MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Performance Report 1 Sept 05 – 31 Oct 06
Computational Models for Belief Revision, Group Decisions, and Cultural Shifts

To: Dr. John Tangney
AFOSR
4015 Wilson Blvd, Room 713
Arlington, VA 22203-1954

1 Dec 06

From: Whitman Richards
MIT 32-364
Cambridge, Ma 02139

Contract No. FA9550-05-1-0321 (MIT 6897876)

PI: Whitman Richards
Voice: 617.253.5776
wrichards@mit.edu
Fax: 617.253.0415

The Stata Center, Building 32
32 Vassar Street, Cambridge, Massachusetts 02139-4307
PHONE +1.617.253.6218 FAX +1.617.258.8682 WEB csail.mit.edu

Computational Models for Belief Revision,
Group Decisions, and Cultural Shifts
Annual Report
1 Sep 05 - 31 Oct 06

Whitman Richards
Mass. Inst. of tech.
32-364

Contract No: FA9550-05-1-0321
Comp. Sci. & Artificial Intell. Lab
Cambridge, MA 02139

wrichards@mit.edu

voice: 617-253-5776

Objective: Our aim is to develop computational models of the role of beliefs in decision-making and how beliefs are revised when individuals or groups within a culture are subject to external pressures. Some of these models focus on individual belief revision, others on group dynamics and decision-making, and a later set will address the effects of cultural shifts. Data under study, or being collected, include the sacred values, behaviors and evolution of both non-violent and violent social networks in a variety of cultures. Model development ranges from understanding clique formation in social groups, to exploring the role of trust in network stability, the effect of role models, or of other types of influence on decision-making, especially in strategic planning.

Activities: The initial step in the MURI was to establish collaborations and to explore joint research projects. To this end, many meetings among individual MURI members took place in the first 6 months of the start date. These meetings jump-started several joint projects which are now well underway: e.g Medin (Northwestern) – Tenenbaum (MIT); Forbus & Medin (Northwestern) – Winston (MIT); Richards (MIT) – Atran (UMich/JohnJay); Pfeffer (Harvard) – Tenenbaum/Richards (MIT). These collaborations are being facilitated by students and post-docs engaged in joint efforts. Also important was a 1-1/2 day meeting held at MIT in Jan 06 , which led to the exploration of several additional potential collaborations. Other collaborations have now been developed, some still in the formative stages, with the specific directions influenced by progress in the most active areas of the project.

1. **Models under study**

Over the past year, we have been developing and applying six main types of computational models to understand belief revision, decision-making, and network structure and evolution. These are:

(i) Infinite Relational Model (Tenenbaum, Kemp) which is aimed at recovering social network structure from data about beliefs in the society (e.g. Medin-Atran Guatemalan studies) and an extension on Learning Relational Concepts.
(ii) Story Structure & Causal Models (Forbus, Winston, Finlayson) whose long-tern target is automatic abstraction of themes, role-models, etc. from traditional stories in a culture.
(iii) Consistency-Conformity Model (Page, Bednar) which show conditions for convergence to stability for groups with heterogeneous interests;
(iv) Probabilistic Social Network Model (Richards, Atran, Kasturirangan) which is designed to explore the fragility of a network subject to both internal and external pressures.
(v) Multi-agent Influence Diagrams (Pfeffer, Gal), which is a framework for understanding actions and action sequences in strategic settings.
(vi) Ideal Performance Measures (Stankiewicz), which allows an experimenter to evaluate the optimality of performance in sequential decision-making tasks.

Summaries appear in sections below, preceded by an update on data being collected for analysis.

2. Cultural Studies: Sacred & Protected Values

Sacred (Fiske & Tetlock, 1997) and protected (Baron & Spranca, 1997) values refer to two closely-related psychological constructs describing a subset of values that are highly resistant to trade-offs, particularly with secular goods (such as money and other goods readily exchangeable with money), even in cases where a willingness to sacrifice such values would seem to lead to greater net benefits. This runs counter to most normative models of rational choice and has obvious implications for understanding cross-cultural conflict in cases where one or both adversaries make decisions that appear contrary to their own interests.

In the protected-values literature, one explanation for this resistance to tradeoffs has been that protected values (PVs) are about deontological rules (moral prohibitions such as, "do no harm") more than about the consequences of those rules (e.g., the net amount of harm itself) (Baron & Spranca, 1997; Ritov & Baron, 1999). Actions that violate certain prohibitions are seen as unacceptable even if by this violation they lead to better consequences. The sacred-values literature does not distinguish between rules and consequences, but instead suggests that outcomes themselves can be of two distinct types, the first readily convertible to monetary values and exchangeable with other goods, and the second distinctly not so, such that attempts to trade these goods with secular ones will be seen as taboo (Tetlock et al., 2000). Although some sacred and protected values are more or less universal, such as those pertaining to innocent human life, there are generally widespread cultural and individual differences as to whether or not a particular valued good will be sacred or protected, and even in the more universal cases as to what defines the category, e.g., what constitutes *innocent* life (Fiske & Tetlock, 1997; Shweder, et al., 1997).

The study of sacred and protected values is in its infancy, and differences in foci, empirical evidence, and explanatory mechanisms between sacred and protected values leave many questions unanswered. Little is known about the psychological processes leading to different behavioral patterns when people do or do not hold protected and sacred values, about why people with protected values sometimes appear less rather than more concerned with the consequences of their actions, about how people make tradeoffs when confronted with conflicts between different kinds of sacred values, or about the variability and universality of sacred and protected values cross-culturally, as well as of the action tendencies in response to their violations. The following set of projects aim to answer these questions.

The scope of networks under study is quite broad, but can be conveniently divided into two parts: projects involving non-violent social groups and those focused more on violent networks. The non-violent communities being interviewed or modeled include the earlier Guatemalan Itza' & Ladino; and more recenttly Menominee Indians, a small Amish community, evangelical Christians, and hunters and fisherman (both Menominee and Caucasian), all living in two neighboring counties in northeastern Wisconsin (Medin et al.). We have also studied two groups engaged in fostering education in underdeveloped regions outside the US (Kasturirangan, Richards). In addition to these non-violent social groups, we have augmented earlier data on Indonesian, European, and Middle Eastern networks

considered as terrorist risks (Atran et al.). Results are summarized in the next sub-sections.

2.1 Non-Violent Communities (Medin, Atran et al.)

Three broad questions have motivated much of this research: *First*, what role does the domain or content of the protected values (PV) play in the decision processes associated with it? *Second*, how does culture contribute to both the PV's domain and to the decision processes? *Third*, what are the dynamics of the decision processes themselves and how do they depend on culture and domain?

We have begun to explore these questions, by conducting interviews with members of three diverse cultural groups living in rural northeastern Wisconsin (a small Amish community, Menominee Indians, and evangelical Christians, as well as a sampling from other members of the Caucasian population living there.) The interviews contained scenarios relating to distinct moral domains (life, the environment, and human rights), followed by a set of questions measuring PVs and how participants understand them. Such questions included (a) the role of religion in the participants' lives, (b) how the participants made recent moral decisions, and (c) when appropriate, how participants thought forests in the area should be managed (a controversial topic for Menominee Indians, majority culture hunters, and some of the farmers who also managed forests).

An initial analysis of responses to the scenario questions have helped us identify PVs for the different cultural groups as well as some of the characteristics of these PVs. Across groups, PVs were associated with less willingness to compromise, lower likelihood that positions would change over time, and higher ratings of personal importance. At the same time, there were strong cultural differences with regards to which scenarios received PV responses and why.

We are still in the process of analyzing these interviews. Perhaps the most important preliminary finding is that few if any of the reasons participants fail to consider costs and benefits include a *lack of concern* with consequences. Among other things, an unwillingness to consider consequences is related to (1) a pre-existing commitment as to the best choice in the given scenario that precludes the need to weigh hypothetical costs and benefits, (2) a recognition by participants of their own lack of knowledge or certainty with regard to the stated costs and benefits that thus diminishes the validity of weighing them and enhances the value of following moral rules derived from

(subjectively more reliable) sources outside themselves (e.g., from their traditions or religion)[1] , and (3) a recognition by participants of their own lack of knowledge or certainty with regard to the scenario that thus diminishes the validity of weighing costs and benefits and enhances the value of following moral rules derived outside themselves (e.g., from their traditions or religion), and (3) failure among researchers observing apparently non-consequentialist choice to identify the appropriate sources of utility. This suggests that PVs may be more utility-driven than has previously been thought, further suggesting that some apparently irreconcilable conflicts may largely depend on a better understanding of why particular groups hold the PVs they do so that the right kinds of sacrifices and requests can be made.

