ABSTRACT

AJMERI, NIRAV. Engineering Multiagent Systems for Ethics Aware and Privacy Respecting Social Computing. (Under the direction of Dr. Munindar P. Singh).

A socially intelligent personal agent understands and helps its user navigate the norms governing the user's interaction in a society. This research seeks to advance the science of privacy by tackling nuanced notions of privacy, understood as an ethical value, in personal agents. It addresses the research question of how we can engineer social intelligence in a personal agent such that it selects ethically appropriate actions and delivers a pleasant and privacy-respecting social experience. We develop multiagent system techniques to engineer such socially intelligent and ethical personal agents.

This research develops (1) Arnor, a software engineering method to engineer privacy-aware personal agents by modeling social intelligence via norms, (2) Poros, an approach that enables personal agents to reason about shared contexts and infer contextually relevant social norms that preserve privacy, and (3) Ainur, a decision-making framework to design personal agents that can reason about values and act ethically.

Arnor goes beyond traditional software engineering methods to engineer personal agents by systematically capturing interactions that influence social experience. We claim that (1) Arnor supports developers in engineering intelligent personal agents, and (2) personal agents engineered using Arnor provide a better privacy-preserving social experience than agents engineered using a traditional software engineering method. We evaluate Arnor via a developer study and a set of simulation experiments and measure the social experience via metrics of norm compliance and sanction proportion.

A personal agent may deviate from norms in certain contexts. Poros (1) enables personal agents deviating from norms to share the context of a deviation with other agents and (2) provides personal agents the ability to reason about shared contexts. We make two claims about the impact of context sharing and reasoning in Poros. First, the ability to reason about deviation contexts helps a personal agent accurately infer contextually relevant norms and act in a privacy-respecting manner. Second, by acting according to such contextually relevant norms, a personal agent yields higher goal satisfaction to its users than an agent that does not reason about shared contexts. We demonstrate these claims via social simulations involving agent societies of varying sizes and diverse characteristics reflecting pragmatic, considerate, and selfish agents.

Privacy, values, and ethics are closely intertwined. Preserving privacy presumes un-

derstanding human values and acting ethically. Ainur equips a personal agent with an understanding of values such as pleasure, privacy, recognition, and security, that are promoted or demoted by the agent's actions. This understanding of values helps personal agents select ethically appropriate actions especially in scenarios where either the norms conflict or the value preferences of the users are not aligned. We empirically evaluate Ainur via simulation experiments seeded with data from a user survey. We find that agents developed using Ainur produce ethical actions that exhibit fairness and yield a pleasant social experience to the agents' users.

Engineering Multiagent Systems for Ethics Aware and Privacy Respecting Social Computing

by
Nirav Ajmeri

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2019

APPROVED BY:

Dr. Jon Doyle	Dr. William H. Enck
Dr. Jessica N. Staddon	Dr. Laurie A. Williams

Dr. Munindar P. Singh
Chair of Advisory Committee

DEDICATION

To Chhitu dada and Keshav dada.

BIOGRAPHY

Nirav Ajmeri was born to Vina and Suresh Ajmeri in Vadodara, the cultural capital of the state of Gujarat, India. He grew up in the holy city of Mathura in Uttar Pradesh, India, and later lived in New Delhi, Vadodara, Thiruvananthapuram, and Pune in India.

He obtained a B.E. in Computer Engineering from Sardar Vallabhbhai Patel Institute of Technology, Gujarat University, and an M.S. in Computer Science from NC State University. Prior to joining North Carolina State University for his doctoral studies, Nirav worked as a researcher in Software Engineering Lab at Tata Research Design and Development Center, India. During his doctoral studies, Nirav interned at HERE Technologies (formerly Nokia Maps) with its CTO Research team as a research intern.

Nirav loves his family, likes cricket and arcade games, and follows infrastructure development forums. In his free time, he runs an arcade gaming website and a social bookmarking website.

ACKNOWLEDGMENTS

This work has benefited from collaborations with several people. First and foremost, I express my deepest gratitude to my advisor Dr. Munindar Singh for his astute guidance, support, and encouragement. I am forever grateful to him for everything that I learned.

I am sincerely thankful to members of my advisory committee, Drs. Jon Doyle, William Enck, Chris Mayhorn, Jessica Staddon, and Laurie Williams. My work has greatly benefited from the interactions with them and their valuable advice over the years.

