moor proposes a classification of four types of ethical agents: ethical impact agents, implicit ethical agents, explicit ethical agents and full ethical agents (Moor 2006).

Ranging from agents that have an impact by their very existence to full blown, human-like reasoning agents that have consciousness, intentionality and an ability to provide plausible justifications for their actions.

Given our current understanding of moral reasoning, artificial intelligence, epistemological devices, etc. the best we can try to construct are explicit ethical agents that can make some ethical judgments that are not hard-wired into their make-up and have some ability to provide an account of how they arrived at their judgment.

In terms of sophistication we are now looking to implement second generation moral agents. The first generation artificial moral agents were ground breaking in that they for the first time implemented moral decision making (a paradigmatic example in case Danielson 1992) . They are limited in their orientation (moral issues only), their internal make-up, and interoperability and interconnectedness. The current generation will have support for forms of interconnectedness (team work for example), be multi-purpose (mixture of moral and non-moral goals) and still be very limited in embodiment (mainly electronic) and epistemology (only strong typed). A next generation will probably bring physically embodied agents with strong moral epistemological capabilities.

Design principles and requirements

In implementing and experimenting with artificial constructs we follow a methodology that focusses on the translation from one environment - theory and application - to another - the artificial setting (Wiegel 2006b). This translation happens in several steps. Key is that the theory and application examples are modeled in a (semi)formal language that is close to the implementation environment. The requirements that follow from this methodology are amongst others:

- neutrality - support for various (meta-)ethical theories
- executability
- comprehensibility
- formality
- configurability
- scalability
- extendibility

(see also Dumas et. al., 2002, who come up with a similar set of requirements for artificial agents geared for economic transactions, and with whom I share kindred approach).

The environments in which artificial moral agents will operate will not be fully predictable and controlled. And if we want them to be successful they will have to have the

ability to determine their actions autonomously, that is their behavior and action repertoire are not predictable (note that determinism does not imply predictability!). And in time new abilities will be added onto the existing set (dynamically). The agents will be active in different application domains and engage in both non-moral and moral activities. They will pursue different goals at the same time, and have to make decisions within limited time and restricted information. The list of requirements following from these observations are:

- mixed moral and on-moral goals support
- mixed moral and on-moral activity support
- bounded rationality, time & resource constraint
- extendibility

This is not an exhaustive list of all requirements, but it does capture several key ones.

Following from these requirements a set of design principles can be formulated (from which I omit the software engineering oriented ones):

- agents behave both pro-actively, goal driven and reactively
- behavior is build from small action components that can be put to various uses and re-configured at run-time to form different behavior patterns
- agents can decide if and when to update their information base
- agents interact with each other and the environment

These principles are in support of what Korniek and Uzgalis observed as important characteristics of successful biological systems, which I feel are directly relevant in this context (Korniek and Uzgalis 2002:85).

- emergent behavior – not all behavior is specified upfront
- redundant degrees of freedom – more ways to achieve a particular goal
- no central director of action
- think local, act local – local scope of control

These are characteristics agents should have, and/or display in order to be able to meet to above requirements.

Building blocks: modeling & implementation

To model agent behavior the belief-desire-intention model, BDI-model (Bratman, 1987) provides a good foundation. It captures both bounded rationality, and the goal oriented aspect that is required for autonomous agents. There are two important elements missing in the BDI-model to make it suitable for modeling artificial moral agents: the deontic element and the action element. Therefore the BDI-model is extended through the deontic-epistemic-action logic framework, DEAL framework (Van den Hoven and Lokhorst, 2002). The deontic logic covers the deontic concepts of 'obligation', 'permission', and 'forbidden'. Epistemic logic expresses the things we know and belief. And the action logic allows us to reason, through the STIT – see to it that – operator to reason about actions. Combined we can construct propositions like

- $\text{Bi}(G(\Phi)) \rightarrow \text{O}([i \text{ STIT } \Phi])$

meaning if *i* believes that ' Φ ' is morally good than *i* should act in such as way that ' Φ ' is brought about

The modal logic in this approach is used as a specification language rather than as formal logic for theorem proving. My sentiment in this respect is very much that of (Halpern, 2000:1)

“McCarthy want robots to reason with a formal logic. I'm not sure that is either necessary or desirable. Humans certainly don't do much reasoning in a formal logic to help them in deciding how to act; it's not clear to me that robots need to either. Robots should use whatever formalisms help them deciding. Robots should use whatever formalisms help them make decisions. I do think logic is a useful tool for clarifying subtleties and for systems designers to reason about systems (e.g., for robot designers to reason about robots), but it's not clear to me that it's as useful for the robots themselves.”

Of course, software agents and robots do need logic. It is inconceivable that they reason without logic. But they need not necessarily have a complete and sound logic system.

Using the BDI and DEAL modeling tools agents can be attributed both moral and non-moral desires, have beliefs about what is morally obligatory or permissible, form multiple intentions, decide to act on them, and actually enact them.

