Designing Ethical Personal Agents

Nirav Ajmeri
North Carolina State University

Hui Guo
North Carolina State University

Pradeep K. Murukannaiah
Rochester Institute of Technology

Munindar P. Singh
North Carolina State University

Editor:
Munindar P. Singh
m.singh@ieee.org

The authors consider the problem of engineering ethical personal agents. Such an agent would understand the applicable social norms and its users' preferences among values. It would act or recommend actions that promote preferred values, especially, in scenarios where the norms conflict.

As personal agents weave themselves into the very fabric of our lives, it is crucial that those agents respect their users' values and act ethically. We understand a *value* as what is right or good according to an individual and *ethics* as a system of values. Rokeach[1] proposed two types of values—*terminal* values, referring to desired end-states of existence, and *instrumental* values, referring to modes of behavior or means to achieve the terminal values.

A socially intelligent personal agent (SIPA) would understand social contexts, including applicable norms, and help its users flexibly navigate those norms. Additionally, an ethical SIPA must understand terminal values, such as security, happiness, and recognition, and its actions must respect instrumental values such as honesty, helpfulness, and forgiveness.

Engineering ethical SIPAs faces two main challenges. First, a SIPA must recognize the relevant values and reason about the users' preferences over those values in order to choose an ethical action. A SIPA's action may simultaneously promote and demote different values.[2] For instance, a SIPA's action to share its user's location with family members promotes safety but demotes privacy.

Second, since people may have conflicting preferences on values,[3] a SIPA's decision about which values to promote or demote affects other users. For example, a teenager may prefer privacy over safety, but his parents may prefer the reverse. A SIPA's action to share the teenager's location affects both the teenager and the parents. Thus, an ethical SIPA must reason not only about its user's values and preferences, but also about those of others in the social context.

SOCIAL NORMS

Social norms are central to a social context. A norm characterizes interactions between autonomous parties. We adopt Singh's representation[4] in which a norm is directed from a subject to an object, as a conditional relationship involving an antecedent (which brings an instance of the norm in force) and a consequent (which brings the norm instance to completion). A new instance is generated whenever a norm applies. This representation yields clarity on who is accountable, when, for what, and to whom. A norm has four core elements, expressed as N(subject; object; antecedent; consequent), where N specifies the norm type. We consider norms of three types:

Commitment, C(subject; object; antecedent; consequent), means that its subject commits to its object to ensuring the consequent if the antecedent holds. For instance, consider a user, Aron, and his mother, Eevee. (We draw names from Pokémon anime.) Aron, who has poor night vision, could be committed to his mother, Eevee, that whenever he is out, he will keep Eevee informed of his location. Therefore, Aron is accountable for sharing his location to Eevee whenever he is out at night, which we write as:

$$C(\text{Aron}, \text{Eevee}, \text{notHomeAron} \wedge \text{evening}, \text{shareAronLoc})$$

Authorization, A(subject; object; antecedent; consequent), means that its subject is authorized by its object for bringing about the consequent if the antecedent holds. Although the authorized party can decide not to take up the authorization, the authorizing party must support the authorized condition if called upon.[5] That is, the authorizing party is accountable for ensuring success of the authorization's consequent if its antecedent holds. For example, Aron could authorize Eevee to access Aron's location if he is not at home before evening, which we write as:

$$A(\text{Aron}, \text{Eevee}, \text{notHomeAron} \wedge \text{evening}, \text{accessAronLoc})$$

Prohibition, P(subject; object; antecedent; consequent), means that its subject is forbidden by its object from bringing about the consequent if the antecedent holds. The subject is accountable for ensuring the consequent remains false. For instance, Eevee could be prohibited at all times by Aron from sharing his location to someone else, which we write as:

$$P(\text{Eevee}, \text{Aron}, \top, \text{shareAronLoc})$$

A *sanction* is an action, positive or negative, by a subject toward an object in response to the latter satisfying or violating a norm.[6]

SIPAS AND VALUES

To illustrate our ideas, consider Pikachu, a location sharing SIPA. Pikachu may share its user's geolocation and social context, including place (such as a bar or theater), companions, and activity. Importantly, Pikachu must ethically decide whether to share the user's details with no one, everyone (public), or specific people.

