

Representing and Revising Norms and Values Within an Adaptive Agent Architecture

Nick Lacey and Henry Hexmoor
Computer Science and Computer Engineering
University of Arkansas
Fayetteville
AR 72701
USA
{nlacey,hexmoor}@uark.edu
Telephone : 501-575-{4024,2420}
Fax: 501-575-5339

February 22, 2002

Abstract

This paper presents an adaptive agent architecture which is capable of representing and manipulating social notions such norms and values. The architecture stems from a coherence-based approach to perception and belief revision, in which high level explanations and constraints are used to guide the search process to the most coherent hypothesis. The addition of norms and values to this architecture will lead to the creation of agents which can form, update, and revise complex beliefs concerning the intentions and probable behaviour of other agents.

Keywords: Multi-agent systems, Knowledge Representation

1 Introduction

In this paper we present a novel method of representing data concerning the social configuration of agents within a multi-agent system. The architecture we present builds on a tested implementation but has not yet been implemented itself. This approach combines two fields of research, namely:

1. The representation of knowledge concerning

social modalities, and

2. The implementation of agents based on coherence

In [5], Hexmoor and Beavers discuss extending the traditional BDI approach to agent design by also considering the concepts of values, obligations, and norms. They conclude that adding these new modalities increases the robustness and flexibility of the agents that are produced. In this paper we incorporate norms and values directly into the agent architecture, but we consider obligations to be implicitly represented.

In the field of epistemology, the coherence approach to justification holds that our beliefs are organised in a web-like structure of mutual support [12], [2]. This approach to justification has made its way from philosophy to AI, where it forms the basis of the coherence approach to belief revision [4], [3]. Elsewhere, we have shown that implementing agents based on coherence yields systems which are tolerant to noise and are capable of re-organising belief structures in the light of new information [7], [9].

This paper presents an agent architecture which exploits the powerful adaptive nature of systems based on coherence by constructing a

coherence-based ontology which includes the social notions of values and norms.

Section 2 describes work which has already been done concerning the implementation of an agent capable of coherence-driven belief revision. Section 3 then describes how this architecture is enhanced by the addition of the social concepts of values and norms. Finally, Section 4 examines directions for future work, and presents some conclusions that can be drawn from this research.

2 Belief Revision Based on Symbolic Coherence

This research builds on a previously implemented agent, called **SH**, as it was based on strong holism and coherence. The primary method of inference used by the agent was that of *Explanation-Based Backward Chaining*. This method of inference is central to our proposed architecture and so is described in this section.

The ultimate goal of the backward chaining process is to find the *Final Interpretation* for the current set of sensory inputs. The value of the Final Interpretation may be derived from several alternative explanations, each one representing an alternative high-level conception of the current state of the agent's environment.

The first explanation that the agent will attempt to backward chain is the *Default Explanation*. This represents the most coherent explanation of all information received by the agent prior to the present moment. If the agent is able to find the value of the *Default Explanation* without violating any constraints, then it has no need to investigate any of the alternative explanations, and can adopt the default explanation.

Alternatively, the agent may find that no value for default explanation can be found without violating constraints. This indicates that the agent's internal model is out of step with its environment, so some adjustment is required. If this occurs, as many alternative explanations as necessary are explored. The alternative explanations are derived and then holistically evalu-

ated.¹ Here we are following Thagard and Millgram, who note that human agents make decisions by holistically assessing competing hypotheses [13].

Explanations are *Meta-Level* beliefs, meaning that they concern the relationships between other beliefs, rather than directly representing states of affairs in the agent's environment. Every explanation may incorporate other, lower-level explanations, as well as domain-dependent beliefs concerning facts and relations about the agent's environment. Alternative explanations will usually provide differing mechanisms to derive the same pieces of data. This means that if a piece of sensory data that is necessary for the default explanation is unavailable, the agent may be able to use its knowledge of its environment to derive the missing data from another source.

If the ontology of the agent were to be visualised as a sphere, the central beliefs would be the meta-beliefs while the outermost beliefs would be state dependent assertions. This structure is based on Quine's concept of the web of belief [11]. By associating each belief with the ontological level to which it belongs, the agent has access to a computationally inexpensive method of deriving the relative importance of each belief.

This technique also addresses a common criticism of coherence based justificational structures, namely that they are unable to account for the differing epistemological importance attached to difference classes of belief. The spherical model provides a clear basis for holding that high-level core beliefs are more important to the agent's ontology than perceptual beliefs on the periphery, and hence are less likely to be revised.

3 Extending the Design to Include Values and Norms

We propose extending the agent described above by integrating norms, values, and intentions into its belief revision strategy. This extended agent

¹Limited space prevents a detailed explanation of method used to compare competing explanations. See [9] for further details.