The implementation of these models is done using the JACK agent language (JACK). This abstraction layer on top of the Java programming language provides support for multi-agent systems, based on the BDI-model. It provides components for agents, team, modalities of belief, desire and intention. Plans are sequences of actions that provide low level components for behavior. Plans are connected at run-time to provide complex behavior. To facilitate meta-level reasoning there are various mechanisms ranging from hard-wired, instantaneous choice to explicit reasoning about options and preferences. Beliefs are implemented as n-tuples (first-order relational models) that support both open-world and closed-world semantics. Plans can be reconsidered based on new desires (events), and new information becoming available. Plans can be linked to beliefs using logic variables for which agents can find multiple bindings. An intricate mechanism that allows to cater for multiple instances of an objective.

Some results

These building blocks proved to be a powerful basis to capture, model and implement aspects of moral agency, though still a great deal needs to be done (Wiegel 2006a). Below I list some examples of the (support for) insights gained.

1) Negative moral commands are different in nature from the way we talk about them. 'Thou shall not...' is a general form of moral command telling you what not to *do*. Trying to implement such obligations proved rather hard. The reason is that these commands are in fact not about acts as we talk about about them: do not kill, do not lie, etc. There are many ways (in fact infinite) in which one can kill, and identifying each of them is impossible. These moral commands are actually about classes of action that are characterized by their outcome, e.g. bringing about a state of affairs in which someone does not have a beating hart, with all the qualifiers about intentionality, etc. This observation implies that agents must have an explicit conception (right or wrong) about the outcomes of their actions, and the ability to classify them accordingly.

2) Morality must act as both restraint and goal-director. Moral reasoning in artificial agents much function within a large context. An artificial *moral* agent in its own right does not have much practical relevance when it comes to application. An agent can have as one of its goals or desires to be a moral agent, but never as its only or primary goal. So the implementation of moral reasoning capability must always be in the context of some application in which it acts as a constraint on the other goals and action.

3) Moral decision making or decision making with moral constraints must take various restricting factors into account. One, an agent is open to permanent information feeds. It must be able to decide on ignoring that information or taking it into account. Two, once goals have been set, these goals must have a certain stickiness. Permanent goal revision would have a paralyzing effect on an agent and possibly prevent decision making. Three, different situations require different decision-making mechanisms. In some situations elaborate fact finding and deliberation are possible, in other the decision must be instantaneous.

All these three restrictions refer to resource boundedness and the time consuming dimension of decision making. To cater for these the BDI-model provides a good basis. The software implementation offers three decision making, or meta-level reasoning mechanisms: hardwired, ordinal and cardinal ranking and explicit reasoning.

4) Moral epistemology will prove to be the biggest challenge for the implementation of artificial moral agents. (Meta-)ethical theories make strong epistemological claims. E.g. if moral qualities are supervenient on natural qualities how does an agent perceive these natural qualities and deduce the moral dimension from them? If moral qualities are non-reducible how does an agent intuit or perceive them? How does an agent come to approve of something, and hence call it morally good?

These are very hard questions for which no clear answers are available that can be formalized. Moral epistemology of artificial agents will have to be strong typed for the near future. This means that perception, events, facts, etc. have

to be typed at design-time. This means for example that events will need to have an attribute identifying its type, which then allows the agent to interpret it.

5) Walk-of-life scenario testing is a good way to test the agents. To better understand the above findings and to test artificial moral agents a Walk-of-life scenario provides a good tool. It details 'a day in the life of' for an agent in which it pursues various goals and is confronted with different events that require various form of deliberation.

With all the promising results a note of reserve is called for. With Wooldridge, writing on his logic for rational agents, I would like to say (Wooldridge, 2000:91)

“Belief, desire and intention are in reality far too subtle, intricate and fuzzy to be captured completely in a logic [...] if such a theory was our goal, then the formalism would fail to satisfy it. However, the logic is emphatically not intended to serve as such a theory. Indeed, it seems that any theory which did fully capture all nuances of belief, desire and intention in humans would be of curiosity value only: it would in all likelihood be too complex and involved to be of much use for anything, let alone for building artificial agents.”

This goes for logic, and for any other means to capture the aspects of moral reasoning. With our current understanding of moral reasoning, the challenges of moral epistemology and the state of technology development (far and impressive though it is, still falling short by a long way) we will have to settle for limited artificial moral agents. The explicit ethical agent is an ambitious goal though within the realm of the possible in our personal foreseeable future.

References

Agent Oriented Software Pty. Ltd, JACK, www.agent-software.com.au

Bratman, M.E., 1987. *Intention, Plans and Practical Reasoning*, Harvard University Press, Cambridge

Danielson, P., 1992. *Artificial Morality*, Routledge, London

Governatori, 2002. A Formal Approach to Negotiating Agents Development, in *Electronic Commerce Research and Applications*, 1 no. 2

Halpern, J., 2000. On the adequacy of model logic, II, *Eletronic News Journal on reasoning about action and change*

Hoven, M.J. van den, Lokhorst, G.J., 2002. Deontic Logic and Computer Supported Computer Ethics in Cyberphilosophy, Bynum et. al. (eds)

Moor, J.H., 2006. The nature, importance, and difficulty of machine ethics (IEEE forthcoming)

Wiegel, V. et. al., 2006a. Privacy, deontic epistemic action logic and software agents, in Ethics and information technology forthcoming

Wiegel, V., 2006b. SophoLab (forthcoming), in Ethics and Information Technology

Wooldridge, M., 2000. Reasoning about Rational Agents, MIT Press, Cambridge