INTERPRETIVE AGENTS: IDENTIFYING PRINCIPLES, DESIGNING MECHANISMS

David L. Sallach
University of Chicago*

ABSTRACT

The present paper takes the position that social action inherently involves meaning and, thus, cannot be adequately modeled without representing the interpretive process among agents. The development of interpretive models is challenging, however, and quickly raises issues of computational tractability. A strategy is developed based on three assumptions (agent focus, continuity reduction and orientation fields) and three mechanisms (prototype inference, orientation accounting, and situational definition). When the three mechanisms are incorporated into an action selection mechanism, they provide a model of the interpretive process of social interpretation with a higher level of verisimilitude than many other approaches.

INTRODUCTION

In communication and action, human actors are oriented by meaning. Accordingly, in every situation, we consider, discern, define, attribute, convey, question, dispute, affirm, reconsider and evolve its meaning in a particular instance. Inevitably, the attribution of meaning is an indexical process: the same participants may view shared situations as having distinctive, or even conflicting, meanings. The process of meaning attribution is dynamic, often shifting rapidly, as the interpretation of the actor shapes and informs the subsequent flow of communications and acts.

The process of modeling orientation and meaning does not stand as a new problem. Many significant strategies in artificial intelligence, including semantic nets, logic-based semantics, rule-based inference, neural networks and subsumption, among others, have sought to address this capability. Therefore, a developing initiative to design interpretive agent models has a responsibility to distinguish its strategies from those already explored.

At present, the boundaries of agent design are bounded by the complexity exemplified by the long search for artificial intelligence and the simplicity of reactive agents. When addressing complex environments, the former typically encounters the limits of computational complexity. Simple agents, on the other hand, tend to require drastically simplified environments, for example, reducing social interactions to the random flipping of binary cultural tags (Epstein & Axtell 1996; Lustick & Miodownik 2000; Lustick, Miodownik & van der Veen. 2001).

The strategy developed below is to capture domain complexity without confronting computational limits by emphasizing *shared social prototypes* and the

* *Corresponding author address:* David Sallach, Center for Complex Adaptive Agent Simulation Systems, Argonne National Laboratory, Argonne, IL 60439; *email:* sallach@anl.gov.

interpretations and actions through which they operate. More specifically, the goal is to design and construct three linked mechanisms that are intended to simulate the fluidity of social interaction: 1) prototype inference, 2) orientation accounting and 3) situational definition. These mechanisms are, in turn, based upon three design assumptions, to which the discussion now turns.

INTERPETIVE ASSUMPTIONS

Agent Focus. One assumption of the present initiative is that agent-based modeling and simulation provides a unique and often effective research domain. The nature of this setting provides several potential advantages upon which the strategy seeks to draw. These include: 1) the ability to control the complexity of the topology and artificial ecology, 2) the ability to define the action and communication capabilities available to the agent, and 3) the opportunity to experiment with a variety of algorithms and methods (including neural networks, genetic algorithms and swarm algorithms). All of these forms of flexibility allow simple, proof-of-concept models to evolve toward more complex and realistic assumptions.

Another advantage of agent models is their natural support for social processes. It is true that some agent models are only minimally social (e.g., models that use individual agents as the fundamental unit of analysis with all state and behavior defined at that level). Yet, even atomic agents relying on simple rules (and therefore limited individual intelligence) can produce a result that simulates implicitly social processes such as local comparison or situated learning. In addition, since the model frequently assumes many agents distributed across space, social networks, etc., there is ample opportunity for experimentation with mechanisms that are more fully and intuitively social.[1] It is possible that the very sociality of such strategies may provide a more realistic and tractable approach to the design of interpretive agents than does the direct attempt to model agent intelligence. The present discussion explores one such approach.

