Towards a Theory of Mind for Ethical Software Agents

Catriona Kennedy[1]

Abstract. We consider the design of an artificial agent that can determine whether a human action is acceptable according to ethical norms and values that typical humans would use in the same situation. Such a decision often depends on whether an action was intended, or on what the actor knows. Therefore the decision-maker needs to reason about mental states of others, a capability known as "Theory of Mind" (ToM). To understand moral scenarios, humans have a rich understanding of mental concepts as a result of experience. In this paper, we argue that many of these concepts can be defined in terms of information processing and mental states in a generic sense, and can be implemented computationally. For example, affective states may be defined in terms of goals, resources and degree of control. We argue that an agent can acquire some understanding of mental concepts and moral norms by developing models of its own information processing on different levels of abstraction and using these models to simulate other minds.

1 Introduction

In complex and fast-changing environments, autonomous agents may have to determine whether humans or other agents are acting ethically. When humans make such decisions, the outcome often depends on whether damage is intended, or on what the actor knows. Therefore the person needs to reason about mental states of others, a capability known as "Theory of Mind" (ToM) [15]. Similarly for an artificial agent to make moral evaluations about issues that humans are concerned about, it needs a non-trivial understanding of mental states and their relation to the scenarios under analysis.

To consider the design of such an ethical agent, we will use the following example scenario. An agent A is given a report about the actions of a human B. While viewing an apartment, B knocked over a vase, damaging something that was valuable to the owner. No further information is given on the context or on the subsequent actions of B. When presented with the report, A should determine how to obtain further information to make a moral evaluation about the actions of B based on what B intended. The decision should agree with human moral evaluations in similar situations.

We assume initially that an agent A has in-built ethical principles that are encoded as a set of requirements (e.g. [2]). The requirements can be regarded as primary "values", which are accepted as "given" (for example, breaking other people's property is wrong). However, at least some of the agent's understanding of mental states and ethical values should be learned as a process of development. Thus, we are aiming to combine a "top down" with a "bottom-up" approach to the design of an ethical agent [16].

[1] School of Computer Science, University of Birmingham, UK, email: c.m.kennedy@cs.bham.ac.uk

1.1 Requirements

The key question for A is whether B *desires* to achieve or preserve the state of affairs that is valued by A. For example, A might have an in-built principle that "damage to other people's property must be avoided". To satisfy the test of moral acceptability, B would have to care about the other person's property. It may not necessarily be successful in always avoiding damage.

In the event of an action by B which violates a requirement R, A will attempt to find evidence that B desires to uphold R. To do this, A must generate hypotheses about B's mental states and test them. For example, "if B cares about R then they will be unhappy that R is not upheld" or "If B asked a question showing their ignorance of actions that can damage R then they did not have sufficient knowledge to satisfy R (they might still care about R)".

In the case of the vase scenario, the following are possible explanations for B's behaviour if B is innocent:

1. B didn't see the vase or didn't know that it was fragile;
2. B knew about the vase but was distracted (e.g. looking out the window).
3. B may not have wanted to break the vase, but still broke it deliberately because of some circumstances A doesn't know about. (e.g it was fake and the owner was going to sell it as a counterfeit).

We will focus particularly on hypotheses 1 and 2 in this paper, but we will also discuss some of the challenges posed by 3.

The decision-making process also needs to be transparent. In addition to generating possible hypotheses (as above), A should also explain why it considers one of these to be a possibility and what evidence might support or falsify its hypothesis.

It is important to note that we are not considering the ability to solve moral dilemmas (such as the "trolley" dilemmas [10]) but instead the ability to *be concerned* about the dilemma because the agent cares about human life (or other things valued by humans). Therefore, whatever the decision in a "trolley" dilemma, the agent only fails the test if there is evidence that they did not *care* about human life. If they attempted to apply a moral principle, they pass the test. The agent A need not "blame" the decision-maker B even if the decision is not the one that A would make itself.