Example 1 *Aron values safety. Also, he has a commitment to his mother, Eevee, that he will share his location with her when he is not home. Sharing locations promotes safety. One evening, Aron meets a friend at The Flying Saucer, a local pub. Knowing Aron's commitments and values, Pikachu shares with Eevee that Aron is at The Flying Saucer with a friend.*

$$\text{C-share-AE} = C(\text{Aron}, \text{Eevee}, \top, \text{shareLocWithEevee})$$

$$\text{shareLocWithEevee} \Rightarrow \text{Sat}(\text{C-share-AE}) \wedge \text{safety} \uparrow$$

Example 2 *Aron values safety and social recognition, and commits to Eevee as before. Aron is attending a scientific conference in Stockholm. Sharing Aron's location with Eevee satisfies his commitment and promotes safety. Sharing Aron's location publicly additionally promotes social recognition. Thus, Pikachu shares publicly that Aron is in Stockholm attending a scientific conference.*

$$\text{shareLocWithEevee} \Rightarrow \text{Sat}(\text{C-share-AE}) \wedge \text{safety} \uparrow$$

shareLocWithAll ⇒ Sat(C-share-AE) ∧ safety ↑ ∧ social-recognition ↑

Example 3 *Continuing Example 2, Dr. Drampa, Aron's academic advisor, is attending the same conference. Dr. Drampa values privacy and prohibits his students from sharing location publicly when they are with Dr. Drampa. Now, by sharing Aron's location publicly, Pikachu promotes Aron's social recognition, but demotes Dr. Drampa's privacy and violates Aron's prohibition by Dr. Drampa. In contrast, by sharing his location with Eevee, Pikachu does not promote social recognition, and does not violate the prohibition or demote Dr. Drampa's privacy. Since Aron fears potential sanctions for violating Dr. Drampa's prohibition more than he prefers social recognition, Pikachu shares Aron's location only with Eevee.*

P-privacy-AD = P(Aron, Drampa, SameLoc, ShareLocWithAll)

shareLocWithAll ⇒ Sat(C-share-AE) ∧ Vio(P-privacy-AD) ∧ safety ↑ ∧ social-recognition ↑ ∧ privacy ↓

shareLocWithEevee ⇒ Sat(C-share-AE) ∧ Sat(P-privacy-AD) ∧ safety ↑ ∧ social-recognition ↓ ∧ privacy ↑

Example 4 *Aron is with Chansey on a midnight hike at Pilot Mountain. Chansey values privacy, and prohibits location sharing with all (just as Dr. Drampa does). However, Aron prefers safety to privacy in this context. Knowing these, Pikachu shares Aron's location with all his* friends *(which includes Eevee). Note that sharing with friends, is both safer and less privacy violating than sharing with all and does not violate Aron's prohibition from Chansey.*

P-privacy-AC = P(Aron, Chansey, SameLoc, ShareLocWithAll)
shareLocWithAll ⇒ Sat(C-share-AE) ∧ Vio(P-privacy-AC) ∧ safety ↓ ∧ privacy ↓
shareLocWithFriends ⇒ Sat(C-share-AE) ∧ safety ↑ ∧ privacy ↓

These examples demonstrate the complexity of ethical decision making. To act ethically, a SIPA must (1) acquire information about context, social norms, and values; (2) reason about actions despite conflicts among and between norms and values; and (3) potentially communicate its reasoning (arguments) to other SIPAs to avoid sanctions.[7] We need a systematic method to support SIPAs in accomplishing these nontrivial tasks.

VALAR: A FRAMEWORK FOR ETHICAL AGENTS

We propose Valar to engineer SIPAs that can understand preferences among values and reason about them to make policy decisions as exemplified above. Valar extends Arnor[7] with values and provides a four-step method to model stakeholders, contexts, social norms, and values.

Stakeholder modeling identifies the stakeholders, their goals, and relevant actions of a SIPA. A SIPA's *stakeholder* is either its user or someone affected by its actions. A stakeholder's *goal* defines what states he or she prefers. An *action* represents a step a SIPA may take.

