Socially Intelligent Agents to support Ethical Decision-Making

Anita Raja[1], Catriona Kennedy[2], and Roger Hurwitz[3]

[1] Department of Software and Information Systems, The University of North Carolina at Charlotte, Charlotte, NC 28223, anraja@uncc.edu
[2] School of Community-Based Medicine, University of Manchester, UK, catm.kennedy@gmail.com
[3] Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, rhhu@csail.mit.edu

Abstract. Recent projects on computational support for public policy deliberations, e.g., automatic classification of comments to rule-making agencies, have addressed some needs of ethical decision-making. They achieve this by expanding the values considered and the outcomes of a decision for different groups. Since their aim, however, has not been the systematic inclusion of the relevant interests, values and consequences in policy deliberations, computer supported decision-making in such environments may yet be narrow, short-sighted and unethical. To address this problem, we are developing a reusable means of bringing the viewpoints of power-limited elements of the population and usually overlooked dimensions of value into the public policy decision space. Our multi-agent approach involves designing value-aware BDI agents with conflict resolution capabilities while harnessing human-machine complementarity. We claim that new kinds of policy innovation may result when humans and agents act together.

1 Introduction

Models and simulations developed in the social sciences are frequently used to inform policy-making. These may range in complexity from the two person game theoretic models that are a basis for nuclear deterrence strategies to econometric models for forecasting the national economy and supercomputer simulations of flu transmission that can guide public health officials [1]. However, many such models, especially those dealing with the effects of a policy on the welfare of individuals, may fail to take into account the values and concerns of some people whom the policies might affect. For example, tests and models of drug efficacy may predict that a drug does not sufficiently improve the treatment of a disease to warrant its licensing by the Food and Drug Administration (FDA) for sale to the public. Some patients and their families will likely object that such models use crude numerical measures or do not consider the possibilities of helping some cases like theirs. Complaints like these have been quite noticeable with regard to cancer drug candidates, e.g., laetrile. This general case may be classified as an ethical problem since it raises issues of justice and fairness.

Our response to this problem is to develop representations and algorithms that incorporate into decision-making contexts the **multiple, differential perspectives** arising from the different life experiences and needs of diverse population groups. We address the challenge of under-represented groups by defining ethical agents that represent values on two different levels in accordance with principles for public participation and value advocacy [2]. These are ”consensus values” that are held by all participants and ”stakeholder values” that are only held by some. Consensus values include concern for process and content. Furthermore, as the number of ethical concerns (or social values) increases, people are more likely to differ in their evaluations and objections to policy proposals. Entities representing these concerns should be able to revise initial proposals, formulate new ones or introduce new rules (other than utility maximization) for guiding decisions through **conflict detection and communication**. This will allow the model to go well beyond a utility maximization model that comprehends all the values held by the affected populations, reduces these values to measures on a single scale of utility and selects the proposal that maximizes the aggregated utility.

In this paper, we present our approach towards developing a multiagent system to help decision-makers and researchers evaluate policies. Our development of the ethical agent is also guided by the notion that an agent can object to policy proposals if it can show that under-represented (limited power) groups will be treated unfairly. It is composed of the following objectives: **Value-aware agent-directed analysis and simulation**; and **Human-machine complementarity**. We describe our first steps towards a multiagent toolkit based on the Belief-Desire-Intentions (BDI) [3] framework, a popular model for autonomous agents, to explore representations of human values and ethical decision making so that agents can use these as parameters to control simulations and associated data analysis. We emphasize the human-machine complimentarity because we believe that software agents can provide surprising results because of values that are under-represented and may be hidden. The humans, on the other hand, could use negotiation and visualization to understand the complexity of the problem and intervene to correct the agents if necessary. Newly discovered values and concerns can be integrated into an existing knowledge representation, helping with both agent and human learning. We hypothesize that new kinds of policy innovation may result when humans and agents act together. Our longer term aim is to find ways of involving under-represented groups directly in the interaction.