Chapters 2, 3, and 4 are based on joint works with my colleagues Dr. Pradeep Murukannaiah and Hui Guo. Interactions with Dr. M. Birna van Riemsdijk and Pietro Passoti form the foundation of Chapter 4. Innumerable discussions and iterations with Pradeep and Hui have shaped and improved this work.

I have benefited from several other collaborators at NC State University and outside. Although not all works I completed with them are included in this dissertation, their knowledge and insights have undoubtedly influenced my thinking. I thank each of them. Collaboration with Drs. Chung-Wei Hang and Simon Parsons introduced me to argumentation theory. Discussions with Dr. Shams Al-Amin, Dr. Tina Balke-Visser, Dr. Emily Berglund, Dr. Jon Doyle, Dr. Hongying Du, Shubham Goyal, Dr. Anup Kalia, Dr. Luis Gustavo Nardin, Bennett Naron, and Dr. Jaime Sichman helped in understanding nuances of sanctions. Collaboration with Drs. Sibel Adalı, Kevin Chan, Jin-Hee Cho, and Anup Kalia improved my understanding of norms, trust, and emotions in multiagent systems, and taught me how to conduct human subject studies. Work with Chris Allred, Dr. Mark Wilson, and Dr. Guangchao Yuan helped me learn conducting crowdsourcing studies. Discussions with Dr. Pradeep Murukannaiah have influenced my understanding of software engineering, crowdsourcing and creativity. Discussions and works with Dr. Rada Chirkova, Dr. Jon Doyle, Jiaming Jiang, and Dr. Özgür Kafalı helped me learn formalization and improved my understanding of sociotechnical systems and cybersecurity. I thank Drs. Shams Al-Amin, Hongying Du, and Mehdi Mashayekhi for helping me learn tooling multiagent simulations. From Hui Guo and Dr. Zhe Zhang, I learned text mining and language processing. Collaboration with Karthik Sheshadri and Dr. Jessica Staddon has influenced my understanding of privacy. I learned various aspects of software requirements engineering and knowledge engineering while working at Tata Research Design and Development Center (TRDDC), India. I am grateful to my mentor Dr. Smita Ghaisas, and my colleagues Preethu Rose, Manish Kumar, Manas Agarwal, Riddhima Sejpal, Kumar Vidhani, Manoj Bhat, Manish

Motwani, and Shashikant Sharma at TRDDC. My internship at HERE Technologies introduced me to location and trajectory privacy research. I am thankful to my mentors Drs. Matei Stroila, Raghavendra Balu, and Bo Xu.

I am also thankful to my other colleagues at the Multiagent Systems and Service-Oriented Computing lab including Samuel Christie, Zhen Guo, Mu Zhu, and Shrey Anand. I have learned from each of them and sincerely appreciate their encouragement, discussions, and support.

My experience would not have been memorable without the friendship I have made throughout my studies. I thank (in no particular order) Keyur Patel, Hardik Amin, Hitesh Makwana, Gaurav Varshikar, Khushali Khadiwala, and Prachi Agarwal for being my constant source of inspiration. They have patiently listened to my random ideas and have always given worthy feedback. Raleigh would not have been lively for me without (in no particular order) Harsh Patel, Divya Mehta, Vandit Khamker, Abhinav Sarkar, Neeraj Badlani, Sarvesh Rangnekar, Arvind Telharkar, Ashwin Shashidharan, Aruni MK, Anup Kalia, Sweta Rout, Pradeep Murukannaiah, Indumathi Srinivasachari, Anant Raj, and Prerna Prateek. I thank them for all of the discussions, dinners, games, movie nights, and outings.

I am deeply indebted to my parents, Vina and Suresh, for helping me become who I am today. Sacrifices they have made for me are beyond measure. My wife, Rucha, stood by me throughout my graduate school journey. I am forever grateful to her for her love, support, and encouragement. I also thank my family, particularly, Baa, Dada, Nana, Hetal, and Arpit for their unconditional love and support. Although, Dada and Nana are not with us anymore, I know they are proud. I express sincere gratitude to my parents-in-law, Meenakshi and Haresh, and their family for their well wishes and unwavering support. Special thanks to Aru for all the happiness. This journey would have neither started nor concluded without all of their support.

Lastly, I thank the US Department of Defense for support through the Science of Security Lablet at NC State University and the Laboratory of Analytic Sciences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

My Thesis

Software developers can engineer personal agents that deliver to its stakeholders an ethical and privacy-respecting social experience by modeling and reasoning about social norms, social contexts, and value preferences of the stakeholders.