2 Architectural Building Blocks

The *Polyscheme* architecture ([5] has been used to model an agent with ToM [4, 3]. Central in the Polyscheme framework is the idea of multiple "worlds" in which a statement can be true or false. Some of these worlds may be counterfactual, where an agent uses internal simulation to "imagine" a situation that is false in the real world. To understand another agent, it first creates a counterfactual world, where it imagines itself to be the other agent B and makes initial

assumptions that some of its own beliefs will be held by B. This process of assuming that some true statements in the real world are also true in the counterfactual world is called "inheritance". As the agent detects differences between itself and B, it "overrides" some of the inherited assumptions.

If we consider Hypothesis 1, agent A imagines itself not seeing the vase and knocking it over; it knows that it would be unhappy and would apologise (because of its moral norms R). Therefore it also imagines that B would feel the same way if it did not see the vase and that it would also apologise (inheritance). As a result, A can test the hypothesis to detect any differences between itself and B, which may override its initial inherited assumptions.

In order to simulate the mental states of B, A first needs to imagine what it would do itself, and what its mental states would be. However, since we are aiming for a developmental approach, it is not practical to just "give" the agent a set of propositions about its own mental states in multiple situations. Its understanding of itself should ideally come from its own experience. This would provide robustness and flexibility in its attribution of mental states to others. Such experience would also allow it to generate autonomous explanations about why it reached a certain conclusion [7]. We will argue that meta-reasoning can provide a foundation for an agent to build a model of its mental states.

2.1 Meta-reasoning

Meta-reasoning is a computational paradigm for "thinking about thinking" [8]. Typically a meta-reasoning component (or "meta-level") monitors and evaluates an agent's problem-solving processes (or "object-level"). In the case of an apartment viewing agent, the object-level is the main reasoning component that collects information on the condition of each room, while evaluating quality and making decisions about whether to ask more questions etc. The meta-level monitors and evaluates the performance of the object-level. For example, is the object-level making predictions (expectations) that are being contradicted? In addition to meta-level monitoring, meta-level *control* makes decisions about object-level processes, such as what goals need to be generated, how much attention should be given to a problem and what needs to be learned.

If we apply meta-reasoning to the Polyscheme representation of ToM, the meta-level is the part of the system that generates and controls simulations, while detecting differences between self and other. The object-level is the actual reasoning within the simulated worlds.

2.1.1 Reasoning traces

For an agent to inspect its own reasoning, a reasoning trace is required [6], which acts like an episodic memory [9] of mental events. A reasoning trace can be in the form of an "audit trail" that is left by an object-level process. Different kinds of trace may be generated. For example, the following information might be recorded in a trace T1:

- What did the agent know initially? What did it see? What information did it consider to be relevant?
- How certain was the agent about its subsequent inferences (if any)?

A different kind of audit trail T2 might be a sequence of "decision events" or "branch points" (as in [14]) where each decision event includes the following kind of information:

- Current goal and the options that were being considered;
- How the options were evaluated (positively or negatively);
- Which option was chosen, and why?

We propose to use the notion of a reasoning trace to represent the "fine structure" of mental states and processes.

2.1.2 How are reasoning traces used?

There are different ways in which a meta-reasoning process can use a reasoning trace in a cognitive architecture. The following are two possibilities:

- Integrity-checking: Meta-reasoning component M1 monitors object-level O1 and checks if the trace satisfies a required pattern. When inspecting T1, the meta-reasoner compares the actual trace with what it *ought* to be. For example: where the assumptions correct? Did it consider all the information? Did it miss out any options when making a decision? This is approximately the approach taken in [12], which emphases distribution of meta-levels to ensure that all reasoning processes are satisfying the requirements.
- Failure diagnosis: the meta-reasoning component checks if the current trace matches a known pattern of reasoning failure. This is the approach taken in [6], which uses a taxonomy of different types of failure. An example failure type is "contradiction between expected and actual observations".

