Agents for Trustworthy Ethical Assistance

Catriona M. Kennedy

School of Computer Science,
University of Birmingham, UK
C.M.Kennedy@cs.bham.ac.uk

Abstract

We consider a hypothetical agent that informs humans about potential ethical problems, such as human rights violations. It may be argued that such an agent has to be embodied in the human world, with human-like experiences and emotions. Otherwise, it would be unreliable due to its "indifference" to human concerns.

We argue that a non-human-like ethical agent could be feasible if two requirements are met: (1) The agent must protect the integrity of its own reasoning (including its representations of ethical rules etc.). Therefore it requires a reflective architecture with self-protection. (2) The agent's world should generate events that can be related to ethical requirements in the human world. A step in this direction is intrusion detection based on "policy"(e.g. stating which network hosts can talk to each other using which protocols). The policy requirements can be translated into "acceptable" patterns of network events in the agent's world and the agent can learn to recognise violations. A key question is whether the "policy" can be abstracted to the level of general ethical principles (e.g. specifying honest business relationships) and whether the agent can learn these principles by associating them with events in its own world.

Key words: concerns, control systems, grounding, policy, proactivity, self-protection.

1 Introduction

An ethical assistant agent can be understood as an agent that assists humans in promoting the implementation of ethical rules, for example in society or within an organisation. In particular, the agent could inform humans about potential violations of ethical rules in circumstances where this may not be obvious (for example, due to information overload or emotional states preventing recognition of a problem). Hypothetical scenarios involve varying levels of complexity and range from conflict of interests or distortion of scientific results to human rights violations.

We assume that the ethical rules themselves are not in dispute, and that the agent has a representation of these rules, allowing it to explain its reasoning. The agent must be trustworthy: if its ethical specification is modified by an unauthorised party, it should not just indifferently follow the new rules. For example, if there is a new rule that denies privileges (e.g. human rights) to a group of people the agent should refuse to comply. This "indifference" problem is discussed in more detail in [Kennedy, 2000].

The problem may be overcome if the agent is embodied in a human world and experiences human-like emotions. We argue that such a human-like agent may not be necessary, but that the following requirements are important:

- The agent architecture should provide reflection and self-protection; the agent must care about itself in order to care about other things.
- The software agent's world should contain events produced by human activity, and requirements about such activity should correspond to patterns that the agent can learn to recognise.

2 Concerns, Control Systems and Reflection

A control system provides a starting point from which to develop models of agent "caring" or "concerns" [Frijda, 1986]. An agent acting on behalf of human concerns may be regarded as a control system if it has to maintain the world in a state that satisfies human specified requirements [Kennedy, 2000]. For example, the control system may be a robot that keeps a room tidy while occupants are continually creating disorder. The robot has rules (or other specification) defining satisfactory tidiness (e.g. what objects are allowed to be on the floor, what are acceptable places for other objects?).

Alternatively the agent may be a policy-based intrusion detection and response system (as in e.g. [Balepin et al., 2003]) where the policy is a specification of satisfactory behaviour of objects in a network (for example, what hosts can talk to each other using what protocols?). If the network exhibits behaviour that contradicts the requirements, the agent will attempt to identify the problem, isolate it and remove it if possible, so that the network behaviour returns to a satisfactory state (a capability called "intrusion-tolerance" [Dacier, 2002]). A "state" in the network security domain is normally a sequence of events in short term memory, showing behaviour (e.g. all conversations started, continuing or closed in the last n milliseconds).

In a hostile environment the agent's rules defining acceptable states may be illegally modified or subverted, or its capability to preserve these states may be interfered with. Therefore the agent needs a secondary layer of concerns (a "meta-level" control system) whose purpose is to maintain the agent's own integrity (including its specification rules, software and hardware). Without this reflective layer to "care" about itself, the agent cannot reliably "care" about any human desired states in an environment.

The meta-level is an example of "meta-management" [Beaudoin, 1994] and is also related to self-protection in autonomic computing [Steinberg, 2004]. To avoid an infinite regress of meta-levels, an architecture with distributed (mutual) meta-levels has been introduced and demonstrated as a proof-of-concept [Kennedy and Sloman, 2003].

Therefore, the agent needs a description of its own components along with their allowable behaviour, which is in turn derived from the policy. Consequently, the agent's own actions are governed by the requirements, due to its meta-management.

3 Proactivity and Independence

In addition to self-protection, the agent should have an independent understanding of the human concerns it should act on. In addition, an ethical agent needs to be proactive so that it can point to situations that humans are failing to recognise.

The agent's descriptions of desired states should be associated with low-level patterns detectable by the agent's sensors, and which it can learn to distinguish from undesirable states. For example, "all plates stacked in cupboard" may be satisfied or violated by many configurations in the visio-spatial world, thus giving the robot a large space of experiences and options to explore. Similarly, in the intrusion detection domain, "client C must use protocol P when talking to server S" would be satisfied by many different patterns of activity that are detectable by currently available sensors (e.g. SYN packet logger). It can also be violated in many different ways, each of which can be learned by the agent and counteracted accordingly. This requirement for an agent to experience the "meaning" of a specification relates to the concept of semantic grounding. Examples include [Harnad, 1990] and recently [Gorniak and Roy, 2004] in the visio-spatial world.