Reduction of Continuity. A second assumption of the present strategy is that a significant fraction of agent innovation can be represented by a translation process from a continuous environment to discrete internal models that provide the basis for inference. More specifically, the present model assumes that agents are situated in a complex environment that makes available multiple simultaneous cues or, in a semeiotic sense, signs that are used by the agent in defining the salient features of the current situation, including a complex forms of communication that might, for example, include tone and emphasis of voice, facial expression and body language. Such situational cues are subtle, textured, and may be represented as defined on continuous domains. However, given bounded rationality, each agent "collapses" the richness of the setting into more discrete classifications, rules and schemas.[2] Agents must, for example, determine whether a statement is a fact or a misrepresentation, whether a quip is a joke or an insult, whether a

[1] As has often been argued (Minsky 1987; Sallach 1988; Gasser 1991; Axtell 2003), social algorithms and mechanisms may come to be increasingly important in defining the foundations of computation and information science.

[2] The comparative advantage of particular forms of internal representations comprises an active area of research in itself.

response is indifferent or a threat, and many other complex communications that themselves occur in a larger (ecological, technological, structural) context.

Considering the range of possible interpretations in their virtually unlimited combinatorics is beyond the capacity of the agent and is, thus, inevitably reduced to a finite set of possible alternatives. However, the prospective *responses* of agents may be subtle and nuanced and, therefore, may again be reasonably represented in complex and continuous forms.[3] It follows from the translation assumption that, at every step, the communicative and interpretive processes are a possible source of misunderstanding between (among) agents.

Orientation Fields. A third assumption of this approach is that agents (dynamically) maintain an orientation field with an emotional valence for every relevant agent, object and resource, and that this field forms the context within which inference occurs. In addition to concrete referents, an orientation field also contains typifications of various types and at diverse levels of granularity, which are the focus of agent affectivity at varying levels of intensity.

In general, there is considerable stability in affective commitments. However, events, along with associated cognitive reclassifications, can influence emotional valence (e.g., a trusted employee becomes a competitor). Accordingly, it is assumed that emotion and cognition are integrated into a co-evolving orientation field that shapes successive agent behavior.

INTERPRETIVE MECHANISMS

In the context of these three assumptions, three mechanisms are identified that can be used to (at least begin to) simulate the constitution of meaning. The three mechanisms are: 1) prototype inference, 2) orientation accounting, and 3) situation definition. These will be described in turn.

Prototype Inference. A primary mechanism is the use of conceptual prototypes in comprehending and drawing inferences about the world. Rather than a set of facts, assertions or beliefs, human concepts take the form of prototypes. That is, agent concepts, both individual and collective, manifest a core/periphery structure with concept exemplars, and departures from that prototype, varying along the dimensions that together define the prototype concept.

This conceptual structure is well known and fully documented by cognitive science research (Rosch 1978; 1983; Hahn & Chater 1997), but understanding the process of drawing inferences from a network of prototype concepts is in an early stage. Accordingly, in the near future, methods are likely to be exploratory, and perhaps domain-specific. However, the incorporation of naturally-occurring data structures seems likely to add robustness and plausibility to agent models.

Orientation Accounting. A second mechanism, orientation accounting, is inspired by social pragmatism (Mead 1934; Mills 1940) and ethnomethodology (Garfinkel 1967; 2002). A simple accounting mechanism models the facts that: 1) in preparing a communication or act, an agent considers the likely response of significant others, and 2) if recent communications or acts are challenged by others, the decision must be defended,

[3] For present purposes, it is not assumed that such complex responses will be expressed in natural language.

initially by invoking an anticipatory rationale. Orientation accounting locates the mechanism withinin the previously discussed orientation field. While there are cognitive dimensions in orientation accounting, emotional anomalies must be resolved to achieve relational stability. If the challenge continues, for example, it may be necessary to elaborate the defense and have the elaboration (minimally) accepted, accept disruption of the relationship, or acknowledge a misjudgment. Orientation accounting is implemented as a capability (set of methods) shared by all interpretive agents, and driven by emotional orientation.

Situational definition. A third mechanism is situational classification. The specificity of circumstances, in conjunction with the agent response to those circumstances, provides direct input into the action selection mechanism. An ability to define the salient features of a situation is a vital skillset for human actors, and has been so recognized since the work of W.I. Thomas (1967). Accordingly, it is important to model this skillset, in prototype terms and in sufficient detail to clarify domain specific implications. Situation theory (Barwise1989; Devlin 1991) provides a formalism with which to represent the process.