Following more detailed analysis of the data, the next step will be to conduct further interviews, most of which will include the same participants. These interviews have yet to be designed and will depend on upcoming analyses, so it is premature to state what they will contain, but it is likely that they will include questions to (a) distinguish between diverse motivations for non-consequentialist thinking, and (b) resolve the paradox that PVs are associated both with heightened and diminished concern with consequences, as discussed elsewhere in this report.

[1] Literature on bounded rationality and adaptive heuristics (e.g., Gigerenzer et al., 1999; Gigerenzer & Selton, 2001; Simon, 1955; 1957) as well on social proof (Cialdini, 1984) and adaptive cultural evolution (e.g., Boyd & Richerson, 2001) provide theoretical and empirical support to for the reasonableness of such an approach.

2.2 Protected Values and Cost-Benefit Incentives (Medin, Atran et al.)

2.2.1 Laboratory Studies (Medin)

In laboratory-based experimental studies, we have also actively pursued ideas related to our field studies. For example, Bartels & Medin (in press) and follow-up work by Medin, Bennis, and Bartels suggests that whether morally-motivated agents (with PVs) make decisions by following moral rules of right and wrong or by weighing costs and benefits is susceptible to manipulation: contexts directing attention to net benefits induce consequence-focused choice; contexts directing attention to the permissibility of actions that violate moral rules induce rule-based choice.

In other work, Iliev & Medin (in preparation) looked at some cognitive properties of protected values. Two popular cognitive phenomena were chosen: conjunction fallacy, which is judging a conjunction of events to be more likely than one of its constituents; and anchoring, which involves the influencing of subjects' answers by the presentation of an irrelevant numerical value. They found that participants who had protected values on abortion, for example, did not differ from the rest on their performance on a neutral conjunction fallacy task, but showed a significantly higher fallacy rate when the scenarios were relevant to their values. Alternatively, participants with protected values showed less "non-normative" behavior when an anchoring task was used as a measure of performance. The participants who had protected values on abortions, or found the issue important, showed lower levels of anchoring compared to the rest.

2.2.2 Field Studies: Israel-Palestine Dispute (Atran et al)

Adversaries in political conflicts often conceptualize the issues under dispute as sacred values, rather than simply applying a cost-benefit analysis. What is the interaction between these two parameters of decision-making? Just as some religions forbid any mingling of the sacred with the profane, we speculated that people follow a deontological rule or intuition that forbids any attempt to measure moral commitments to sacred values along an instrumental metric. We have interviewed over one thousand individuals with representative samples from Palestinian members of Hamas, Palestinian students, Palestinian refugees and Jewish Israeli settlers. We predicted that those who hold sacred values would be less antagonistic to compromise over those values if the adversary suffers a similar loss over their own sacred values, even if the adversaries' loss does not instrumentally alter the compromise deal at hand. Specifically, our results show that violent opposition to compromise over issues considered sacred is (1) increased by offering material incentives to compromise but (2) decreased when the adversary makes materially irrelevant compromises over their own sacred values. Thus, simple instrumental incentives for peaceful resolution of violent political and cultural conflicts may fail when those involve threat-contested issues as sacred values.

These conclusions were based on surveys that measured emotional outrage and propensity for violence in response to peace deals involving compromises over issues integral to the Israeli-Palestinian conflict: exchanging land for peace (in experiments with settlers); sovereignty over Jerusalem (in experiments with Palestinian students); the right of Palestinian refugees to return to their former lands and homes inside Israel (in experiments with Palestinian refugees); and recognition of the validity of the

adversary's own sacred values (in each sample). These deals were all hypothetical, but involved compromises that are broadly typical of the types of solutions that are frequently offered within political discourse in the region. In our experiments all participants were opposed to compromise over these issues. In addition, a subset of participants indicated that they had transformed this preference into a sacred value, opposing any trade-off over the relevant issue in exchange for peace no matter how great the benefit to their people. (See Ginges, Atran & Medin, 2006 for details.)

Our results show that when people cognitively view issues and resources as sacred values they may not reason instrumentally at all (i.e. such as to maximize profit.) This result has powerful implications for understanding the trajectory of many cultural, resource and political conflicts, implying that when people view a resource (such as land), an activity (such as hunting a particular animal or farming a certain crop) or an idea (such as obtaining a nuclear weapon) into a sacred value, attempts to solve disputes by focusing on increasing the costs or benefits of different actions can fail. Specifically, instrumental incentives for taboo tradeoffs necessary for peace backfired when Israeli settlers, Palestinian refugees and Palestinian students had transformed issues in the dispute into sacred values. Instead, when dealing with conflicts involving sacred values, culturally sensitive efforts at identifying tragic trade-offs that involve equitable gains or losses over those values may open up new channels for peaceful resolution of otherwise intractable conflicts.

3.0 Modeling emotional factors in decision-making (Forbus, Medin)

Emotional outrage, whether linked to violations in sacred values or not, often plays a dominant role in decision making under adverse pressures. Our modeling effort begins with a representation for linking emotional factors to decision-making goals. Current computational models that aim to appraise the emotional state compute declarative objects with numerical properties that represent emotional evaluations of events and entities in the world with respect to an agent's own goals, beliefs, and intentions. The impact of such evaluations is computed by essentially gathering up those available conceptual structures that have affective content and combining the numbers associated with them using some scheme. It seems reasonable to assume that such information is encoded into long-term memory along with everything else that is stored.

One computational hurdle is to retrieve this information, or, more specifically, information about situations similar to the decision-making task at hand. Our MAC/FAC model has this capacity. (Forbus, Gentner, & Law, 1995.) Given this platform, we now have several interesting directions that we have begun to explore. First, we can assume in addition that emotional states should influence what situations are considered similar. (This seems consistent with the literature on state-dependent recall.) Second, we can assume that experience plays a role in decision-making, and that past experiences frame current decisions in situations similar to those experienced. In our model, both the MAC/FAC components stages would be sensitive to such conceptual situations. Finally, if the affective content in the MAC or FAC stages of processing is accessible to the gathering processes in emotion evaluation, then emotions could be affected even without being able to articulate why, since, for instance, there are no structural bindings in the MAC stage. (That is, "I've got a bad feeling about this" might happen before the reminding is actually available.)

4.0 Concept Map System for Qualitative Models (Forbus & Dehghani)

Imagine an interview scenario inquiring about a change in a forest population, with questions initially focusing on immediate changes, and subsequent questions focusing on long-term effects on the ecosystem. One would like to see how the variables covered in the interviews were related, perhaps as a graphical structure or some other representation that makes clear the mental model of the individual interviewed. Specifically, we would like to know how people in these different cultures reason about moral dilemmas (e.g. Medin & Atran studies.) Our concept map system provides a psychologically plausible formalism for capturing such models. Consequently, we are building a system to enable scientists to more easily construct qualitative representations from protocol data, providing a platform that will facilitate analysis, interpretation, and simulation. (See Forbus et al 2006.)

Our system is being built on an earlier system that was limited to a single qualitative state and aimed at middle-school classroom use. The new system (as yet unnamed) will enable elements of the model to be annotated with specific text from the interviews they were drawn from. It will support multiple qualitative states, thereby providing the ability to capture sequences of qualitative states, and alternative outcomes. At any time, the scientist sees a concept map, but the maps are constructed in such a way that they can be directly translated to formal representations for simulation work.

5.0 Beliefs and Social Network Structure (Tenenbaum & Kemp)

The most obvious method of inferring the social network of a community is to gather information about "who talks to whom", or more generally, "who associates with whom." We have applied our Infinite Relational Model to such data collected by Medin, Atran & Ross (2004) to infer the graphical form of a Guatemalan community. (See Itza' example below.) This model is now powerful enough

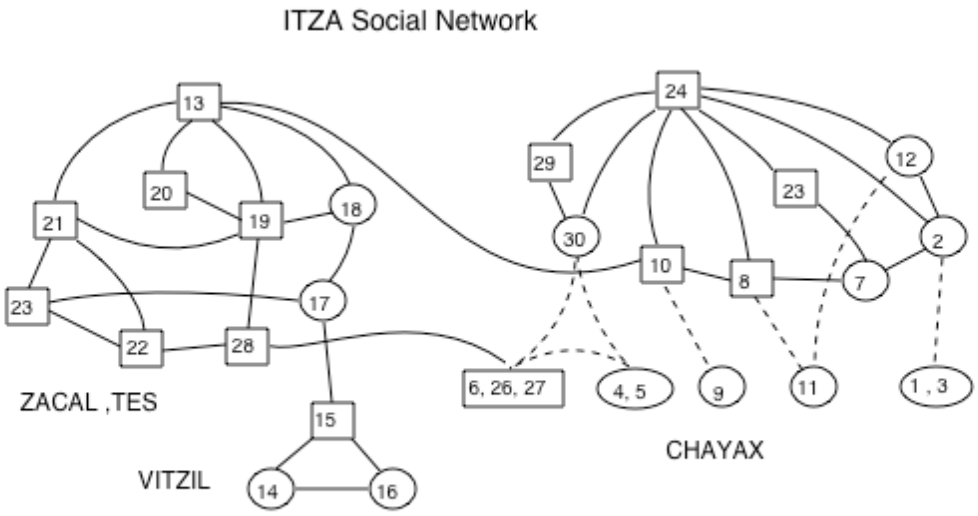

Figure 1: The social network inferred from interviews of Itza' members who were asked to name people "most important for your life." (Atran, Medin and Ross, 2004.)

to recover the structure of a stable social network, given adequate data. (Further work in this area is now aimed at recovering the dynamic structure of evolving networks.)