We plan to use content analysis and machine learning methods on publicly available documents to discover new policy concerns as well as to learn different decision rules depending on the context, for instance, maximizing fairness instead of maximizing utility. We model the reconciliation of conflicts between policy objectives arising from the different values as an intention reconsideration problem [4] in BDI agents. We are currently designing decision-theoretic decentralized algorithms for intention reconsideration and leveraging a suite of agent negotiation techniques including argumentation in software agents to resolve conflicts. The novelty of our approach is in the agent control of modeling, simulation and data analysis in a way that is ”value-aware”.

The remainder of this paper is organized as follows: In Section 2, we motivate our work using an urban renewal example. We then present related work and discuss background concepts. In Section 4, we discuss our ongoing work for developing value-aware ethical agents that harness human-machine complementarity.

2 Motivating Example

We sketch here how a decision regarding urban renewal could be handled in the system we are developing. Suppose in a particular geographical region, there are a large number of residents in substandard housing who cannot afford to move out to alternate housing. There is also the concern that if new housing is built in the region, the loss of green space can diminish the quality of life (health) of existing residents in the region. Our motivating question is what type of local government intervention should handle this situation. We assume the following two policy proposals: Proposal 1 is to ”Build New Affordable homes”; the associated action is ”Build more houses and move residents to new housing. But this requires building on green space”; the identified costs are ”Loss of green space; construction costs; disposing unsafe homes costs”. Proposal 2 is to ” Demolish and Regenerate”; the associated action is ”Demolish existing houses and build new ones”; the identified costs are ”temporary relocation costs; construction costs; disposing unsafe homes costs”.

Agents A1 and A2 act on behalf of P1 and P2 respectively. Agents can test policies by running simulations in a virtual world using UrbanSim [5]. The simulated world is a geographical grid divided into regions and neighborhoods. Each region will have associated models for economic levels, health, demographic change etc. We assume that both agents have the same background knowledge about the housing scenario, which is expressed using the BDI formalism. This includes the consensus values in the housing scenario, which we assume to be health, safety, affordability, occupancy rate and total costs. Agents also have access to the same simulation models and supplementary data. However, agents will focus attention on different variables because they are acting on behalf of values corresponding to their respective policy proposals. A1’s main concern is the experience of the resident (e.g. they don’t want temporary relocation). A2’s concern is green space. For each agent, the flow of control takes the following form:

Step 1: Agent A1 recommends a method to test its preferred policy in a virtual world (using the UrbanSim interface). This includes the specification of an initial state and the execution of a model to calculate the effect of making the proposed changes. The parameters for visualization and data generation from the model should include the agent’s main concern, but should also include the other consensus values. Each of these recommendations can be overridden by a human user. The user also has the opportunity to give reasons that can be added to the agent’s beliefs.

Step 2: The new data generated from the simulation run is also shared with A2. Both agents interpret the data according to their respective concerns. Data analysis algorithms may be used to summarize the data so that logical statements are produced [6]. Although A1 may be satisfied with the result, A2 will detect that too much green space will be used if this policy goes ahead. A2 raises its objection by showing (using inference) that conservation of green space is an example of an environmental conservation principle (consensus value), thus drawing A1’s attention to this problem. Visualization also allows human users to raise objections and make counter-proposals at this point.

Step 3: A2 recommends a way to test its counter-proposal (repeat steps 1 and 2 for Policy 2). But A1 is dissatisfied with the result of this.

Step 4: Agents agree that the conflict is about two different costs (green space and resident experience). Detection of conflicts and determining the nature of the conflict is done by humans initially. Once the causes of the conflict have been identified, the negotiation can be focused on the parameters that are conflicting.

Step 5: Each agent will pursue its local intentions while adhering to the globally accepted norms for policy negotiation and action enforcement. For example, the new policy may be proposed where residents are moved out in a staggered fashion. Intention reconsideration will use a Decentralized Markov Decision Processes (DEC- MDP)-based mechanism [7] while conflict resolution will be achieved using negotiation. Once an agreement about a policy is achieved, the new policy is tested using steps 1 and 2.