Social computing has underpinnings from both computational sciences and social sciences [Wang et al., 2007]. It involves interplay between computing and social entities including people and relationships between them. Existing research limits the scope of social computing to social network analysis. In this research, we expand its scope to developing computing technologies and infrastructure via which we can understand social reality including social behaviors, social contexts, and social expectations.

Privacy in social computing encompasses both technical and technical aspects. But much of the literature in privacy has focused on these aspects as two different goals. Some research aims to design secured systems with the help of cryptographic protection. Other research aims to protect personal information by facilitating informed choice options to an individual and assume that policies and regulations are enforceable. This research tackles the science of privacy from a sociotechnical viewpoint that bridges the two goals.

Human interactions in a society are driven not merely by personal needs and expectations as Chapter 2 explains. Others around us and their expectations play a prominent part on the way we act and interact. A personal agent (a social computing technology) acts and interacts on behalf of its human user.

A *socially intelligent personal agent* (SIPA) adheres to *social expectations* of its *primary*

and *secondary stakeholders* (defined later in Chapter 2), adapts according to the circumstances or *social context* [Dey, 2001], acts on behalf of its user (primary stakeholder), and provides a pleasing social experience to all of its stakeholders as opposed to an individual experience to its user.

The key objectives of this research are: (1) to engineer personal agents such that they deliver a pleasant social experience, and yet preserve their stakeholders' privacy, and (2) to make engineering of such personal agents efficient and effective for developers. We recognize social norms, social context, and value preferences of stakeholders as important factors that influence the working of such personal agents.

1.1 Preliminaries

We now provide some necessary background on values, privacy, and social norms in multiagent systems.

1.1.1 Values

The concept of *value* has two main connotations: one is about the economic worth of something and the other, more broadly, refers to what people consider important in their lives [Friedman et al., 2008b]. We adopt the later connotation in this research.

Values are mostly universal across human societies, as stated by Schwartz [2012] and Rokeach [1973]. The values in Schwartz's [2012] work are broad motivational goals, such as stimulation, achievement, security, and benevolence. Rokeach [1973] proposes two types of values—*terminal* and *instrumental*. Terminal values, such as security, freedom, happiness, and recognition, refer to defined-end states of existence. Instrumental values refer to modes of behavior or means to promote the terminal values. Ethicists subsume ethics in the theory of values [Friedman et al., 2008a]. We recognize an ability to understand these values as an important aspect in a SIPA for it to deliver an ethical experience. Dechesne et al. [2013] define values as ideals worth pursuing. They observe that these ideals could conflict since they may not be preferred equally by each individual (that is, each SIPA stakeholder, in our research).

1.1.2 Privacy

The concept of privacy encircles several areas. An individual's notion of privacy is partly based upon the society's notion of privacy and partly based upon his or her personal experiences [Westin, 1967, 2003]. The attitude toward sharing personal information varies from one individual to other.

Westin [1967] classifies individuals based on their privacy preferences: *privacy fundamentalists* are the individuals who are extremely concerned about their privacy and are reluctant to share personal information; *privacy pragmatists* are concerned about privacy but less than fundamentalists and they are willing to disclose personal information when some benefit is expected; and, *privacy unconcerned* do not consider privacy loss when disclosing personal information [Westin, 1967]. The pragmatists are further grouped into *identity-aware* and *profile-aware* individuals [Spiekermann and Cranor, 2009]. Identity-aware individuals are those who are more concerned about revealing identifying information such as e-mail or physical address rather than revealing their interests. Profile-aware individuals worry more about sharing their hobbies, age, interests, or preferences.

Privacy has been a topic of debate and has no globally agreed-upon definition [Smith and Shao, 2007]. Theorists from the areas of law, philosophy, sociology, and computer science have defined privacy in their respective perspectives. The importance of privacy and what it brings to an individual is also debated. Although there is no single definition, researchers acknowledge the idea of privacy, and view it as a collection of concepts instead of one specific concept [Smith and Shao, 2007].

Privacy has been considered a right, and is protected by regulations. Prosser [1960] discusses the right of privacy from a legal perspective. Solove [2006] provides a taxonomy of activities that can violate privacy. The purpose of Solove's taxonomy is to aid the development of privacy laws, in protecting the right of privacy. Spiekermann [2012] lists the challenges of Privacy by Design, a proposed solution to the regulation of privacy, including the differentiation between privacy and security and detailed methods to incorporate privacy into system design. Westin [2003] dissects the values of privacy in modern societies from political, sociocultural, and personal dimensions. He defines privacy as a claim of an individual and, when recognized by law and social convention, a right, to determine the revelation of his or her information.