To show more clearly the scope of our model, consider an oblique and indirect source of data about individuals in a community: the values they place on their natural resources and their environment. As mentioned above, such differences in beliefs and protected values differ dramatically between cultures. For example, in the Medin et al Wisconsin studies, the regard for animals and forest life by Menomee is found to be quite different from that held by Caucasians in the same region. Analogous differences in belief relations among individuals also can be

used to identify groups within a culture. For example, as part of the Guatemalan study, data were collected about the beliefs held by members on the relationships between themselves, animals and plants in their environment. We can show that there are correlations between social group structure and the belief systems people have about biological categories, and how these categories relate,

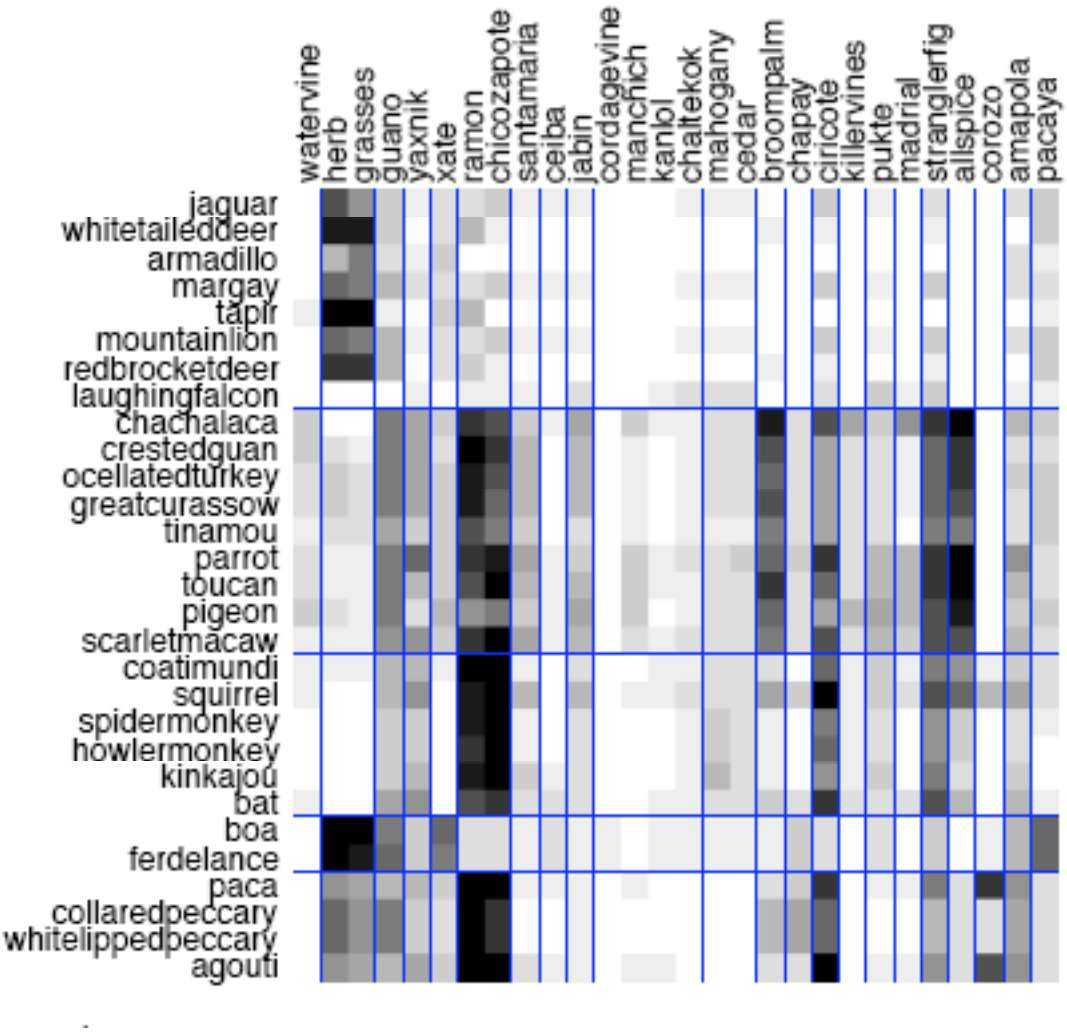

Figure 2 above shows correlations among animals and plants held by the Itza'. (Results are averaged over all members interviewed.) We then extend our examination of such correlations to examine the beliefs of individuals from three cultural groups: Itza', Q'eqchi' and Ladinos. Our model now discovers clusters (or

cliques) of individuals that share similar beliefs. Illustrative members of four of these clusters are shown in Figure 3.

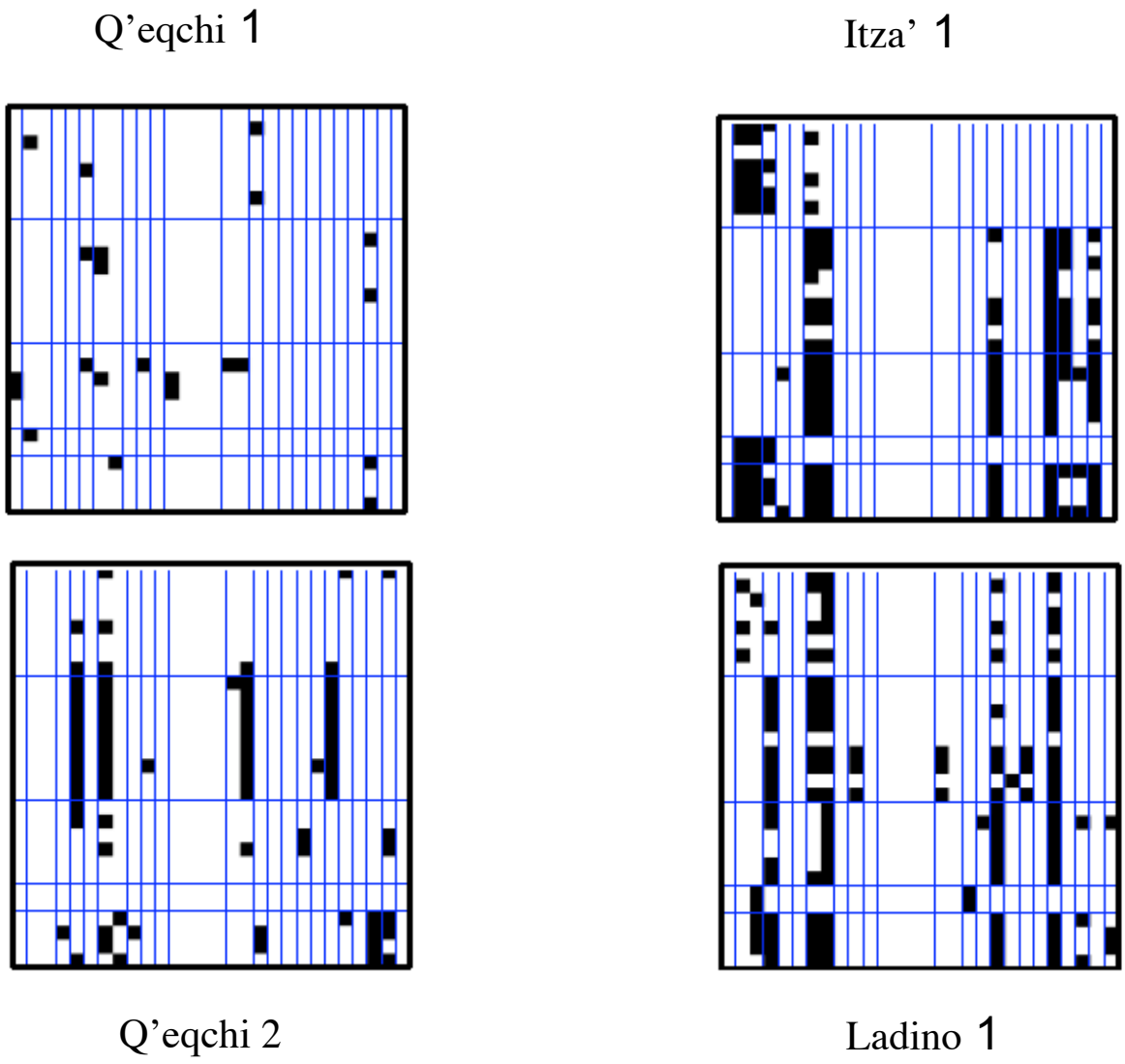

Fig 3: Clustering of sub-groups in a Guatemalan community, based on correlations of individual beliefs of relationships between plants and animals

The importance of this new result is that the individual belief relationships can be used to categorize members of a community. This will greatly facilitate comparisons of the basic structure of social networks, not only across cultures, but also of communities within cultures, and their evolution (in progress.)