In both cases, a set of generic trace patterns is held in long-term (semantic) memory, while specific instances of traces (audit trails in episodic memory) are matched against the patterns. This is how an agent monitors the integrity of its reasoning or "makes sense" of its experience, depending on the respective paradigm. Both of these approaches may be combined.

Different kinds of meta-reasoning may use different paradigms and trace information. For example, one meta-reasoner (M1) might specialise in detecting lack of knowledge or understanding, while another (M2) specialises in detecting distraction or forgetting due to competing pressures. Detailed reasoning traces can be generated and inspected in a language such as *Funk2* [14].

2.2 Developing Representations of Mental States

The traces T1 and T2 represent mental states on a high level, and do not include the computational "fine structure". To determine B's experience, the agent needs to simulate what it means to "know" or to "see" something. One solution is to provide a mapping from a low level trace (the fine structure) to a high level mental concept or process, which may itself be embedded in a high level trace (such as T1). This originates from the agent's own understanding of its information processing. Therefore, we also need a process by which the agent *learns* to understand its mental states (a self-model), since we are aiming for some bottom-up development in the agent's ability to make ethical decisions.

2.2.1 Mental concepts as trace patterns

Mental concepts in T1 and T2 may be defined in terms of lower level trace patterns, of which specific instances are actual histories. For example, a trace pattern might define the concept of "knows about x" as "repeatedly able to retrieve with certainty the details of x when questioned". An information retrieval system (object-level) can leave

a trace of its actual success or failure in answering queries with a certainty level. A meta-level can then evaluate this trace by comparing it with the ideal pattern (representing "knowing") to give the agent an understanding how well it knows or can remember a concept. Therefore A can test if B knows something using this definition, since the concept is also associated with patterns of external behaviour that can be observed.

Similarly, forgetting can be defined as failure to retrieve an item that the agent can remember being able to retrieve previously. In this case the trace pattern in the semantic memory can refer to the content of a previous episodic memory ("I can remember knowing about x, but now I've forgotten"). Degrees of certainty can be defined in terms of contradiction between expectation and reality (see for example, [6]).

2.2.2 *Learning self-models*

An agent can develop a model of its own information processing by self-observation, allowing it to learn general patterns from its reasoning traces (on different levels). Such a self-model can enable the agent to predict its mental state in a hypothetical or future situation [13]. In this way the agent can build up self-familiarity so that it can simulate its predicted mental states in counterfactual reasoning.

2.3 Reasoning about caring

Mental states that are relevant to moral reasoning particularly involve values and goals. We are assuming that A's values are determined by R, the set of requirements that constitute the moral norms of A (encoded using deontic logic or other representation). If we use the integrity-checking paradigm for meta-reasoning, R can also include requirements to be satisfied by mental traces such as T2, which records the decisions made and why. For example, what things *ought* to be valued positively or negatively? What goals are acceptable in what situations? The problem is more complex if the agent is to test hypothesis 2 above (distraction). In this case it also needs to understand about adverse circumstances affecting mental states, such as limited resources and conflicting goals. For this it requires experience and self-observation over time.

For the purposes of low-level computational representation, we can define "caring" as *persistence* in attempting to satisfy a goal in the presence of conflicting goals or resource pressures. In such a situation the agent will also be creative and autonomous in the way that it attempts to satisfy the goal. In information-processing terms it will *spend computational resources* searching for different ways of solving the problem; it will try to acquire new information (e.g. by asking questions and exploring).

On the other hand if it makes a decision that something is not important, it will de-allocate mental resources to it. In the apartment example, the agent A can look for evidence on whether B was attentive and walked slowly through the apartment asking questions, or whether B was multi-tasking (e.g. making a phone call) while walking between rooms. If B cares about R it does not need to be successful in satisfying R. However, if B fails it will evaluate the resulting state negatively and its behaviour will make this clear.