3 Related Works

UrbanSim [5] is a simulation toolkit to assist with urban planning. It is composed of different models, which can interact together in the simulation environment. The models include household and business location choice models, economic models and demographic change over time. Models can be used to predict the outcome of proposed policies. Such policy proposals are called ”scenarios” in UrbanSim. The UrbanSim GUI allows users to select datasets and models, choose scenarios and explore results.

One of the key problems in the design of BDI agents is the selection of an intention reconsideration policy. A BDI agent uses such a policy to determine the circumstances under which it will expend computational resources deliberating over its intentions. Simari and Parsons [4] show that intention plans in a single BDI agent can be mapped to a MDP policy and vice-versa.

In recent work [8], we have addressed two related issues in the context of problems modeled as DEC-MDPs: a) how to handle a decentralized learning situation in which there is a very large search space for each agent? and b) how to resolve conflicts among the learned policies of different agents? We study these problems in the context of a tornado tracking application where each agent controls a set of radars and the goal is to maximize the overall utility for a given configuration of radars. We use a multiagent reinforcement learning algorithm to learn stochastic policies for the DEC-MDP and a decentralized negotiation mechanism to resolve the conflicts among agent policies in a partially global perspective. During learning, instead of starting with an MDP that contains all the possible states, each agent unrolls its search space. The space is selectively expanded and explored based on the conflict resolution performance in the agent's neighborhood (set of agents that have frequent interactions).

We plan to use decision theoretic methods to improve BDI reconsideration for value-aware contexts in the ethical agents described in this paper. These methods will also serve as a basis for conflict resolution. Emele et al. [9] combine argumentation, machine learning and decision theory to learn underlying social characteristics (e.g. policies/norms) of others and exploit the models learned to reduce communication overhead and improve strategic outcomes. We plan to build on this work as part of our negotiation approach by using methods that build better models of other agents. Also, Kamar et al. [10] introduce a decision-theoretic formalism for deciding whether to help other agents in collaborative planning with partial information. A key component is a probabilistic recipe tree, an efficient representation of the other agent's beliefs, which they use as a control mechanism to avoid combinatoric explosion. This leads to the effect that in negotiations or even pre-negotiations only those values or reasons that could change believed probability of an agent choosing one proposal over the other should be advanced by another agent. In our context, this means that announced opposition to a plan can be understood as helpful information and an agent who wants to give a second agent a reason to support it's proposal might use collaborative filtering on comments.

In the AIMSS project [11], we developed a proof-of-concept software agent that controls and interprets a simulation and checks if the simulation predictions are in agreement with reality. As an example case study, we used the housing scenario described in the earlier motivating example. In the proof-of-concept, datasets are generated from the simulation and from the real world observation data. The real world dataset is simplified as it represents selected features from the large and noisy observation data that can be compared with the simulation dataset. We applied Association Rule Mining to both datasets to produce two sets of generalized logical statements. We then applied logical satisfiability checking to look for inconsistencies between them. One limitation of this work is that the simulation represents a single view of reality, which we are addressing in the current work.

Gutmann and Thompson [12] propose a principled process for stakeholders to deliberate about proposals despite disagreement. The procedure is based on three values that should be observed in the deliberations and two principles regarding content of proposals. Our approach will harness this procedure. The deliberation is based on three procedural principles that guide the process and two content principles that guide content disagreements; The procedural principles are reciprocity, publicity and accountability. Reciprocity means people are offering reasons for their positions that have some possibility of being accepted by other people, who are also offering reasons for their positions. Publicity means that facts and relevant information of the matter need to be disclosed; accountability means people, such as officials, will follow through on commitments. The content principles are liberty and opportunity, where liberty means no decision should take away rights held by individuals and opportunity means the decisions should move in the direction of offering greater and fairer opportunities to the less advantaged people in the society if it increases opportunities for anyone in the society. None of these principles or rules are considered inflexible – what they specify depends on situation and the strength of their application can also depend on situation.