6.0 Evolution of Network Structure with Violent Objectives (Atran, Sageman, Magouirk)

Data on the structure and evolution of social groups with violent objectives is sparse and not freely available for scientific research in the academic, policy, and government communities. Nevertheless, our team has made substantial progress in addressing this problem. We now have a good preliminary base of data on several dozen networks in Southeast Asia, Europe, and the Middle East.

Our database consists of two foundations. The first is a detailed categorization of basic biographical and socio-economic data on members (such as jihadists) that includes date of birth, place of birth, nationality, ethnicity, education, (including links to others such as madrassahs, madrassah types), occupation, and class, as well as detailed information on current organizational affiliation and previous organizational affiliation (both militant and non-militant). This part of the database also details incarceration, release, and death information.

The second database foundation addresses the vast network of connections that form the glue that holds the diverse array of jihadists together. This work includes a comprehensive examination of acquaintance, friendship, family, madrassah, and terrorist training (e.g. Afghanistan, southern Philippines) ties. These ties are rigorously documented based on a methodology created to discern differences in the strength of ties over time and in the reliability of the ties based on the available open-source information. All ties are meticulously sourced with a focus on primary documents. This time-series connection data will allow us to examine how counter-terrorist activities affect terrorist network structures. Specifically, it will allow us to test the hypothesis that al Qaeda, Jemaah Islamiyah (JI) and other jihadist groups are moving from a hierarchical organizational model in which a centralized leadership structure directs the overall organization activities, towards a leaderless resistance model in which small groups engage in resistance or violent activity without central coordination.

One of our objectives in creating this database is to ascertain the importance of leaderless resistance in the context of jihadist groups across the world. Specifically we find support for the hypothesis that although counter-terrorist activities by the United States and its allies have decapitated the leadership structure of many terrorist and jihadist groups such as al Qaeda and Jemaah Islamiyah (JI), organizations continue to

thrive using less hierarchical organizational structures. Focusing on al Qaeda and JI and affiliates, we hypothesize that jihadist groups are moving away from hierarchical organizational models towards leaderless resistance models. Under a leaderless resistance model (popularized by the former Klansman and Aryan Nations member Louis Beam), small groups engage in resistance or violent activity independently without central coordination.

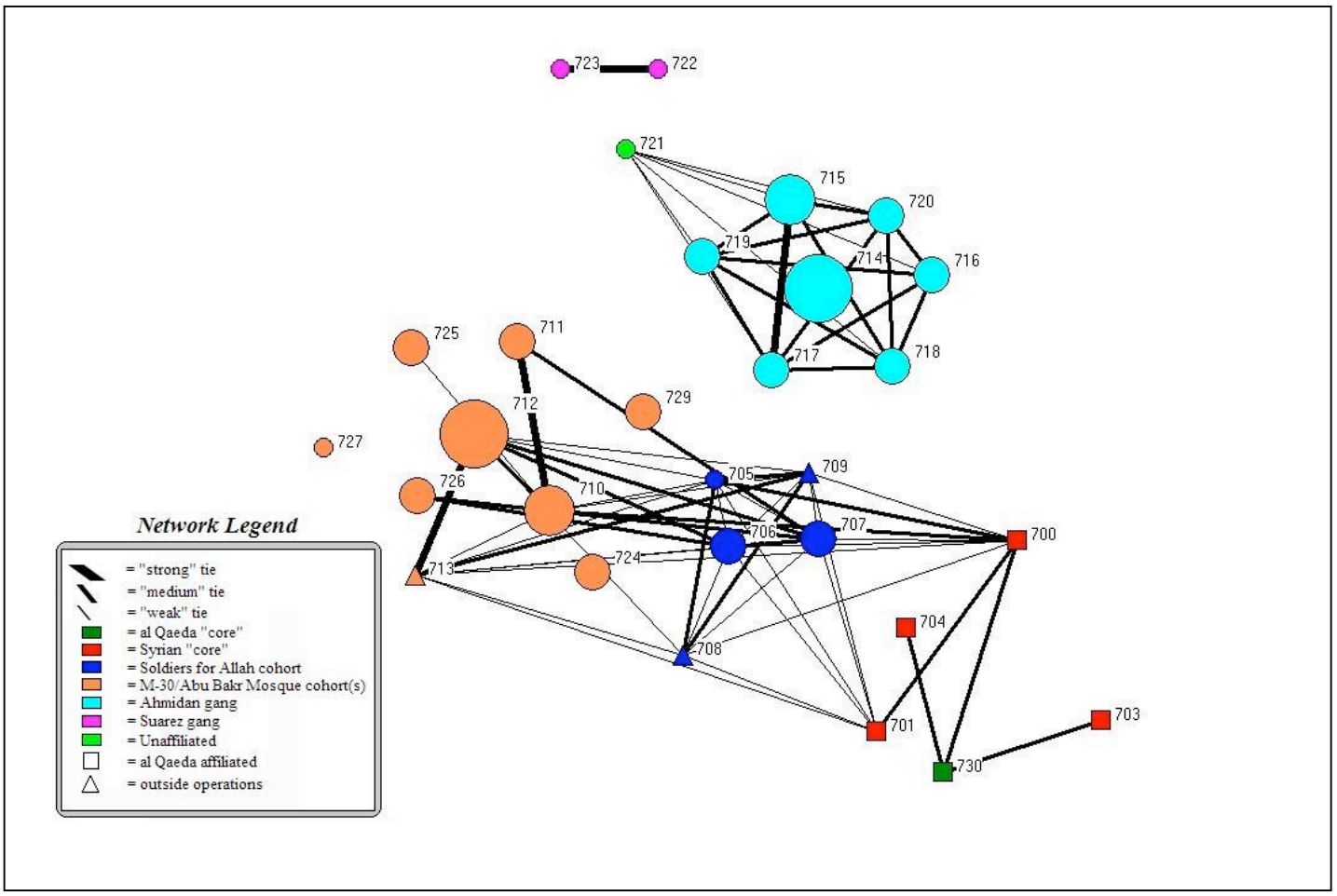

Fig. 4. A snapshot in 2002 of the evolving Madrid

A good example of an effective, disjoint small group is based on social network analysis of the Madrid bombings in 2004. We have recreated the evolution of this network since its early days in 1990. The final form involved the merger of two separate groups, roughly the idealists (Abu Bakr Mosque cohorts) with the Suarez gang members, who could provide the necessary resources. Each group had between 7 and

15 active members, developing separately until early 2003. By mid-2003, negotiations of mutual benefit to each were completed, with a merger under dual leadership. In its final form at the time of the bombing in March 2004, the network still had only about two dozen members. A snapshot of the pre-merge structure is shown in Fig. 4. Based on current data, this network is very typical of many other violent groups (e.g. Bali, Leeds, Netherlands.) Hence in this MURI, our theoretical models are currently focused on the development and fragility of networks of this size. (See Section 8.0 below.)

7.0 The Consistent Conformity Model (Page, Bednar)

We have been examining the interplay between two competing forces upon an individual's behavior when placed in a social setting: a desire for social conformity versus a desire to preserve individual consistency. Obviously, the interplay between these two forces will depend upon the social setting and the culture within which the group resides.

The general structure of our model is to assign attributes to agents. Each attribute has a value taken from a set of integers. Agents conform with one another when the value on an attribute matches that for the same attribute for another agent. An agent's desire to be consistent corresponds to each agent matching their value on one attribute a value (or values) on another of their attributes. Thus we have a tension between an agent achieving the same values on all of its own attributes, versus matching the attribute values of other agents in the community.

Let agents start with a random assignment of values. We then model the dynamics (and equilibrium conditions) for consistency or conformity alone. We can show that when individuals (agents) are driven only by the desire for consistency, the model produces consistent individuals, but no intra-culture homogeneity: i.e. a society consisting of individuals who exhibit behavior consistent only with themselves. In contrast, if individuals (agents) ignore consistency and choose instead only to conform to those around them, then all individuals converge to the same set of inconsistent attributes. Hence each force alone creates a simple dynamical system.