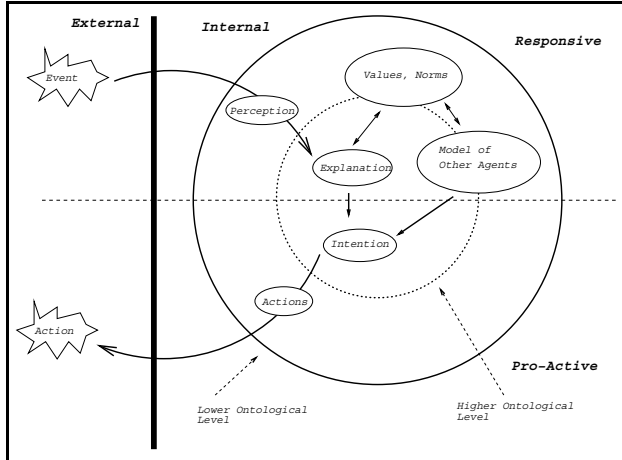


Figure 1: Adding Values And Norms Intentions to Agent **CX**

has been named agent **CX**, as it is based on an extended version of coherence compared to agent **SH**.

Jennings, Sycara and Wooldridge [6] define a responsive system as one which is able to perceive its environment and respond to changes which occur in it. As such, responsiveness can be contrasted with pro-activity. A pro-active agent does not merely respond to its environment, but is able to exhibit goal-directed behaviour.

As shown in Figure 1, the architecture on which agent **CX** is based distinguishes responsive processing from pro-active processing. The diagram also shows how values and norms play a part in the perception explanation process.

Deliberation occurs during both the agent’s responsive and pro-active cognitive phases. However, the deliberation which occurs during both phases is different. During the first phase, the agent is attempting to ascertain the state of the environment. During the second phase, the agent is deciding on the best course of action, given the current state of the environment. Both phases involve a regulatory feedback loop to control processing.

The operation of the second phase has already been described in detail elsewhere [8]. In this paper we confine our attention to the mechanism of coherence-based perception within a multi-agent environment.

3.1 Perception and Explanation

There are two related classes of external events which agent **CX** will be required to perceive:

- Events not caused by active agents
- Events caused by active agents in the environment

Agent **SH** is able to form coherent explanations of passive events using the techniques described in Section 2. Forming explanations accounting for the second class of events involves interpreting and explaining the actions of other agents in the environment.

In our design of agent **CX**, the following steps will be used to interpret the actions of other agents:

1. Perceive the actions of another agent O_n
2. Using its default concept of the norms and values of agent O_n , form competing alternate interpretations which explain the actions of agent O_n
3. If multiple consistent interpretations are possible, use a suitable metric to choose the most coherent interpretation of agent O_n ’s actions.

Constraints are used to test the consistency of the agent’s model of the values, norms, and intentions of the other agents in its environment. These constraints will be provided by the system designer. Their role can be seen as three-fold:

1. Constraints can be viewed as a shortcut, as they allow the agent to detect an inconsistency as early as possible in the knowledge base derivation process.
2. Constraints allow the high-level representation of states of affairs that cannot obtain in the agent’s environment. If a particular constraint is violated under the current explanation, the agent can immediately infer that the explanation is flawed, and thus can invest its energies elsewhere.

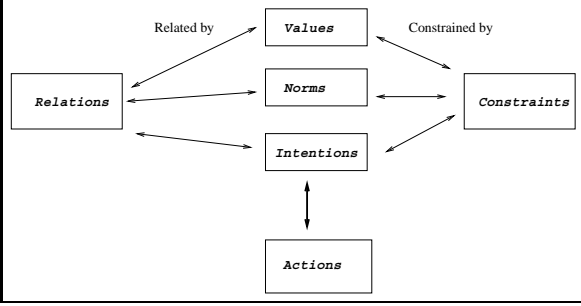


Figure 2: The Relationship Between Values, Norms, and Intentions

- Due to the high-level nature of constraints, it is possible to associate a particular recovery plan with every constraint. By following the recovery plan, the agent may be able to produce a new explanation which successfully resolves the current problem.

Constraints represent a form of pre-parsed domain knowledge which the system designer provides for the agent. As Pollock notes [10], this technique obviates the agent from having to deal with the complexities of the frame problem, at the expense of requiring that domain knowledge be hard coded into the agent at design time.

As shown in Figure 2, values, norms, and intentions are encoded as separate objects within the agent's ontology. Constraint and relation objects are used to encode the relationships between these objects. For example, suppose agent **CX** perceives another agent O_n engaged in the action of driving on the wrong side of the road. This action by itself means very little. In order to determine what action to take, agent **CX** must decide what is the most likely *intention* of agent O_n . Any of the following intentions might be associated with O_n driving on the wrong side of the road:

- O_n is turning across a lane of traffic and as such is temporarily on the wrong side of the road.
- O_n is overtaking another vehicle
- O_n is intending to drive on the wrong side of the road

In order to decide how best to react to its perception that O_n is driving on the wrong side of the road, agent **CX** must first decide which of these is the most likely *intention* of O_n . Once this determination has been made, agent **CX** will be able to determine how best to respond to the situation.