Situations are frequently extremely dynamic; however, with successive communications or actions having the potential to redefine how participants categorize and respond to the situation. As Sawyer has extensively documented, situations emerge, and are attended to, using the improvisational skills of the participants. The resulting experiences, which cannot be assumed to be the same, even among common participants, then feed back into the structure the agent's orientation field.

The three mechanisms work together to shape meaning in communication and action. Communications and events create a new situation that is defined by agents in terms of existing prototype situations. Inference is then made about causality, constraints and probable outcomes. In the generation of communication and action alternatives, emotional accounting considers how best to justify a course of action, including the possible alteration of the action to improve the response of significant others. After a situational response, and possibly at other points in the future, prototypes are reclassified and emotional valences are adjusted. The operation of these three mechanisms does not exhaust the components of interpretive agents, but they provide a preliminary nucleus.

TOWARD IMPLEMENTATION

Since salient social entities and events are complex, multi-dimensional, variegated and fluid, agents must have a coherent and computationally tractable means of reasoning about them. Prototypes have explored by empirical psychology during the last several decades, but they have not yet been incorporated into agent simulation. For the purpose of this project a prototype is defined by: 1) assembling a set of dimensions[4] along which a particular entity (or related set of entities) may vary, 2) identifying clusters of core values that, relative to which, one or more salient social entities are defined, and 3)

[4] Mathematically and computationally these dimensions will be represented as relational domains, with this extension: in addition to attribute values, domains may aggregate complex entities (cf., Codd 1979) and, especially, prototypes. However, every component of a complex entity must ultimately be reducible to (values defined upon) a relational domain.

reasoning about the location of entities (or sets of entities) and how they may influence agent orientations. These three components represent agent actions available to interpretive agents, and will be designed and implemented during the course of the present project.

An orientation is an emotional valence toward a salient social entity. Orientations tend to be preserved, but can be called into question by events. Major or consequentially-timed events can result in a wholesale reorganization of agent orientations. One constraint on the easy or frequent restructuring of orientations is the fact that they are shared with groups of other agents, those of which the agent is a member, or with which s/he identifies. Both as individuals and groups, these *are* salient social prototypes. This means that the actions constituting orientation management include anticipation of significant agent responses, selection and calibration of possible actions to satisfy such constraints and the generation of accounts in which the consequences of actions for salient individuals or groups can be justified. Together these possible actions define an aspect of agent behavior to be implemented as part of the present initiative.

Interpretive agents, complete with prototypes, orientations and the actions with which they are managed, find themselves in situations which must be interpreted. The agents use prototypes and orientations to generate expectations and act accordingly. Subsequent to the situated event (itself composed of the actions of multiple agents, as well as possible exogenous developments), the event is assessed and prototypes and orientations are realigned accordingly. This process of realignment is constituted by possible actions available to the agent. These actions will be designed and implemented during the course of the present project.

Together agent actions that allow the management and use of prototype inference, orientation accounting and situational interpretation result in an *interpretative aspect* that will be available to the agents used in this project. Since prototypes and orientations are shared among groups, but individually aligned, they provide a basis of common action but are also a source of possible misunderstanding among agents and groups. Such misunderstandings must be negotiated, or otherwise responded to, for coordinated social action to emerge.

Thus, prototypes and orientations constitute a shared social heuristic by which coherent social behavior may be simulated without creating unbounded computational demands. The prospect of modeling interpretive agents and interpretive social processes carries the potential of a new type of social simulation that can capture the complexity inherent in meaning-oriented systems, while still remaining computationally tractable.

DOMAIN COUPLING

While these features and capabilities are designed to be generic and, thus, broadly applicable, their use requires that they be mapped to particular topical domains. Each domain has its own entities, events, types and structures that must be specifically represented in order to develop a reliable model. Thus, both designing generic mechanisms and embedding them are both part of the overall design process.

To illustrate, the generic mechanisms described above will be applied to a stylized two-stage electoral process. More specifically, agents will be programmed to

apply prototype inference, orientation accounting and situational definitions to the issues addressed by political actors facing elections. The interacting interpretations of diverse agents will then generate the creation and dissipation of (partially) shared political orientations relative to their political options.