2.4 Reasoning about control

Being distracted or forgetful implies a *lack of control* over mental processes. The following example traces record events relating to control. T3: changes in working memory over time:

- History of top-down attention focus: things which were deliberately added to working memory.
- History of "salience" events: bottom-up emergence of ideas or noticing of details.

T4: Perception of the difficulty of a problem, or pressures:

- History of salience events which were disruptive;
- History of changes in subjective difficulty of a task due to other pressures;

These traces must be generated computationally. Therefore we must ask how an agent can detect that it has control over its mental states? For example, how does the agent know whether an item appearing in working memory is a result of deliberately remembering or imagining something, or whether it just appeared because it noticed something (salience).

This problem might be solved using causal tracing [14]. This can be used to track a chain of decisions and inferences which originated from an initial decision. In the apartment example, the initial decision to view the apartment can lead to a choice of which room to enter first, resulting in entering the kitchen. This in turn leads to a choice of what appliances to inspect first, how to evaluate them and what questions to ask. Each decision is a branch point in the trace.

Causal traces can be applied at different levels of abstraction, and do not only apply for modelling introspection. The lowest level might be conditional branches in a piece of executing code. On a higher level, an agent can generate an intention to remember something and then the resulting item in memory can be traced back to the intention. If an unexpected item appears in working memory, it can be attributed to a distraction that the agent is not currently "in control" of. For example, distractions may be due to bottom-up perceptual processes that are allowed to interrupt the "top down" control in situations where the interruptions are important for survival. Cognitive architectures with variable attention filters [17] are relevant in this case, where emotions in particular are modelled as "interruptions".

In the apartment example, if A has experience of situations where it fails to maintain attentional control in the presence of distractions, it can also attribute these states to B and check B's subsequent behaviour that would be consistent with this explanation.

3 Grand Challenge: a Turing Test for Moral Cognition

The above architectural building blocks might help an agent to make decisions that are similar to human moral decisions in restricted scenarios. The longer term challenge is a more general system that would pass a "Moral Turing Test" (MTT) [1], where the agent's decisions in a wide range of scenarios would be compared with a human's decisions. If an observer cannot distinguish between the two, the agent would pass the test.

Designing an agent that can pass an MTT provides an opportunity for detailed analysis of human cognitive and emotional mechanisms involved in moral decisions [16]. In particular, the role of *empathy* is important, as well as the capability to make exceptions to a rule.

3.1 Simulation and Empathy

Polyscheme allows for a process of "backward inheritance" [3] where new information populating the simulation of B may be inherited back to A's "real" world, allowing A to actually "feel" what it is like to be B. The backward inheritance process might be useful

in triggering more inferences about possible explanations for B's behaviour, since A is allowing itself to be affected by the simulation as if it were a real percept. Both forward and backward processes may be important characteristics of empathy, which has a role in moral cognition [11].

3.2 Autonomy and Flexibility in Understanding Moral Norms

The need for flexibility and willingness to learn from the other agent is a significant challenge in meta-reasoning. Similarly, the ability to extend and revise R autonomously may be necessary in situations where B's actions do not match any known scenario of guilt or innocence (Hypothesis 3 in the vase-breaking example). In this context, a willingness to learn more about the experience of the other agent implies some "respect" for B. Such respect may be implicitly consistent with R, if R includes social rules of "fairness" and listening. However, the behaviour of B may contradict an explicit rule in R (deliberate breaking of property). Once A understands the new situation, it can experience empathy for B (due to backward inheritance) and conclude that an exception can be made in this case. If the backward inheritance is an in-built feature of its architecture, it cannot choose to suppress empathy, but it might override some rule in R that is not consistent with empathy in the new context. Therefore A can have a robust and flexible understanding of the moral norms in R which is grounded in its own "experience" and it might be possible to extend or revise the norms as necessary.