However, when the two forces combine, the resultant tension slows time to convergence. Imbalancing the dominance of one force over the other, surprisingly

can slow convergence further. Furthermore, if the model is expanded to include random attribute assignments (errors), a non-linearity in the system is revealed, and the equilibrium includes substantial heterogeneity. In other words, the two forces in conjunction magnify the effects of error, and tend to push the system away from equilibrium. These results should help to understand aspects of within-culture heterogeneity. (See Bednar et al, 2006.)

8.0 Leadership & Trust (Richards, Atran, Kasturirangan)

As mentioned in earlier sections of the report, we have been collating information about several non-violent and violent social networks in order to determine their typical structure and evolution. All of these networks have roughly thirty or fewer active members. For some (e.g. two cases of groups with educational objectives) there may be an additional thirty, but these members are not as heavily involved of committed, and typically have only transient roles. All of the networks examined (about a dozen) may have more than one leader, but not more than three, with each leader having a following that constitutes a subgroup. The size of the subgroups is usually less than ten, and all subgroups are connected, but more loosely than members within each group. The Itza' and Madrid networks illustrated earlier are specific examples that are within the range of what we consider to be a small group network.

Our aim is three-fold: (1) to determine whether violent and non-violent networks have similar structures, and the range of parameters that characterize them; (2) the evolution of these networks; and (3) a (theoretical) measure of their stability – how easily they might be shattered. Although data collection continues, our current assessment is that both violent and non-violent networks are very similar in their structure and evolution. A summary is in preparation.

In parallel to data collection and assessment, we are proceeding to develop a model for network stability. The important parameters we are using for network structure are based on our preliminary studies and include the number of bonds between members (about 50% connected) and members to leaders (about 70% connected), and leaders to leaders (100% connected.) Our measure of network stability is the likelihood that the current leader's proposals will be accepted by the majority. Such acceptance will be undermined if there is doubt about the current leader's abilities. The source of this doubt can be varied: lack of trust, a change in context (perhaps a new scenario),

conformity vs consistency pressures, influences arising from new enrollments, etc. When such concerns arise, the support for the current leader erodes.

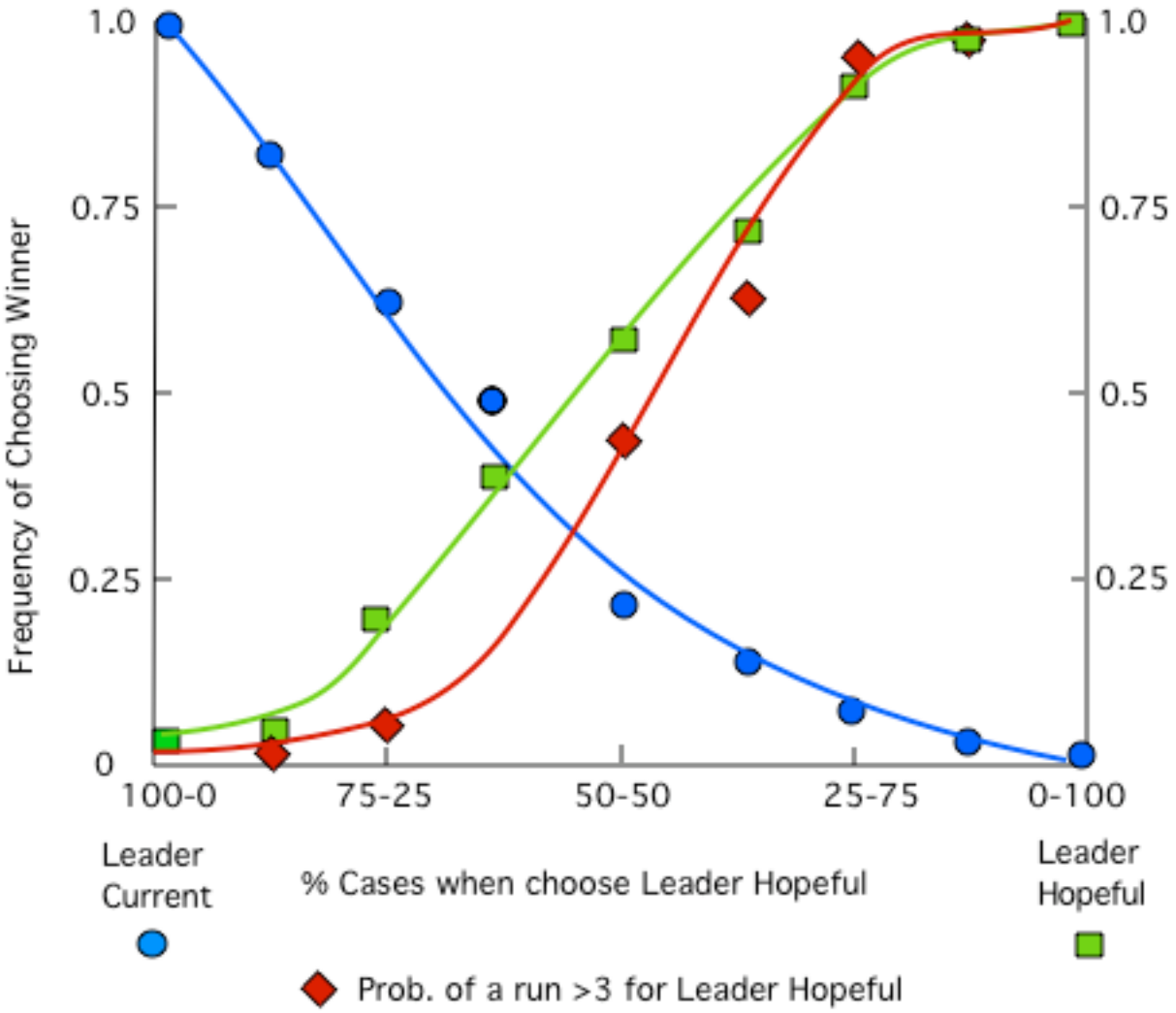

Fig. 5 Effect of lack of trust in Leadership on network stability. The ordinates show the likelihood that either the current leader (left, blue curve) or an alternative leader (right, green curve) will win a majority vote. The variable along the abscissa is the level of doubt about the current leader's abilities. The red curve shows the probability that there will be a sequence of three tallies in a row with the majority either stalemated, or favoring the alternate Leader_hopeful's proposals.

To simulate this scenario, we model each node in the network as an individual with a preference for a particular action or outcome. We assume that the greater the separation of individuals in the network, the more dissimilar their choices (See Richards

et al, 2002, Richards, 2005.) Using this distance metric, we can determine which of any two proposed alternatives an individual will favor. If this distance metric to the current leader is small (specifically one edge step in the network), the individual will support the leader, otherwise the individual voter will be indifferent. In this latter case, we assume that, with some probability representing the level of doubt, the individual will go along with an alternative proposal made by the "leader hopeful." The probability that an indifferent vote will be cast in favor of the "leader hopeful" is our independent variable.

Fig. 5 shows the general form of our result. Consider the most obvious case when the voter has an indifferent choice (i.e. in the network that voter lies more than one edge step from the leader). If the current leader is trusted (or respected, etc.), then the voter will always go with the leader's proposal. In this case, the current leader will win the vote almost 100% of the time, as shown at the left side of the figure. But what if that same indecisive voter has doubt about the current leader's abilities? For example, what if the level of doubt or lack of trust rises to about 50%? Then all such indecisive voters will favor the Leader_hopeful half the time. For the networks similar to those in Figs 2 & 4, the Leader_hopeful now wins consensus 60% of the time to 20% for the current leader. More disastrous for achieving consensus, the likelihood of a stalemate increases. This is implied by the red curve, with the odds for a sequence of three or more outcomes against the current leader's proposals rising to about 40% in 100 votes. Clearly, the current leader is losing control, in favor of another member equally regarded in the network. This is taken to be an instability in reaching a consensus. Stability returns only when the Leader_hopeful has become the new leader.

This form of the result illustrated in Fig. 5 is very insensitive to the choice of parameter values of the small networks we are examining. A corollary of this robustness is that removing a few edges or nodes (including the leader) will have little effect on our measure of network stability.

9.0 Ideal Observers & Strategic Play (Stankiewicz)

Our lab has been developing tasks and models in the area of Belief Updating using Bayesian updating algorithms. More recently, we have added the process of Trait Modification. To this end, we have been working on a task that involves adversarial decision making with uncertainty in which each player has the opportunity to invest their resources into processes that have immediate, medium-term or long-term benefits.