The situated model containing these complex interactions is implied by the nested political games developed by Tsebelis (1990). Nested games occur when two or more games are played at different but interrelated levels, with one forming a context in which the other is conducted. Strategy choices that are apparently suboptimal may actually be a result of multi-layer interaction.

For example, a two-phase election such as in the French Fifth Republic has an initial partisan phase, followed closely by a coalition phase, in which the previous competitors must close ranks within a week in order to successfully contest the general election (Tsebelis 1990, pp. 187-232). During the first round, candidates are typically motivated to attack the candidate from within their coalition (their competitor) because they both seek support from the same pool of voters. During the second round, potential for the coalition to win the office provides an incentive to support and vote for the coalition candidate.

Tsebelis' analysis of historical trends indicates that a coalition's ability to close ranks in the second phase is influenced by how competitive the coalition partners are with each other and by their prospects for winning the seat. Thus, the results of the partisan phase define a context in coalition phase decisions are made. Tsebelis' analysis is that the less competitive the parties in the coalition are, and the better their prospect of winning the office, the more likely supporters of the losing candidate will transfer their support in the second phase. However, when this pattern occurs, it is because individuals and organizations have pre-existing orientations, and interpret events in particular ways. There are also occasions when Tsebelis' generalization does not obtain, and those patterns also result from the orientations, situational definitions and interactions of participating agents. The challenge is to construct simulation models that can capture the closing of ranks *as well as* the failure to close ranks in their situated specificity.

Tsebelis constructs a game theoretic model of these tensions. The resulting framework is sometimes criticized, however, for failing to adequately address equilibrium selection. In a broader sense, its virtues are also its faults, in the sense that the additional complexity introduced by the concept of game nesting results in a model that is analytically intractable. The nesting of layers presses the design of agent simulation beyond the standards of current practice as well, in that agents may hold views that are ambivalent and internally inconsistent. Thus, the example may usefully illustrate the prospective advantages of interpretive design as well.

Imagine, for example, that there are thousands of voters and dozens of issues of varying salience on which they may hold opinions. Multiple parties and their leading candidates attempt to attract voters who share their positions and, in some cases, evolving toward positions that will attract key segments of the electorate. A larger number[5] of newspapers serve a particular reading audience and take positions on a variety of issues. Together, and through their interactions, these individuals and organizations define multiple orientation fields: their own and those the share. The result gives rise to a dynamic co-evolving process.

[5] The exact number of issues, parties and newspapers are parameters to be varied across multiple runs.

In defining an interactive model, we can further assume that there are three small group settings in which actors influence each other: the party committee, the editorial board and the neighborhood group. The latter might be extended to work groups, religious groups, etc., where the form of interaction is a sharing of orientations as each tries to convince others and/or learn others' information relative to the question of who to support in the relevant phase of the election. In each setting, individuals bring their personal orientation field, which may evolve based on interaction with others in the process of arriving at shared strategies to achieve group goals.

In sum, when the structure of a two-phase electoral process is translated into a simulation model, where heterogeneous voters from across the political spectrum are confronted with a variety of candidate strategies and first-phase outcomes, as well as the commentary of various editorials, the contribution of situational models becomes evident. Specifically, because each agent has a potentially distinct social background and particular position within the field of contested policies, each voter castes a ballot, not only in the context of a private (subjective) state, but also of a unique situation.

REFERENCES

Axtell, Robert L. 2003. Economics as distributed computation. Pp. 3-23 in T. Terano, H. Deguchi & K. Takadama, eds., *Meeting the Challenge of Social Problems via Agent-Based Simulation.* Tokyo: Springer.

Barwise, Jon., 1989, *The Situation in Logic*, CSLI, Stanford, California.

Collins, Randall., 1981, "On the Microfoundations of Macrosociology," *American Journal of Sociology*, 86 (5), pp. 984-1014.

Collins, R., 2000, "Situational Stratification: A Micro-Macro Theory of Inequality," *Sociological Theory*, 18 (1), pp. 17-43.