4 Work in Progress

Currently, we are harnessing our experience in housing simulations, while adding value-sensitive design principles to UrbanSim parameters. We will start with simple policy scenarios, where the principal concern is assuring that values of those affected by a public policy decision are represented among the options being discussed. We shall then move to more complex scenarios of emergent conflict among value-based options and their consequences. We will also develop an ontology for agent beliefs, to be used for BDI agent reasoning, control of simulation and data interpretation. We are addressing the following challenges.

Automated Content Extraction: Initially, agents are given a specification of values they will focus on (for example green space). In subsequent iterations, agents may use content analysis algorithms to discover values in arguments proposing policies. Content analysis can also be used to bolster an agent's rationale on why another agent should accept its values. In order to determine other policies and viewpoints about the urban renewal problem, we plan to apply unsupervised learning techniques to publicly available comment archives on urban renewal to extract previously unknown values, for e.g., moving people even temporarily out of the region could lead to job losses and thereby affect occupancy rates: Jobs would then be a new value to learn for this situation. These alternate policies and viewpoints are publicly accessible in online archives. Researchers [6] have demonstrated how online archives can be effectively classified and mined for sentiment and beliefs. Of particular interest for our purposes is Regulations.gov which hosts electronic dockets for eight federal regulatory agencies, including EPA, and has comment archives regarding issues of land use and environmental regulation similar to scenarios for UrbanSim. We plan to mine them for the range of values that citizens bring to bear in evaluating proposed regulations in these matters to assure the adequacy of our software agents in this respect.

Simulation Control: Unlike agent-based simulation, where agents inhabit a simulated world, the software agents in this case can use simulation to support argument. The agents interact with the real world since they also analyze real world data. During a simulation, the agent interprets machine-readable data corresponding to a series of simulation snapshots (screens) so that the agent and the humans are ”looking at” the same predicted states. Agents will interact with the simulation engine of UrbanSim so that they can query the simulation results at specified intervals. For example: every 50 cycles, an agent can query the number of houses built on green space and the number of residents still in unsatisfactory housing. Alternatively it may simply dump the current values of all variables in a given snapshot. This information is then shared with other agents and used in the negotiations (as in the motivating example). At any time, the user may interrupt the process and modify the simulation options, and may also pass this new information to the agent. For example, a user may decide that they are also interested in jobs in a neighborhood. This would modify the agents concern (value) representation and constrain its use of analysis algorithms.

Human-Machine Complementarity: The agents that use simulation control, content analysis and conflict resolution have the potential to reveal new values and policy options, providing a learning experience for human users. On the other hand, humans have the opportunity to intervene and provide corrective feedback if an agent does not accurately represent a value, or if there are problems due the lack of ”common sense” of agents. We expect that humans and agents will jointly propose policies and direct the configuration of simulations and content analysis. The precise level of autonomy requires experimentation. In Step 1 of the example, the agent initially recommends an UrbanSim configuration to a user but the user also has the opportunity to configure models and scenarios manually and pass the configuration to the agent.

Automated Conflict Detection and Resolution: In the initial phase of our work, human users determine what type of conflict exists and direct the conflict resolution process accordingly. In later phases we plan to introduce autonomous conflict detection. In general, an autonomous ethical agent might favor a proposal because it enhances the values it supports; it then realizes that its current proposal might not be adopted because other agents have different values or weightings of the same values. Having discovered this conflict, the agent revises its intention, either by producing a new proposal or favor another agent's proposal that has less opposition. It can be supported in this process by the reasons provided by another agent on why other value(s) might be important. This results in the other agent's values being added to the first agent's values for its evaluation of scenarios, constituting a change in agent preferences that arise in negotiations.