Furthermore, participants have the opportunity to modify how ``aggressive'' they are in addition to how ``trustworthy'' they are. On each turn, the player can modify these traits (Trait Modification), which modifies the actions that are available to them. Using this task, we are interested in two primary questions: First, what behavior leads to specific Trait Modifications (e.g., more aggressive to less aggressive) and second, how does the modification of the Trait lead to investing into processes that have different temporal payoffs. The goal of the game is to maximize personal payoff after a set period of interactions.

In addition to Trait Modification, we are also examining the belief updating process. Because the actions that are available to the participant are dependent on the Trait State of the player, one can use the observations generated by the opponent's actions to infer the opponent's current Trait State. To measure the performance of a player's belief updating strategy each player explicitly indicates what their Belief is about their opponent(s) Trait State (Aggressiveness and Trustworthiness). We are interested in understanding how well humans are able to infer these Traits and how well they use that information in this adversarial game.

Finally, in addition to understanding this task behaviorally, we wish to be able to model this behavior. Our current approach is based on a Bayesian model using Partially Observable Markov Decision Processes (POMDP.) This model will provide us with the ideal performance for these tasks, against which the players will be judged.

10.0 Language for Negotiation and Strategic Play (Pfeffer, Gal)

Individuals and groups involved in interactions with each other are often uncertain about each other's situations, goals and intentions. In these situations, values, norms, preferences and reasoning patterns attributed to cultural affiliation as well as individual personality traits affect people's behavior, as well as their interpretations of each other's actions. Atran et al. have found that people's cultural values affect their reasoning about the consequences of their actions on themselves and their adversaries.

Our work has focused on modeling and understanding the preferences and reasoning patterns of agents that underlie their behavior in strategic situations. We have investigated both game-theoretic models and extensions of game-theoretic models. As in some of the previous sections, our models apply to both human and computer agents,

but we are focusing particularly on human agents. Our work has proceeded in four directions:

1. Identifying the motivations and patterns of reasoning that lie behind agent's actions in strategic situations.
2. Modeling agent's reasoning about other agents in the face of partial information about other agents.
3. Modeling reciprocal behavior in ongoing interactions.
4. Extending the Colored Trails platform for experimental studies

10.1 Patterns of Reasoning (Pfeffer, Gal)

We have identified four different patterns of reasoning that lie behind an agent's decision making process. These are, briefly,

i) Taking an action because of its direct effect on the agent's outcome
ii) Taking an action that has an effect on another agent's outcome, in order to influence the second agent to respond by taking an action that is beneficial to the first agent.
iii) Taking an action in order to communicate information that the first agent knows to the second agent, in order to influence that agent.
iv) Taking an action in order to cause the second agent to find out something that the first agent does *not* know.

Each of these reasoning patterns has been identified with a diagrammatic pattern of paths in a Multi-Agent Influence Diagram, which is a graphical representation of strategic situations. The figure below shows four Multi-Agent Influence Diagrams, each illustrating one of the patterns of reasoning. Ellipses indicate chance variables, square nodes represent decisions made by agents, and diamond nodes represent outcomes. There are two agents named Alice and Bob, represented by A and B in the diagrams. In the first diagram, there is a direct path from Alice's decision to her outcome, so she will take an action because of its direct effect. In the remaining diagrams, there is no direct path from Alice's decision to her outcome, so the only way she can affect her outcome is to influence Bob's behavior. In the second diagram, Alice influences Bob by taking an action that directly influences Bob's outcome. In the third diagram, Alice influences Bob by communicating information about a chance variable C, which Bob cares about, thus changing Bob's behavior. The fourth pattern of reasoning is the most surprising. It turns out that Alice can cause Bob to discover something that Alice herself does not know, and thereby influence his behavior. Here,

D is something that Bob cares about. C depends on D, and C is known to Bob, so he possibly has some knowledge of D. But this knowledge is modulated by Alice's action's effect on C. Thus Alice influences Bob's knowledge of D.

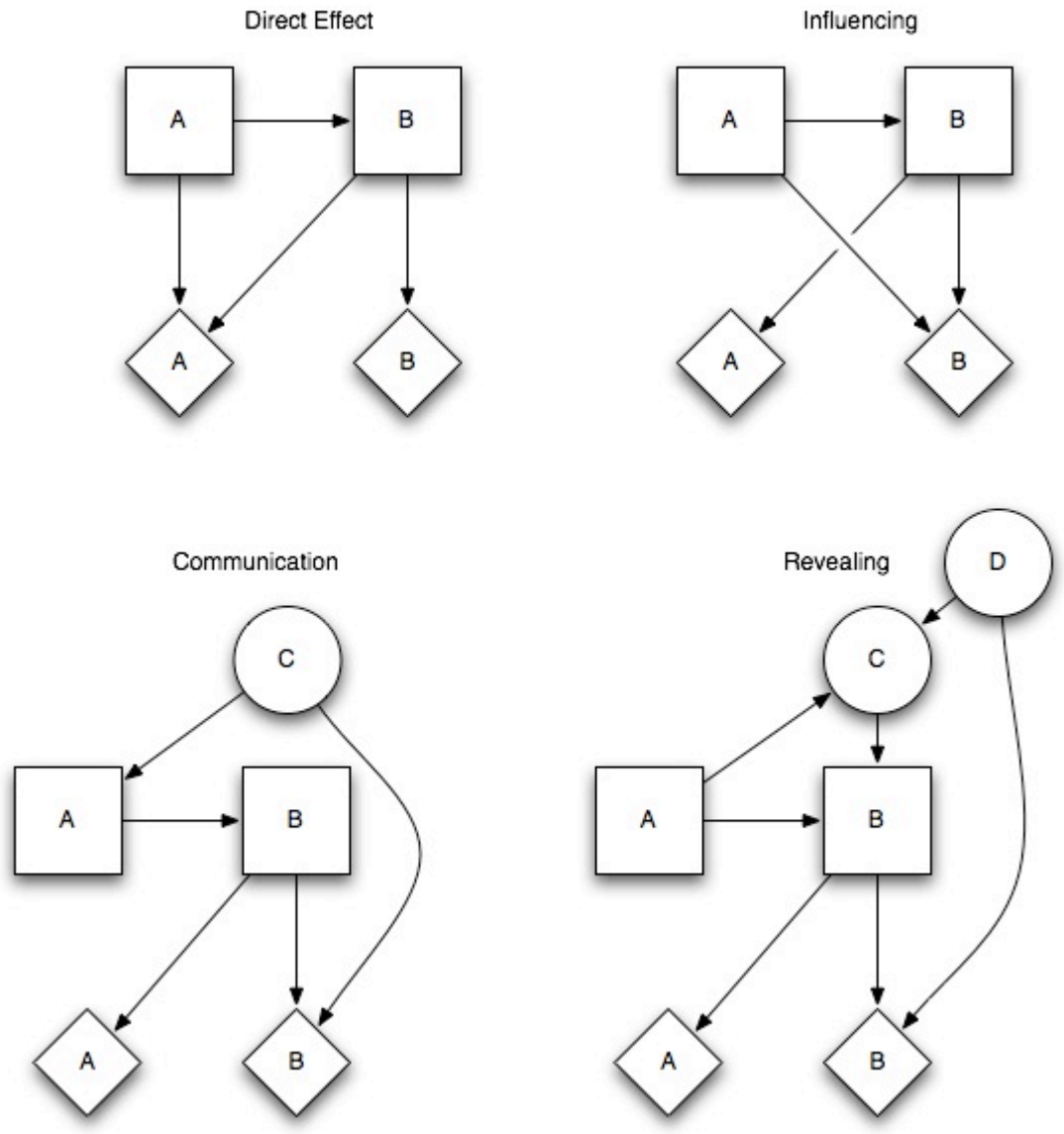

Fig. 6. Multi-Agent Influence Diagrams for the four reasoning patterns

We are currently trying to show that our categorization is complete, in that every situation in which an agent has reason to make a decision falls into one of our

categories. We are also investigating whether there are inherent advantages, either in representation or in inference, to prefer one form of reasoning over the other.

10.2 Partial Information (Pfeffer)

In many situations, agents will have uncertainty about other agents. This uncertainty may be about preferences of other agents, their goals, tasks, resources and reasoning patterns. We are investigating how agents reason strategically in the face of such uncertainty. In particular, we are focusing on the case where agents have partial but incomplete information about other agents.

We have formulated several hypotheses about how an agent, in particular a human agent, reasons about information about other agents. The simplest hypothesis is that it ignores this information and acts solely based upon its own situation. The next hypothesis is that it uses this information in some heuristic manner. A third hypothesis is that the first agent tries to reason about possible situations of the second agent, and tries to maximize its expected utility with respect to these possible situations. One can carry this further. A fourth hypothesis is the first agent reasons about a second agent's reflexively reasoning about the first agent.