Context modeling identifies contexts in which stakeholders interact. A *context* refers to the relevant circumstance of decision making, and it is crucial in determining which goals to bring about and which actions to perform.[8]

Social modeling identifies the norms and sanctions (see sidebar) associated with a stakeholder's goals and a SIPA's actions. The social norms and sanctions characterize the social architecture in which SIPAs act and interact.

Value modeling identifies the relevant values and stakeholders' preferences among those values, and how each action by the SIPA promotes or demotes the identified values. A stakeholder's *value* preference specifies what outcomes are morally superior to others in the stakeholder's judgment. Stakeholders' preferences among values provide a basis for choosing which goal to bring about or which norm to satisfy.

Figure 1 illustrates the main components of a Valar SIPA. A SIPA maintains (1) a model of the stakeholders, including their goals and values; (2) a world model, including its current state (context), and preconditions and effects of available actions; and (3) the social model, including applicable norms and sanctions. Using this information, the SIPA's decision module determines

an ethical action that would be most compatible with its stakeholders' value preferences and the applicable norms. The SIPA may perform the determined action or recommend it to its user depending on the application.

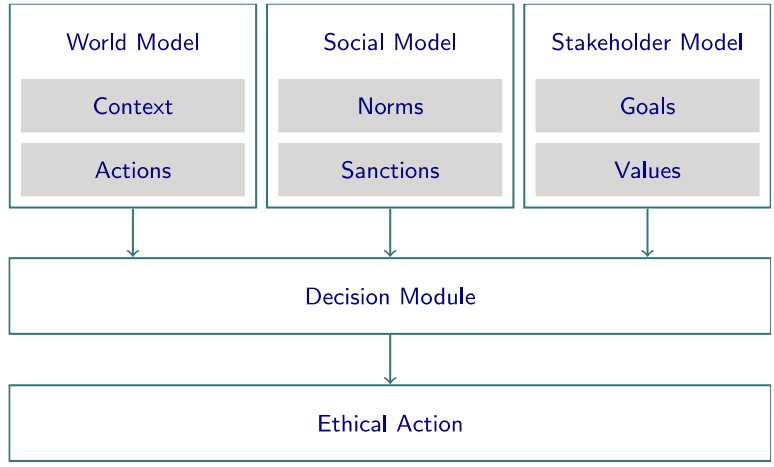

Figure 1. A conceptual model outlining decision making by a Valar SIPA.

Reasoning. A SIPA can choose to satisfy or violate norms by identifying stakeholders' contextual preferences among the values that these norms promote or demote. Following Sotala's approach,[9] a SIPA learns to maximize a reward function based on its stakeholders' values. For simplicity, a SIPA maintains each stakeholder's preferences as vectors of numeric weights on the various values—the higher the weight, the more important the corresponding value is for that stakeholder. Therefore, we can compute the extent to which an action promotes a stakeholder's values, or the aggregated value gain, as a weighted sum. A SIPA maintains the weight vector of different values under each social context, and respects values by choosing an action that produces the maximum aggregated gain.

EVALUATION: POTENTIAL BENEFIT OF VALUES

Evaluation is a challenge with any approach that involves informal, subjectively defined concepts such as ethics and values. We conducted a small empirical study to investigate if understanding the values promoted and demoted by a SIPA's potential actions and the stakeholders' preferences among the values could guide the SIPA to select actions that yield a pleasant social experience to its stakeholders.

Twenty-four graduate and nine undergraduate computer science students participated in our study, which was approved by North Carolina State University's Institutional Review Board (IRB).

We asked the participants to imagine they were in a given context—a combination of place (first column of Table 1); time of day of visit; and companions (alone, a colleague, crowd, a family member, or a friend). Each context was tagged as safe, unsafe, sensitive (disclosure of which may be harmful to the participants or their companions), or not sensitive.

Each participant completed two surveys to select a check-in policy (action) appropriate for that context. The first survey did not provide awareness of the values promoted or demoted by a sharing policy; the second survey provided awareness of the relevant values. Each survey asked for (1) a *check-in* policy ordered from high to low privacy preservation: share with *none*, *companions*, *common friends* (of companions), and *all*; and (2) a *confidence* in the selected check-in policy on a Likert scale of 1 (very low) to 5 (very high).