Once we have the basic translation from high level policy to low level events, the following properties can enhance the agent's independence:

- Actions: The policy should include allowable actions by the agent, where each action may be implemented by many different motor sequences. (In a network, a "motor" sequence may be a sequence of reconfiguration actions, for example shutting down processes and starting new ones.) The "best" sequence is determined by the agent and depends on its self-protective concerns, which are states of its own software or hardware that the agent found to be most desirable (e.g. efficient) for upholding the policy.

- The agent should learn about states that satisfy or violate the *intended* requirements, independently of the formal requirements themselves. For example, it may observe normal use of a network in the absence of intrusions and acquire its own model of acceptable behaviour. (This is important in an ethical domain because any formalisation of ethical rules is expected to be incomplete; there are always exceptions.)

- The classes of events associated with human desired states should also be "desirable" from the point of view of the agent's self-protection, so that it will be proactive in support of human desired states. This should happen if the self-protective concerns are derived indirectly from the human concerns (as in all examples here). (Exceptions may include situations where a robot is required to crash onto a planet's surface).

4 From Policy to Ethics

A policy-based intrusion detection system may be regarded as a special case of an ethical assistant agent where the "ethical requirements" are limited to behaviour in the network in which the agent is situated.

If we try to extend the notion of policy to cover ethical behaviour in the human world, the software agent no longer has direct access to the events in question. Instead, it must rely on reports and other data produced by humans (A more detailed discussion is in [Kennedy, 2000]. Similarly, its "motor" actions are mostly limited to recommendations it makes to humans.

The similarities and contrasts between an ethical agent and the other two kinds of agent as summarised in the table below.

Type of state	Agent sensing of state	Problems it must detect	Agent actions
Network activity	Network sensor information	Undesired state, unreliable sensors	Reconfigurations, recommendations
Human activity	Human-produced reports	Undesired state, inaccurate reports	Seek more information, recommendations
Physical world	Robot sensor information	Undesired state, unreliable sensors	Physical actions

In the world of human-reported events, "sensor" reliability is a complex issue, with problems ranging from inaccurate reporting to weaknesses in content analysis technology. However, deliberately inaccurate sensors or untrustworthy analysis software is also a problem for a network monitoring agent and to a lesser extent for a robot. In the content analysis case, some limited reconfiguration actions are possible in its own software world, (e.g. focusing attention on interesting information sources).

From the table, we can identify two different layers of ethical requirements.

1. Information reliability and general trustworthiness. For example, if there is a massive increase in the inaccuracy of reports above the normal noise level then this in itself is a suspected ethical problem.

2. Rules about behaviour or desirable states in the external world.

The first problem is already addressed in current research. A security policy is a particular way of ensuring trustworthiness and information reliability. These requirements are also easy to relate to the agent's self-protective concerns (because the system to be protected includes itself) as well as to fault tolerance in general (e.g. the agent needs multiple independent sources to cross-check accuracy and to detect above-average inconsistency).

The second problem relates to more general issues, such as recognising equality of rights of other agents. A speculative area of future work might investigate whether an agent can be taught the meanings of simple ethical concepts by embedding it in a world of other (simulated) agents and showing it examples of desirable and undesirable behaviour in well-chosen scenarios (e.g. involving conflicting goals or limited resources). The aim would be to teach the agent to recognise problems and intervene by finding imaginative ways to act within the ethical rules in its specification.

5 Conclusions

We can conclude that a trustworthy ethical assistant agent is technically feasible, although some challenges have to be overcome. In particular, we need to investigate whether simulated ethical scenarios can provide a form of ethical "grounding" for the agent that promotes its independence and proactivity in the same way as the grounding of requirements in a network or visio-spatial environment.

It is also possible that a non human-like agent may have advantages over an agent with human-like emotions because the former may detect problems overlooked by humans due to stressful or emotional situations. The agent will, however, have its own concerns in the form of preferences for certain states over others. The interaction between the agent's self-protective concerns and its model of human intended requirements is also an area of future work.

References

[Balepin et al., 2003] Balepin, I., Maltsev, S., Rowe, J., and Levitt, K. (2003). Using specification-based intrusion detection for automated response. In *Sixth International Symposium on Recent Advances in Intrusion Detection (RAID 2003)*, Pittsburgh, PA, USA.

[Beaudoin, 1994] Beaudoin, L. P. (1994). *Goal Processing in Autonomous Agents*. PhD thesis, University of Birmingham.

[Dacier, 2002] Dacier, M. (2002). Malicious- and Accidental-Fault Tolerance for Internet Applications (MAFTIA) – Design of an Intrusion-Tolerant Intrusion Detection System. Project deliverable - D10.

[Frijda, 1986] Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

[Gorniak and Roy, 2004] Gorniak, P. and Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21.

[Harnad, 1990] Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

[Kennedy, 2000] Kennedy, C. M. (2000). Reducing indifference: Steps towards autonomous agents with human concerns. In *Proceedings of the 2000 Convention of the Society for Artificial Intelligence and Simulated Behaviour (AISB'00), Symposium on AI, Ethics and (Quasi-) Human Rights*, Birmingham, UK.

[Kennedy and Sloman, 2003] Kennedy, C. M. and Sloman, A. (2003). Autonomous Recovery from Hostile Code Insertion using Distributed Reflection. *Journal of Cognitive Systems Research*, 4(2):89–117.

[Steinberg, 2004] Steinberg, D. H. (2004). What you need to know now about autonomic computing, part 1: Introduction and overview, at: `http://www-106.ibm.com/developerworks/ibm/library/i-autonom1/`. Last consulted March 2004.