External attribution of intentional notions to explain and predict agent behaviour

Conference Paper · January 2001

DOI: 10.1145/375735.375865 · Source: DBLP

Catholijn M. Jonker
Delft University of Technology

Jan Treur
Vrije Universiteit Amsterdam

External Attribution of Intentional Notions to Explain and Predict Agent Behaviour

Catholijn Jonker
Vrije Universiteit Amsterdam
Department of Artificial Intelligence
De Boelelaan 1081a
1081 HV Amsterdam The Netherlands
Tel. +31 20 444 7743
jonker@cs.vu.nl

Jan Treur
Vrije Universiteit Amsterdam
Department of Artificial Intelligence
De Boelelaan 1081a
1081 HV Amsterdam The Netherlands
Tel. +31 20 444 7763
treur@cs.vu.nl

Wieke de Vries
Universiteit Utrecht
Institute of Information and
Computing Sciences
Padualaan 14, De Uithof
3584CH Utrecht
wieke@cs.uu.nl

Categories and Subject Descriptors

I.2.11 [**Computing Methodologies**]: Artificial Intelligence – *languages and structures.*

General Terms

Human Factors, Theory, Verification.

Keywords

Beliefs, desire, intention, BDI, temporal representation.

As agent behaviour often goes beyond purely reactive behaviour, nontrivial means are needed to understandably describe their behaviour. An attractive feature of intentional notions (cf. [3], [10], [11]) to describe agent behaviour is that these notions offer a high level of abstraction and have intuitive connotations. An agent decides to act and communicate based on its beliefs about its environment and its desires and intentions. These decisions, and the intentional notions on which they are based, generally depend on information just acquired by observations and communication, but also on information acquired in the past.

The formal analysis of intentional notions in agents as performed in this paper can be compared to the study of animal behaviour. By field studies and experiments a biologist gathers a large amount of data on the actions of the animal in various situations. Based on this empirical data, explanations are designed of why the animal acts like it does, and then these explanations are tested in new situations. The approach followed in this paper is similar: based on a set of externally observed behaviour traces of another agent, notions such as beliefs, desires and intentions are attributed in such a way that an easy to understand explanation can be given of observed behaviour, and behaviour can be predicted. It is shown how, on the basis of formally defined criteria, the process of attribution can be automated, and built in as a capability of an agent. The agent designed in this manner is able to autonomously and dynamically attribute intentional notions to model other agents' behaviour.

In the formalisation introduced, externally observed behaviour traces of the agent are formalised as temporal sequences of the agent's input and output information states. A temporal language is introduced to express properties on behaviour. In terms of this temporal language, formal criteria are identified that express when a (temporal) formula is an adequate representation of a belief, desire or intention describing an agent's behaviour. These criteria are used in the implementation of a component-based agent architecture that indeed is capable of automatically identifying beliefs, desires and intentions of another agent based on observed behaviour.

To be able to characterise internal representations in terms of external notions, temporal definability notions are formulated and related to each other, inspired by Beth's Definability Theorem from classical logic (cf. [2], [8]). It is shown that assuming, among others, a Determinism Assumption, all representations in the agent are definable in terms of external notions.

The formal analysis and implementation presented in this paper differs from the approaches in e.g., [3], [10], [11] in that it relates intrinsically internal notions to external notions, like observations, communications and actions. Criteria (necessary and sufficient) in terms of external notions are presented which a notion has to satisfy in order to be called a belief, desire, or intention. The criteria allow both for externally ascribing motivational attitudes to agents (that may not use any belief, desire or intention internally) by defining these notions in terms of the external behaviour of the agent, and for analysis of internal notions.

Within multi-agent systems this allows an agent to observe the behaviour of other agents, to attribute justified intentional notions to their behaviour, and (partially) predict the behaviour of the other agents. For example, an agent responsible for the safe use of a car might learn to predict that a human standing on the edge of the pavement and looking across the road is probably intending to cross the road.

The approach works well if the time frame is finite, or the natural numbers. But it will work also if the set of (non-discrete) traces is restricted to those for which a more general property called *finite variability* (cf. [1]) holds: between any two time points only a finite number of changes occur, and after each change, a first time point exists for the new state.

An approach that in some aspects is similar in perspective to ours, is that of [12]. They ascribe knowledge to so-called situated automata, which are processes that do not have any internal representation of knowledge. A process with a certain internal state v knows φ if φ is true in all environment situations which are possible when the

AGENTS'01, May 28-June 1, 2001, Montréal, Quebec, Canada.
Copyright 2001 ACM 1-58113-326-X/01/0005…$5.00.

process is in state v. Our approach for ascribing beliefs is different; we relate belief to the acquired information on the environment. Furthermore, Rosenschein and Kaelbling give no account of desire and intention, which is a main contribution of our paper. The same holds for recent work presented in [13], which concentrates on the informational aspects, and abstracts from motivational and temporal aspects; actually, in [13] exploration of the temporal aspects, as presented above, is mentioned as one of the four items on the list of issues for future work.