Making an informed decision. Figure 2 shows the violin plots for reported check-in policies for each of the eight places. We observe that an understanding of values significantly changes

participants' policy choices in the contexts of hiking and hurricane. In these contexts, location sharing promotes safety but demotes privacy, and participants generally preferred the former.

Table 1. The *p*-values indicating the difference in selected check-in policy and confidence when participants are aware and not aware of values promoted by each policy.

Context	Attribute	Policy *p*	Confidence *p*
Graduation ceremony	Not sensitive	0.07	<0.01
Conference presentation	Not sensitive	0.32	0.07
Library	Safe	0.85	0.59
Airport	Safe	0.08	0.23
Hiking at night	Unsafe	<0.01	0.02
Stuck in a hurricane	Unsafe	0.01	0.01
Bar with fake ID	Sensitive	0.83	0.53
Drug rehab	Sensitive	0.14	0.48

Making a confident decision. We observe that participants are more confident in making policy decisions for scenarios where they are made aware of the privacy, fame, and safety values.

We evaluated the corresponding statistical hypotheses via Wilcoxon's ranksum-test. Table 1 summarizes our results for eight conceptual places. The *p*-values obtained indicate that, in some contexts, the participants' decisions before and after they are primed with values are significantly different. Importantly, in some contexts, participants' confidence increases significantly when they are primed with values.

RELATED WORK

Kayal et al.[10] propose a value-based model for resolving conflicts between norms, especially social commitments. Their empirical results indicate that values can be used to predict users' preferences when resolving conflicts. Kayal et al.'s model can supplement Valar, which goes beyond conflict resolution, providing constructs and mechanisms to develop value-driven ethical SIPAs.

Dechesne et al.[3] develop a model of norms and culture, represented by values, to study norm compliance. They concur that values are important in deciding whether or not a norm should be introduced. Borning and Muller[11] motivate Value Sensitive Design to incorporate values in information technology, and highlight that values may differ widely across cultures and contexts.

Riedl and Harrison[12] argue that it is not easy for developers to exhaustively enumerate values, and propose that agents use sociocultural knowledge in stories, such as crowdsourced narratives, to learn values.

CONCLUSION AND FUTURE DIRECTIONS

We propose Valar, an agent-oriented software engineering method, to design ethical SIPAs that can reason about context, norms, values, and preferences among values. The preliminary results from our pilot study indicate that priming with values offers significant guidance to participants in making policy decisions. We conjecture that when SIPAs are made aware of such value preferences, they will choose ethical actions and offer a high-quality social experience to the stakeholders. However, these results are based on a small and biased sample without interaction with a production SIPA.

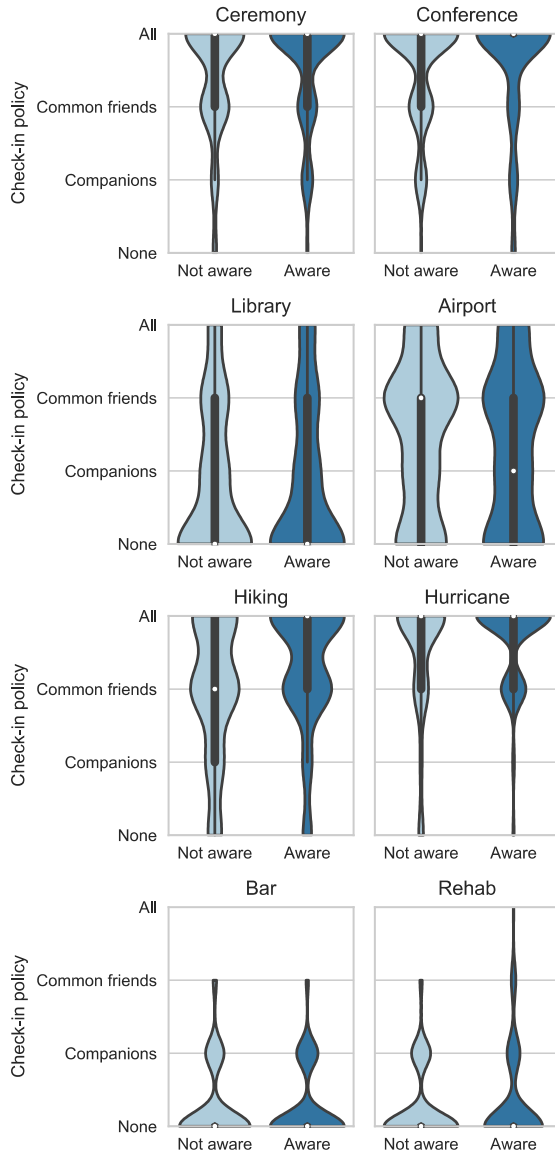

Figure 2. Policy when not aware of values versus when aware of values.