The globally accepted norms (consensus values) will be based on Guttmann and Thompson's [12] well-known process for deliberative democracy described earlier. When local intentions of agents conflict with each other, we will harness both intention reconsideration and conflict resolution algorithms. Intention reconsideration in the BDI agents requires planning and coordination in uncertain, dynamic environments. Online policy learning can be computationally challenging. We plan to harness the decentralized decision theoretic and machine learning techniques developed in our previous work [8, 13] to focus dialogue to pertinent values and also help determine the most persuasive arguments. We will extend the algorithms as well as introduce ontologies to serve as a basis of building better models of the other agents to guide argumentation of ethical agents. The conflict resolution process via argumentation will help guide the agents' intention reconsideration process so that new values can be learned. The reasons, provided in an argument on why a value is important in considering a proposal, can result in an agent adding the values of the opposing agent in its evaluation of scenarios. This constitutes a change in agent preferences that arise in negotiations. It is not included in Friedman et al's [2] more utilitarian approach, which adds modeling to include new values but does not include changes in values in the respective agents' BDI repertoires. This is an advantage of our decentralized autonomous system in contrast to Friedman et al.'s more centralized but open system. As noted earlier, content analysis can be used here to bolster an agent's argument on why its values should be accepted by the second agent.

Evaluation: The basic metrics will be the adequacy and accuracy of our value classifiers, as compared to human coding, on sample comments, and the ability of a software agent to score a proposal according to its value set, compared to a humans doing the same. In addition, we will measure agents capabilities to recognize emergent conflicts among their intentions at the collective level, their revision of intentions in the face of such conflicts or other obstacles, and the generation of new proposals or decision rules to resolve the conflict. It should be noted that decisions the agents reach will be constrained by precepts that no agent should be deprived of a right currently held and outcomes should favor the least advantaged. So, we shall also compare the respective distribution of benefits of decisions so constrained with those possible according to utility maximization.

5 Acknowledgement

This work is supported partially by the National Science Foundation (NSF) under Agreement No. IIS-1018067. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

References

1. Stead, W.W., Lin, H.S.: Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. The National Academies Press Washington, D.C. (2009)
2. Friedman, B., Borning, A., Davis, J., Gill, B., Kahn, P., Jr., T.K., Lin, P.: Laying the foundations for public participation and value advocacy: interaction design for a large scale urban simulation. In: Proceedings of the 2008 international conference on Digital government research. Digital Government Society of North America). (2008) 305–314
3. Rao, A., Georgeff, M.P.: Bdi-agents: From theory to practice. In: Proceedings of the First International Conference on Multi-agent Systems (ICMAS). (1995) 312–319
4. Simari, G.I., Parsons, S.: On the relationship between mdps and the bdi architecture. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems (AAMAS '06). ACM, New York, NY, USA,. (2006) 1041–1048
5. Waddell, P.: Urbansim: Modeling urban development for land use, transportation and environmental planning. **68** (2002) 297–314
6. McIntosh, W.: The digital docket project: Computer assisted textual data analysis of the scotus corpus
7. Bernstein, D.S., Givan, R., Immerman, N., Zilberstein, S.: The complexity of decentralized control of markov decision processes. Math. Oper. Res. **27** (November 2002) 819–840
8. Cheng, S., Raja, A., Lesser, V.: Multiagent meta-level control for radar coordination. Journal of Web Intelligence and Agent Systems (WIAS) (to appear) (2012)
9. Emele, C.D., Norman, T.J., Parsons, S.: Argumentation strategies for plan resourcing. In: The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3. AAMAS '11 (2011) 913–920
10. Kamar, E., Gal, Y., Grosz, B.J.: Incorporating helpful behavior into collaborative planning. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2. AAMAS '09 (2009) 875–882
11. Kennedy, C., Theodoropoulos, G., Ferrari, E., Lee, P., Skelcher, C.: Towards an Automated Approach to Dynamic Interpretation of Simulations. In: Proceedings of Asia Modelling Symposium 2007, in conjunction with Thailands 11th Annual National Symposium on Computational Science and Engineering). (2007)
12. Gutmann, A., Thompson, D.: Democracy and disagreement. Cambridge, MA: Harvard University Pres (1996)
13. Cheng, S., Raja, A., Lesser, V.: Multiagent Meta-level Control for a Network of Weather Radars. In: Proceedings of 2010 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT-2010), Toronto, Canada (2010) 157–164