Agent **CX** will use its knowledge of the domain and of the other agents to interpret the actions of O_n . Agent **CX** will be implemented using the Objects library available in SICStus Prolog [1]. Distinct classes of belief will be encoded as distinct objects. To continue using the example mentioned above, some of the values, norms, and intentions involved in driving a car would be encoded as shown below.

```

Value : V1
Purpose : Obey Traffic Laws
Relations : R1, R2
  
```

Norms are also represented as objects. Some norms which may be applicable to an agent which held V1 might be as follows:

```

Norm : N1
Purpose : Drive on Right
Constraints : C1,C3
  
```

```

Norm : N2
Purpose : Drive on Left
Constraints : C2,C3
  
```

Norm N1 represents driving on the right hand side of the road, whereas norm N2 represents driving on the left hand side of the road.

There are two classes of relationship between values and norms, namely, relations and constraints. Relations are used to represent the relationships between different factual beliefs. Constraints represent conditions which must not be violated if a particular fact is to hold.

In the example above, agent **CX** is interested in determining whether or not agent O_n is still obeying traffic laws. The relationship between value V1 and norms N1 and N2 is shown below:

```

Relation : R1
  
```

```

(If V1
  (If In_US)
    (apply_norm N1)
  (If In_UK)
    (apply_norm N2))

```

Relations and constraints are encoded using a syntax based on Lisp. A parser capable of recursively applying relations of arbitrary depth is used as part of the backward chaining process. Relation **R1** represents the fact that an agent which subscribes to **V1** should drive on the correct side of the road. Thus, it should drive on the right if the US, and on the left if in the UK.

Constraints are used to test the validity of the agent's model of a particular state of affairs. The constraints associated with norms **N1** and **N2** are shown below.

```

Constraint : C1
(ensure_equal
  intended_side_of_road right)

```

```

Constraint : C2
(ensure_equal
  intended_side_of_road left)

```

In order for constraint **C1** to be satisfied, an agent must be *intending* to drive on the right hand side of the road. Constraint **C1** is associated with norm **N1**. This means that any model in which agent O_n holds **N1** but is not intending to drive on the right is inconsistent.

Constraint **C3** represents the fact that it would be incompatible for an agent to hold a set of norms and values which would require **C1** and **C2** to be respected simultaneously.

```

Constraint : C3
(incompatible N1 N2)

```

This example shows that by distributing the semantic content of the agent's intentional and social concepts over several classes of belief, we are able to represent a variety of subtle concepts, as well as the relationships between them.

This example also illustrates the important point that values and norms concern *intentions*,

rather than *actions*. This is because one of the claims put forward to support this research is that intentions give meaning to actions. This means that the relations between values, norms, and other beliefs, must operate at the intention level rather than the action level.

3.2 The Ordering of Norms and Values, and Constraint Violation

An interesting feature of the coherence-based belief revision framework concerns the different ontological status attributed to different beliefs. This allows differing levels of importance to be assigned to different constraints and relations within the agent's ontology. This in turn allows the agent designer to create agents with different behaviour by altering the relative strengths attached to different classes of belief.

For example, suppose that agent **CX** currently believes that agent O_n adheres to **N1** and hence **V1**, and as part of its sensory input it perceives O_n driving on the left while in the US. Through constraint **C3**, the agent is able to deduce that O_n *intending* to drive on the left is incompatible with **N1**. In this situation, agent **CX** is faced with two possible interpretations:

1. Agent O_n is intending to drive on the wrong side of the road, and therefore is no longer subscribing to **N1**. This in turn means that agent O_n is no longer subscribing to value **V1**. In this case, agent **CX** would be well advised to immediately cease all co-operation with agent O_n , and to attempt to avoid it as much as possible.
2. Agent O_n subscribes to **V1**, and therefore cannot be intending to violate **N1**. This in turn means that agent O_n cannot intend to drive on the wrong side of the road. In this case, agent **CX** might determine that agent O_n is in trouble and in need of assistance, meaning that agent **CX** should increase its level of co-operation with agent O_n .

Here we can see how it is possible to create two different agents which, when faced with *exactly* the same input information, will respond in

opposite ways. In the first case, the absence of norm **N1** was sufficient to override the currently held belief that O_n subscribed to **V1**. Thus, in this case, norms are held to be more central than values.

In the second case, however, agent **CX**'s belief that O_n subscribes to **V1** is strong enough to allow agent **CX** to reject the suggestion that O_n is violating **N1**. Hence, in this case, values are held to be more central than norms.