Devlin, K., 1991, *Logic and Information*, Cambridge University Press, New York, New York.

Devlin, K., 1994, "Situation Theory and Social Structure," pp. 197-237 in M. Masuch & L. Polos (Eds.), *Knowledge Representation and Reasoning under Uncertainty* Springer-Verlag, Berlin, Germany.

Devlin K. & D. Rosenberg, 1993, "Situation Theory and Cooperative Action," pp. 213-264 in P. Aczel, D. Israel, Y. Katagiri & S. Peters (Eds.), *Situation Theory and its Applications*, Volume 3, CSLI, Stanford, California.

Epstein, Joshua M. & Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up.* Cambridge, MA: MIT Press.

Epstein, Joshua M. & Robert Axtell. 1996.

Garfinkel, H. , 1967, *Studies in Ethnomethodology.* Prentice-Hall, Englewood Cliffs, New Jersey.

Garfinkel, H., 2002, *Ethnomethodology's Program: Working Out Durkheim's Aphorism.* Rowan & Littlefield, Lanham, Maryland.

Gasser, Less. 1991. Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence* 47, pp. 107-138.

Goffman, E., 1983, "The Interaction Order," *American Sociological Review* 48 (1), pp. 1-17.

Hahn, Ulrike & Nick Chater. 1997. Concepts and similarity. Pp. 43-92 in K. Lamberts & David Shanks, eds., *Knowledge Concepts and Categories.* Cambridge, MA: MIT Press.

Hilbert, R.A., 1992, *The Classical Roots of Ethnomethodology: Durkheim, Weber and Garfinkel*, University of North Carolina Press, Chapel Hill, North Carolina.

Juarrero, A., 1999, *Dynamics in Action: Intentional Behavior as a Complex System*, MIT Press, Cambridge, Massachusetts.

Lustick, Ian S. & Dan Miodownik. 2000. Deliberative democracy and public discourse: The agent-based argument repertoire model. *Complexity* 5 (4): 13-30.

Lustick, Ian, Dan Miodownik, & A. Maurits van der Veen. 2001. Studying performance and learning with ABIR: A research note," In D. Sallach & T.l Wolsko, eds., *Proceedings of the Workshop on Simulation of Social Agents: Architectures and Institutions, October 6-7, 2000.* Argonne, IL: Argonne National Laboratory.

Mead, George H. 1934. *Mind, Self and Society.* Chicago: University of Chicago Press.

Mills, C.Wright. 1940. Situated actions and vocabularies of motive. *American Sociological Review* 5(6): 904-913.

Minsky, Marvin. 1987. *The Society of Mind.* Simon & Schuster.

Rawls, A.W., 1987, "The Interaction Order Sui Generis: Goffman's Contribution to Social Theory," *Sociological Theory,* 5 (2), pp. 136-149.

Rawls, A.W., 1989, "Language, Self and Social Order: A Reformulation of Goffman and Sacks," *Human Studies,* 12, pp. 147-172.

Rawls, A.W., 2002, "Editor's Introduction," pp. 1-64 in *Ethnomethodology's Program: Working Out Durkheim's Aphorism*, Rowan & Littlefield, Lanham, Maryland.

Rosch, Eleanor. 1978. Principles of categorization. Pp. 27-48 in E. Rosch & B.B. Lloyd, eds., *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum.

_____. 1983. Prototype classification and logical classification. Pp. 73-86 in E.K. Scholnick, ed., *New Trends in Conceptual Representation: Challenges to Piaget's Theory?* Hillsdale, NJ: Lawrence Erlbaum.

Sallach, David L. 1988. A comparison of parallel architectures: Neural and social models of mind. *Proceedings of the Second ACM Symposium on Artificial Intelligence.* Norman, OK.

Thagard, P., 1999, *Coherence in Thought and Action*, MIT Press, Cambridge, Massachusetts.

Thomas, W.I., 1967, *TheUnadjusted Girl*, Harper & Row, New York, New York.

Tsebelis, George. 1990. *Nested Games: Rational Choice in Comparative Politics.* Berkeley: University of California Press.