4 Summary and Conclusion

In this paper, we have presented a scenario in which an artificial agent is required to make ethical decisions that are similar to typical human decisions, given a report of human behaviour. We have proposed to combine meta-reasoning with a mental simulation architecture such as Polyscheme. Meta-reasoning can help an agent to build models of its own mental states on different levels of abstraction (self-familiarisation). Such a detailed self-understanding based on reasoning traces can help an agent to generate rich simulations of other minds by imagining that its own mental states apply to the other agent. This in turn helps it to make detailed predictions about the other agent's behaviour.

REFERENCES

[1] Colin Allen, Gary Varner, and Jason Zinser, 'Prolegomena to any future artificial moral agent', *Journal of Experimental and Theoretical Artificial Intelligence*, **12**, 251–261, (2000).

[2] Konstantine Arkoudas, Selmer Bringsjord, and Paul Bello, 'Toward Ethical Robots via Mechanized Deontic Logic', in *In Proceedings of the 2005 AAAI Fall Symposium on Machine Ethics*, (2005).

[3] Paul Bello, 'Cognitive Foundations for a Computational Theory of Mindreading', *Advances in Cognitive Systems*, **1**(1-6), (to appear).

[4] Paul Bello and Marcello Guarini, 'Introspection and Mindreading as Mental Simulation', in *The Annual Meeting of the Cognitive Science Society (CogSci 2010)*, Portland, Oregon, (August 2010).

[5] N. Cassimatis, P. Bignoli, M. Bugajska, S. Dugas, U. Kurup, A. Murugesan, and Paul Bello, 'An Architecture for Adaptive Algorithmic Hybrids', *IEEE Transactions on Systems, Man and Cybernetics, part B*, **40**(3), 903–914, (2010).

[6] Michael T. Cox, *Introspective Multistrategy Learning: Constructing a Learning Strategy under Reasoning Failure*, Ph.D. dissertation, Georgia Institute of Technology, 1996.

[7] Michael T. Cox, 'Metareasoning, Monitoring, and Self-Explanation', in *Proceedings of the First International Workshop on Metareasoning in Agent-based Systems at AAMAS-07*, pp. 46–60, (2007).

[8] Michael T. Cox and A. Raja, 'Metareasoning: an Introduction', in *Metareasoning: Thinking about Thinking*, eds., M. T. Cox and A. Raja, 3–14, MIT Press, (2011).

[9] Nate Derbinsky and John E. Laird, 'Efficiently Implementing Episodic Memory', in *Case-Based Reasoning Research and Development*, LNCS Volume 5650/2009, 403–417, Springer-Verlag, (2009).

[10] Philippa Foot, 'The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices', *Oxford Review*, **5**, (1967).

[11] Alvin Goldman, 'Empathy, Mind and Morals', *Proceedings and Addresses of the American Philosophical Association*, **66**(3), 17–41, (Nov 1992).

[12] Catriona M. Kennedy, *Distributed Reflective Architectures for Anomaly Detection and Autonomous Recovery*, Ph.D. dissertation, University Of Birmingham, Birmingham, UK, July 2003.

[13] Catriona M. Kennedy, 'Distributed Meta-Management for Self-Protection and Self-Explanation', in *Metareasoning: Thinking about Thinking*, eds., M. T. Cox and A. Raja, 138–167, MIT Press, (2011).

[14] Bo Morgan, 'Funk2: A Distributed Processing Language for Reflective Tracing of a Large Critic-Selector Cognitive Architecture', in *Metacognition Workshop of the third IEEE Conference on Self-Adaptive and Self-Organizing Systems (SASO 09)*, (2009).

[15] David Premack and Guy Woodruff, 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences*, **1**, 515–526, (1978).

[16] Wendall Wallach, 'Robot minds and human ethics: the need for a comprehensive model of moral decision making', *Ethics and Information Technology*, **12**, 243–250, (2010).

[17] Ian Wright, Aaron Sloman, and Luc Beaudoin, 'Towards a design-based analysis of emotional episodes', *Philosophy Psychiatry and Psychology*, **3**(2), 101–126, (1996).