We are currently conducting experiments to determine which of these hypotheses is best for people in a three-player negotiation game. In addition, for each of these hypotheses, we have designed learning algorithms to learn how best to play the game with people.

10.3 Modeling Reciprocal Behavior (Pfeffer, Gal)

In the past we have learned models of people's social preferences in single-shot negotiation games. We are currently investigating models for ongoing games. We have developed models that capture notions of reward, punishment and reciprocity. We have developed learning algorithms that learn the parameters of these models from data that we have collected about people's play. Crucial to our model is the tradeoff between the benefit to agents of their current actions and the future ramifications of their actions, brought about by the reciprocation of others. Computing the future ramifications for all actions is computationally infeasible. Our algorithm approximates this quantity by integrating a sample from a representative set of possible future interactions into the learning process.

We are currently conducting experiments to determine whether a player that uses our model with reciprocity is better than one that uses a standard notion such as Fairness Equilibrium (Rabin, 1999). The next step is to investigate whether people's cultural traits, identified by their college major or gender, affect their reciprocal behavior.

10.4 Colored Trails (Pfeffer, Ficici, Gal)

Colored trails is a game played on a board of colored squares with a set of chips in colors chosen from the same palette as the squares. Any square on the palette can be designated as the "goal square" and each agent has a piece on the board, initially located in one of the non-goal squares. At the onset of the game, agents have a set of colored chips. To move a piece into an adjacent square, an agent must turn in a chip of the same color as the square. Chips may be exchanged by the agents, and the conditions of exchange may be varied to model different decision-making situations. (See Gal et al, 2004 for a complete description.)

Although not obvious from this brief description, the Colored Trails game provides a very rich platform for studying decision-making. The chips represent resources which can be traded for one's own benefit, or not, depending upon the player's desire to be cooperate or not. However, unlike tit-for-tat, there are many possible exchanges, many types of play, and many strategies depending upon the particular configuration of the game and the context (two players, many players, or computer agents also engaged in play, single-stage or multi-stage, complete information or partial information, competitive, collaborative or team-based, and so on.) An alpha-release of the Colored Trails system has been made available under the GNU license. We are currently extending the Colored Trails system to support the needs of our experiments. These extensions will be part of future releases of the system.

11.0 Qualitative Reasoning and Story Models (Forbus et al; Winston & Finlayson)

Stories provide valuable information about a culture, such as its traditions and role models. Often, decisions are made by seeing present situations as analogous to the past, and one acts accordingly. Furthermore, sacred values, desired leadership or

individual characteristics and aspirations are passed on through generations through stories including the specific life experiences of participants, their shared texts, and informally shared group stories (aka "urban myths"). Our aim is to be able to convert stories to formal representations that make explicit the mental models and (causal) reasoning in the stories, thereby allowing cross-cultural comparison of reasoning and decision-making processes, as well as automatic extraction of cultural patterns. This collaborative enterprise is multi-layered, with components being built at Northwestern and MIT.

11.1 Story Workbench for Encoding Texts (Forbus, Tomai, Winston, Findlayson)

Much of the material that is used when exploring how decisions are made is textual. Unfortunately, today's natural language technology is nowhere close to the automatic conversion of text to a formal representation useful for story understanding. We believe we can significantly speed up the encoding process by a combination of the following techniques and resources that will be integrated into what we call *The Story Workbench*:

Controlled language. Controlled languages restrict the syntax and vocabulary of what can be stated, to reduce ambiguity and allow automatic processing. Instead of trying to formally encode stories directly in predicate calculus – a process that is known to be time-consuming and error-prone – storywriters will only have to restate them in simpler English. This is used extensively in industry: for example, "Caterpillar English" and "Xerox English" are controlled languages used by document writers at those companies so that a manual can be automatically translated into other natural languages. In prior work, we developed a controlled language, QRG-CE, which supports the creation of qualitative representations from a subset of English. Unlike typical controlled languages, the vocabulary of QRG-CE is not fixed, since acquiring new knowledge via natural language interaction is one of its goals.

Large-scale representation resources. In prior work we were able to use the COMLEX lexicon to provide broad linguistic coverage for syntactic parsing, and the ResearchCyc KB to provide a rich representation vocabulary for semantic processing. We will also leverage the numerous resources that have been developed for statistical syntactic or semantic tagging, for example, statistical word-sense tagging algorithms, or information extraction techniques for agents, events, and dates; work is ongoing to add these capabilities to the Story Workbench.

Interactive feedback. Because controlled languages are often difficult to write without error, and knowing the boundaries of the controlled language can be difficult (especially if it is being revised as the data changes and as new vocabulary is added), we plan to have the Story Workbench automatically check the syntax of the text being entered, and graphically display to the storywriter the formal descriptions it has constructed. Based on this feedback, the storywriter can then modify the text as it is being written to achieve the correct result, rather than relying on a time-consuming write-compile-correct cycle. This kind of interface generally draws on the time-saving tradition of Integrated Development Environments (IDEs) for programming languages, and follow the work of other research groups such as MITRE's Alembic Workbench and Sheffield's GATE.

The key idea is to reduce the barrier to assembling a large database of stories that are computer-readable. Given such a database, story modeling will be facilitated, and in turn, so will the comparisons of traditions and cultural factors depicted by stories. Not only will we be better able to enter formal representations (using a controlled language and real-time interactive feedback), but the amount of rote work a user must perform will be reduced (by leveraging the improvements in statistical syntactic and semantic tagging and the large databases of formally represented knowledge.)

We are building on QRG for our controlled language. Based on a corpus of stories from different cultures that we have collected, we are extending the grammar of this language. In addition, the ResearchCyc representations are being augmented to include the necessary narrative structures. At present, the basic framework of the workbench has been completed, using our Eclipse IDE development platform. The process of assembling the statistical tagging algorithms is currently underway, as well as the design and implementation of interface components for user feedback. Shortly we should also have an interface to the Cyc knowledge base and to the Northwestern EA parser for semantic tagging.

11.2 Automatic Derivation of Cultural patterns (Finlayson &Winston)

Given a large set of stories formally coded by means of the Story Workbench, we will be in a position to discover regularities in cultural story sets. One possible scheme is shown in Fig. 7. The idea is that stories induce a set of story "fragments" which span a space of stories. The appropriate set of such fragments can be recovered using known clustering or statistical learning techniques (see Finlayson & Winston, 2005.)

These fragments can then be pieced together, using analogical mapping, to produce a story consistent with the culture from which the basis set was drawn. This requires an analogical mapping tool to determine how well the stories are covered by the assembled fragments, namely the episodic sets. We can ascertain the effectiveness of this process by making predictions about accuracy of retrieval or probability of re-description for particular stories in human subjects. Our linking hypothesis is that if some piece of story is well "covered" by a piece of the story fragment set, then that piece of the story will not only be better remembered, but more likely to remembered without alteration, as contrasted with another piece of the story set that is less well covered, or not covered at all.

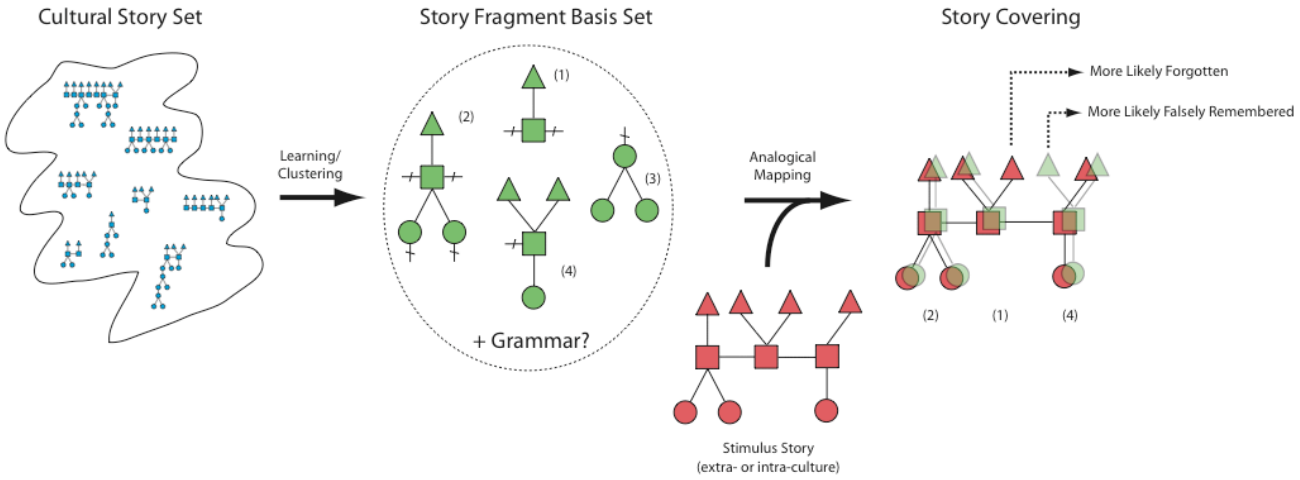

Fig. 7. Story Lines and Workbench mockup

11.3 Broader Relevance of Understanding Cultural Stories

Myths, Folk tales, ballards, stories about heros – and many other forms of stories provide perhaps the most important medium for passing on to future generations the traditions, moral values and role models for a culture. Stories reveal actions and strategies that led to past successes (and failures) against adversity. To mine this wealth of data, we need representational forms and computational models for acts, actors, events and adversities, showing the form of the interactions between these elements of stories. Story Lines and the associated Workbench are constructs that we hope will allow us to place stories in a computational "normal form" that not only

models the stories, but better reveals the important differences between the cultures they depict.