From a fundamental philosophical perspective the approach presented here provides a formalisation of views on the explanation of behaviour, as addressed informally in, e.g., [5]. From this fundamental perspective the characterisations introduced, provide a formally defined bridge between an agent's mentalistic notions such as beliefs, desires and intentions, and materialistic notions such as observation and action performance in the world.

From an application-oriented perspective, in the first place, by means of the implementation of a dedicated agent architecture it has been shown that the defined notions and criteria provide a well-defined basis to develop applications of agents that monitor and interpret the behaviour of other agents.

A second type of application of this work can be found in verification of agents internally designed on the basis of a BDI-model. The criteria presented in this paper can be used to verify whether an internal representation, meant to represent some intentional notion, is a correct formalisation of such an intentional notion.

As a third use from an application-oriented perspective, the results presented in this paper are relevant for Requirements Engineering for distributed and agent systems; e.g., [4], [6], [7], [9]. Requirements for agents often concern behaviour; analysis and specification of such requirements is a difficult process. In practice, specification of requirements for simple reactive behaviour is feasible, but if the behaviours become more complex, requirements specification becomes much harder. The importance of using more abstract notions in requirements specification, as opposed to the more directly formulated behaviour constraints, is also stressed in, e.g., [4]. Ideally, to support reuse of agents, the aim is to specify behavioural requirements without any reference or commitment to the internal structures or states of the agent. However, in practice, when specifying more complex behaviour, often not only reference is made to the dynamics of input and output states of the agent, but also to internal states. This may obstruct replacement and reuse of agents; if another agent is introduced it may have a different internal structure. One possible solution for this problem is to restrict reuse to agents with some comparable unified standard internal structure, for example a standardized BDI-structure.

The solution that can be proposed on the basis of this paper is a different one. It is shown that to be able to use high level concepts in specification of behavioural requirements, it is not necessary that the agent actually possesses these concepts. The paper shows how these concepts can also be attributed from outside, and still have a formal definition in terms of the input and output states, as required within a principled Requirements Engineering process. This combines the best of two worlds: (1) requirements specification at a higher level of abstraction, and (2) not demanding a specific internal structure within the agents.

References

[1] Barringer, H., Kuiper, R. and Pnueli, A., (1986). A Really Abstract Concurrent Model and its Temporal Logic. In: *Conference Record of the 15th ACM Symposium on Principles of Programming Languages*, POPL'86, pp. 173-183.

[2] Beth, E.W. (1953). On Padoa's method in the theory of definition. *Indag. Math.* 15 (1953), pp. 330-339.

[3] Cohen, P.R. and Levesque, H.J. (1990). Intention is Choice with Commitment. *Artificial Intelligence* vol. 42 (1990), pp. 213-261.

[4] Dardenne, A., Lamsweerde, A. van, and Fickas, S. (1993). Goal-directed Requirements Acquisition. *Science in Computer Programming*, vol. 20, pp. 3-50.

[5] Dretske, F.I. (1991). *Explaining Behaviour: Reasons in a World of Causes*. MIT Press, Cambridge, Massachusetts.

[6] Dubois, E., Du Bois, P., and Zeippen, J.M. (1995). A Formal Requirements Engineering Method for Real-Time, Concurrent, and Distributed Systems. In: *Proceedings of the Real-Time Systems Conference, RTS'95*.

[7] Herlea, D.E., Jonker, C.M., Treur, J., and Wijngaards, N.J.E. (1999). Specification of Behavioural Requirements within Compositional Multi-Agent System Design. In: F.J. Garijo, M. Boman (eds.), *Multi-Agent System Engineering, Proc. of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99*. Lecture Notes in AI, vol. 1647, Springer Verlag, 1999, pp. 8-27.

[8] Hodges, W. (1993). *Model Theory*. Cambridge University Press.

[9] Kontonya, G., and Sommerville, I., (1998). *Requirements Engineering: Processes and Techniques*. John Wiley and Sons, New York.

[10] Linder, B. van, Hoek, W. van der, Meyer, J.-J. Ch. (1996). How to motivate your agents: on making promises that you can keep. In: Wooldridge, M.J., Müller, J., Tambe, M. (eds.), *Intelligent Agents II. Proc. ATAL'95*. Lecture Notes in AI, vol. 1037, Springer Verlag, pp. 17-32.

[11] Rao, A.S. and Georgeff, M.P. (1991). Modelling Rational Agents within a BDI-Architecture. In: (J. Allen, R. Fikes and E. Sandewall, ed.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (KR'91), Morgan Kaufmann, 1991, pp. 473-484.

[12] Rosenschein, S. and Kaelbling, L.P. (1986). The Synthesis of Digital Machines with Provable Epistemic Properties. In: J.Y. Halpern (ed.), *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge (TARK'86)*, Morgan Kaufmann, pp. 83-98.

[13] Wooldridge, M.J. (2000). Reasoning about Visibility, Perception and Knowledge. In: Jennings, N.R., and Lespérance, Y. (eds.), *Intelligent Agents VI, Proc. ATAL'99*. Lecture Notes in AI, Springer Verlag, 2000.