This topic suggests interesting future directions. One, to evaluate the effectiveness of Valar via a developer study. Two, to crowdsource data about values and decision making about sharing policies on a much larger scale. Three, to employ machine learning to assist SIPAs in learning value preferences of stakeholders, and accordingly select policies.

REFERENCES

1. M. Rokeach. *The Nature of Human Values*. Free Press, 1973.
2. P. Pasotti, M. B. van Riemsdijk, and C. M. Jonker. Representing human habits: Towards a habit support agent. *Proc. 10th International Workshop on Normative Multiagent Systems (NorMAS)*, 2016.

3. F. Dechesne, G. Di Tosto, V. Dignum, and F. Dignum. No smoking here: values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1):79–107, 2013.
4. M. P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology*, 5(1):21:1–21:23, 2013.
5. G. H. Von Wright. Deontic logic: A personal view. *Ratio Juris*, 12(1):26–38, 1999.
6. L. G. Nardin, T. Balke-Visser, N. Ajmeri, A. K. Kalia, J. S. Sichman, and M. P. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review*, 31:142–166, 2016.
7. N. Ajmeri, P. K. Murukannaiah, H. Guo, and M. P. Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. *Proc. 16th International Conf. Autonomous Agents and Multiagent Systems*, pages 230–238, 2017.
8. P. K. Murukannaiah and M. P. Singh. Xipho: Extending Tropos to engineer context-aware personal agents. Proc. *14th International Conf. Autonomous Agents and MultiAgent Systems*, pages 309–316, 2014.
9. K. Sotala. Defining human values for value learners. *Proc. AAAI Workshop on Artificial Intelligence AI, Ethics, and Society*, pages 113–123, 2017.
10. A. Kayal, W.-P. Brinkman, M. A. Neerincx, and M. B. van Riemsdijk. Automatic resolution of normative conflicts in supportive technology based on user values. *ACM Transactions on Internet Technology*, 2017.
11. A. Borning and M. Muller. Next steps for value sensitive design. *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 1125–1134, 2012.
12. M. O. Riedl and B. Harrison. Using stories to teach human values to artificial agents. *Proc. AAAI Workshop on Artificial Intelligence AI, Ethics, and Society*, pages 105–112, 2016.

AUTHOR BIOS

Nirav Ajmeri is a PhD student in Computer Science at North Carolina State University. His research interests include artificial intelligence, multiagent systems, and software engineering with a focus on security and privacy. Ajmeri has a BE in Computer Engineering from Sardar Vallabhbhai Patel Institute of Technology, Gujarat University. Contact him at najmeri@ncsu.edu.

Hui Guo is a PhD student in Computer Science at North Carolina State University. His research interests include multiagent systems, NLP, text mining, and crowdsourcing. Guo has an MS in Computer Science from East Carolina University, and a BS from Tsinghua University. Contact him at hguo5@ncsu.edu.

Pradeep K. Murukannaiah is an Assistant Professor at Rochester Institute of Technology. He received a PhD and an MS in Computer Science from North Carolina State University. Pradeep's research seeks to facilitate the engineering of intelligent personal agents that deliver a personalized, context-aware, privacy-preserving experience to users. Contact him at pkmvse@rit.edu.

Munindar P. Singh is an Alumni Distinguished Graduate Professor in Computer Science and a co-director of the Science of Security Lablet at North Carolina State University. His research interests include the engineering and governance of sociotechnical systems. Singh is an IEEE Fellow, a AAAI fellow, a former Editor-in-Chief of *IEEE Internet Computing*, and the current Editor-in-Chief of ACM *Transactions on Internet Technology*. Contact him at singh@ncsu.edu.

ACKNOWLEDGMENTS

We thank the US Department of Defense for support through the Science of Security Lablet at North Carolina State University.