It is anticipated that the adaptive nature of this architecture is well suited to allowing agent **CX** to interpret and learn from the behaviour of the other agents in its environment. The fact that this learning occurs at the intention level, rather than the action or perception level, will allow agent **CX** to make decisions at run-time concerning the utility of co-operating or competing with the other agents.

This is because, as part of the perception process, agent **CX** interprets the actions of other agents and ascribes to other agents the intentions which best explain their actions. The constraint satisfaction mechanism which guides this process is ideally suited to detecting discrepancies between presumed and actual agent behaviour. This in turn means that agent **CX** is free to make presumptions about the behaviour of other agents, safe in the knowledge that it will immediately be alerted to any discrepancies.

4 Further Work and Conclusions

We can see three avenues of research that can be pursued from the architecture presented in this paper. The first and most obvious is to complete the implementation of agent **CX**, and to begin experiments concerning agent **CX** in a multi-agent environment.

A second avenue for potential research concerns investigating the effects of placing different classes of belief at different levels within the agent's ontology. As shown in the example above, changing the relative position of values with respect to norms can cause significant

changes in the behaviour of the agent.

Finally, although this paper has not dealt with the issue of multi-agent planning, an agent based on the architecture described in this paper is capable of using the conclusions reached during the responsive cognitive phase while it is planning during the pro-active cognitive phase. The ability of agent **CX** to form adaptive high level beliefs concerning the intentions of other agents in its environment at run-time means that a significant amount of useful information is available to the pro-active intention generation system. Ensuring that this information is used to its full potential is a non-trivial but potentially rewarding exercise.

Thus, we can conclude that the architecture we put forward in this paper will permit the development of agents which are capable of perceiving their environment, responding to perceptions, and pro-actively organising their actions so as to bring about long term goals. The addition of values and norms to the architecture will permit agents based on this architecture, such as agent **CX**, to form and revise high level beliefs concerning the likely social behaviour of other agents.

References

- [1] Jonas Almgren, Stefan Andersson, Mats Carlsson, Lena Flood, Seif Haridi, Claes Frisk, Hans Nilsson, and Jan Sundberg. *SICStus Prolog Library Manual*. Swedish Institute of Computer Science, Kista, Sweden, January 1993.
- [2] Robert Audi. *The Structure of Justification*. Cambridge University Press, Cambridge, UK, 1993.
- [3] Peter Gärdenfors. The dynamics of belief systems: foundations vs. coherence theories. *Revue Internationale de Philosophie*, 44:24–46, 1990.
- [4] Gilbert Harman. *Change in View: Principles of Reasoning*. MIT Press, Cambridge, Mass, 1986.

- [5] Henry Hexmoor and Gordon Beavers. In search of simple and responsible agents. In *Proceedings of the GSFC Workshop on Radical Agents*, MD, 2002.
- [6] Nicholas R. Jennings, Katia Sycara, and Michael Wooldridge. A roadmap of agent research and development. *International Journal of Autonomous Agents and Multi-Agent Systems*, 1(1):7–38, 1998.
- [7] Nicholas Lacey. *Investigating the Relevance and Application of Epistemological and Metaphysical Theories to Agent Knowledge Bases*. PhD thesis, University of Wales, Aberystwyth, 2000.
- [8] Nicholas Lacey, Henry Hexmoor, and Gordon Beavers. Planning at the intention level. In *15th International FLAIRS Conference*, Pensacola Beach, Florida, 2002. FLAIRS. (To appear).
- [9] Nicholas Lacey and Mark Lee. The implications of philosophical foundations for knowledge representation and learning in agents. In Daniel Kudenko and Eduardo Alonso, editors, *Proceedings of the AISB 01 Symposium on Adaptive Agents and Multi-Agent Systems*, pages 13–24, York, UK, 2001. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.
- [10] John Pollock. Perceiving and reasoning about a changing world. *Computational Intelligence*, 14(4):498–562, 1998.
- [11] Willard Van Orman Quine. Two dogmas of empiricism. In *From a Logical Point of View : 9 logico-philosophical essays*, chapter 2, pages 20–46. Harvard University Press, Cambridge, Mass, 1980.
- [12] Ernest Sosa. The raft and the pyramid: Coherence versus foundations in the theory of knowledge. In Paul K. Moser and Arnold vander Nat, editors, *Human Knowledge Classical and Contemporary Approaches*, pages 341–356. Oxford University Press, Oxford, UK, 1995.
- [13] Paul Thagard and Elijah Millgram. Inference to the best plan: A coherence theory of decision. In A. Rom and D.B. Leake, editors, *Goal-driven learning*, pages 439–454. MIT Press, Cambridge, MA, 1997.