12.0 Conclusion

An important vehicle for advancing science is to bring together researchers from different disciplines, each with different skills and perspectives on a problem of interest to all. In the first 18 months, considerable progress has been made on this front, with roughly six such active collaborations in place. These fell into place at our annual meeting in Jan 06. Students and postdoctoral appointments reinforced these joint efforts of the PI's. We hope still one or two more of such collaborations will be sparked by the forthcoming meeting in Jan 07. A further objective of the meeting will be to probe for lacunae where more research activity is needed. Finally the meeting will also give us the opportunity to review among ourselves the overall thrust of the effort, and how the various projects can be woven together more tightly to create a truly unique, integrated endeavor.

Additional Collaborators:

We now have an affiliation with the John Jay Center on Terrorism, where Dr. Atran has an appointment. This begins 1 Nov 06.

Dr. Rajesh Kasturirangan, National Institute of Advanced Studies, Indian Institute of Science, is contributing on the structure, evolution and stability of networks.

Dr. Marc Sageman continues on several of Scott Atran's projects. Also involved are Dr. Reuven Paz and Dr. Khalil Shikaki as consultants on network and PV studies.

Dr. Craig Joseph has recently joined the Medin research group as a postdoctoral fellow. Dr. Joseph has worked closely with the Muslim community in south Chicago (Bridgeview) and will be conducting interviews and other probes aimed at exploring how notions of identity (e.g. what it means to be Muslim and what it means to be an American) foster or undermine alienation.

Dr. Ya'akov Gal has now joined the MURI as a postdoctoral fellow, with joint appointments at MIT and Harvard, collaborating with Pfeffer, Tenenbaum & Richards.

We have also explored or have been engaged in collaborations with Barry Silverman, Univ. Penn.; Alex Pentland, Media Lab, MIT; and IndaSea (specifically, the analysis of alignments among 13 parties preceding and following the Bali bombing.)

Personnel Changes:

See above. Note especially the John Jay collaboration, which may lead to some personnel changes in Fiscal 07.

Personnel Supported (does not include administrative support)**:**

Partial summer salaries for all Principal Investigators.
Partial support for Dr. R. Kasturirangan, Res. Assoc. MIT (now at National Institute for Advanced Studies, Bangalore, India.)

Consultants: Prof. Robert Axelrod, UMich; Dr. Marc Sageman; Dr. Reuven Paz, Israel; Dr. Khalil Shikaki, Brandeis, Dr. Ken Ward.

Postdoctoral Fellows: Y. Gal, MIT/Harvard; W. Bennis, Northwestern

Graduate Students: (at least partial support)
UMich: L. Alattar, A. Bramson, N. Ismail, J. Miller, J. Magouirk, D. Wright
Northwestern: D. Bartels, M. Dehghani, R. Iliev, E. Tomai
MIT: C. Kemp, C. Lawson, G. Pickard, M. Finlayson
Harvard: Y. Gal
Univ Tx: K. Eastman

Undergraduate students: These are hired to assist graduate students.
MIT: L. Kitch, M. Steiffer

Budget:

On track. However, IRB problems at University of Michigan will result in Dr. Atran's subcontract to be moved to John Jay Center on Terrorism. This will not change the budget allocations.

Publications (in preparation or in press.)
(most papers may be retrieved from http://groups.csail.mit.edu/belief-dynamics/)

Atran, S, Medin, D. and Ross, N. (2005) Psychol. Rev. 112, 744 –776.

Bartels, D. M. & Medin, D. L. (in press). Are morally-motivated decision makers insensitive to the consequences of their choices? *Psychological Science*

Bartels, D. M. (2006, November) *Morally-Motivated Judgment and Decision Making*. Symposium organized and to be presented at the meeting of the Society for Judgment and Decision Making, Houston, Texas.

Bartels, D. M. & Bennis, W. M. (2006, November) *Deontology and Consequentialism in Morally-Motivated Decision Making*. Paper to be presented at the meeting of the Society for Judgment and Decision Making, Houston, Texas.

Bartels, D. M. & Medin, D. L. (2006, November) *Are Morally-Motivated Decision Makers Insensitive to the Consequences of their Choices?* Poster to be presented at the meeting of the Psychonomic Society, Houston, Texas.

Bednar, J, Bramson, A. Jones-Rooy, A., and Page, S. (2006) Conformity, Consistency and Cultural Heterogeneity. (submitted.)

Bednar, J. and Page, S. (2006) Can Game(s) Theory explain culture? The emergence of cultural behavior within Multiple games. (*Rationality and Society*, in press.)

Forbus, Invited Keynote, 20th International Qualitative Reasoning Workshop, Dartmouth, NH. Title: Qualitative Representations as a Modeling Language for Cognitive Science, July 2006

Finlayson, M A and Winston, P. H. (2005) Intermediate features and informational-level constraint on analogical retrieval. In B. G. Bara, L. Barsalou, & M. Bucciarelli (edds.). *Annual meeting of the cognitive science society* (Vol. 27) Stressa, Italy.

Forbus, Analogy Workshop, University of Osnabrueck. Title: Qualitative Reasoning Group Analogy Research Overview. October, 2006

Forbus, K., Lockwood, K., Tomai, E., Dehghani, M., and Czyz, J. (submitted). Machine Reading as a Cognitive Science Research Instrument. Submitted to the 2007 AAAI Spring Symposium on Machine Reading.

Gal, Y. Modeling the influence of Task Contexts on Human negotiation. (submitted.)

Ginges, J., Atran, S., and Medin, D. Moral Barriers to the Rational Resolution of Conflict. (submitted to Nature).

Ginges, J, Atran, S., and Medin, D. Sacred Bounds on rational conflict resolution: the Middle East and Beyond. (submitted.)

Ginges, J. (2006, November) *Moral and Instrumental Values in Political Negotiation: Field Studies with Israelis and Palestinians*. Paper to be presented at the meeting of the Society for Judgment and Decision Making, Houston, Texas.

Iliev, R. & Medin, D. (to be submitted to Organizational Behavior and Human Decision Processes). Moral Values and Attention: Anchoring effects and the conjunction fallacy.

Iliev, R., Bartels, D. M., Sachdeva, S., & Medin, D. L. (2006, November) *Cognitive Processing of Morally Relevant Tasks*. Poster to be presented at the meeting of the Society for Judgment and Decision Making, Houston, Texas.

Kemp, C. Tenenbaum, J. B., Griffiths, T.L., Yamada, T. and Ueda, N. (2006) Learning systems of concepts with an infinite relational model. *Proc. 21st National Conference on Artiificial Intelligence* (AAAI-06.)

Page et al (2006) Learning the commons. (in preparation.)

Richards, W., Kasturirangan, R. and Atran, S. Modeling stability in small social networks. (in preparation.)

Richards, W. (2005) Collective choice with uncertain domain models. MIT-CSAIL technical report # AIM-2005-054 (available on-line.)

Stankiewicz , B.J. (Under Review). Understanding the cognitive limitations when making efficient decisions under uncertainty. United States Army Technical Report.

Stankiewicz, B.J. & Pitts, J. (Under review) Remembering How You Got Here: The Role of Path Memory and Landmarks During Localization. Journal of Experimental Psychology: Learning, Memory & Cognition.

Stankiewicz, B.J. & Eastman, K. (Under review) Lost in Virtual Space II: The Role of Proprioception and Discrete Actions when Navigating with Uncertainty. Association for Computing Machinery/ Transactions on Applied Perception.

Stankiewicz, B.J., Cassandra, A.R., McCabe. M.R. and Weathers, W. (In Press). Development and Evaluation Of A Bayesian Low-Vision Navigation Aid. Under Review. Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics.

Tenenbaum, J. B., Griffiths, T. L. and Kemp, C. (2006) Theory based baysian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7) 682-687.