Privacy is inherently a human value [Smith and Shao, 2007; Spiekermann and Cranor, 2009]. In regards to ethics, privacy is considered as an ethical value [Langheinrich, 2001; Taylor, 2002].

Privacy as an Ethical Value.

In this research, we understand privacy as a value with an ethical import [Langheinrich, 2001; Taylor, 2002]. We adopt the nuanced notions (specifically intrusion, appropriation, and disclosure) of privacy as defined in Solove's taxonomy [Solove, 2006].

1.1.3 Social Norms and Multiagent Systems

The concept "norm" is used in different disciplines and thus has variant notions such as social expectations, legal laws, and linguistic imperatives [Boella et al., 2009b]. We adopt the concept of social norms, which describe interactions between principals in terms of what they ought to be, or reactions to behaviors, including attempts to apply sanctions. Thus, social norms regulate the interactions of the principals involved. Norms and normative systems are gaining increasing interest in the computer science community. Meyer and Wieringa [1993] define normative systems as "systems in the behavior of which norms play a role and which need normative concepts in order to be described or specified." Normative multiagent systems as a research area can be defined as the intersection of normative systems and multiagent systems. Norms govern much of our social lives, and therefore are considered as a key element of artificial agents that are expected to behave comparably to humans [Boella et al., 2006].

We adopt Singh's [2013] representation of social norms in which norms are classified as five types: commitment, authorization, prohibition, sanction and power. We consider two main norm types in this work: commitment and prohibition. A commitment norm means its subject is committed to its object to bring about a consequent if an antecedent holds. For instance, *Frank* (subject), a high school student is *committed* (norm) to *Grace* (object), his mother, that he *will keep Grace updated about his location* (consequent) when he is *away from home* (antecedent). A prohibition norm means its subject is forbidden by its object to bring about a consequent if an antecedent holds. For instance, *Frank* (subject), is *prohibited* (norm) by Heidi (object), his class teacher, from *answering phone calls* (consequent) when he is *in a classroom* (antecedent).

1.2 Motivating Example

Example 1 *Consider a ringer manager as a SIPA. The ringer manager installed on Alice's phone decides appropriate ringer modes (loud, silent, or vibrate) for incoming calls. Alice, the phone owner is the primary stakeholder of the SIPA. Bob, Alice's friend who calls Alice often, and Charlie and Dave, Alice's coworkers, who are in her vicinity, are some of the secondary stakeholders. Further, the ringer manager's capabilities influencing its social experience include (1) allowing Alice to be tele-reachable, (2) notifying the caller if Alice is not reachable, (3) enabling Alice to work uninterrupted, and (4) not annoying Alice's neighbors.*

Suppose that Bob calls Alice when she is in an important meeting with Charlie and Dave. Alice is *committed* (a social norm) to answering Bob's phone calls. Another *commitment* is to

keep one's phone silent during important meetings. Alice's SIPA, understanding the norms and knowing that Bob's calls to Alice are generally casual, puts Alice's phone on silent for Bob's call and notifies Bob that Alice is in a meeting; later when Alice's meeting ends, Alice's SIPA reminds her to call Bob.

Should Alice's phone rings loudly during the meeting, privacy implications may follow [Murukannaiah et al., 2016a; Solove, 2006]. A loud ring *intrudes* upon Alice's and other meeting attendees' privacy in that call violates the meeting attendees' reasonable expectation to be left alone. Further, it is likely that meeting attendees frown at Alice (*disapprobation*). If Alice answers the call, those overhearing Alice and Bob's conversation can gain knowledge about her and her interlocutor (*information leak*). If Bob's call were urgent, Bob's SIPA could communicate the urgency to Alice's SIPA, and Alice's SIPA could deliver a different social experience, e.g., set the phone on vibrate to notify Alice of the urgent call and yet not annoy other meeting attendees. Should Alice's phone stay silent for Bob's urgent call, it may affect Alice's and Bob's social relationship.

In the examples above, ringer manager SIPA makes nontrivial decisions influencing social experience of its stakeholders. Existing software engineering methods [Bresciani et al., 2004; Murukannaiah and Singh, 2014; Winikoff and Padgham, 2004] are good starting point to engineer personal agents, however these methods do not guide developers with systematic steps to represent and reason about such scenarios, and thus fall short in supporting agents that adapt to evolving social contexts at runtime.

Social norms inform personal agents about a set of reasonable actions in a social context [van Riemsdijk et al., 2015a]. Norm compliance in a social context is achieved either by (1) conveyance of norms, where SIPAs are made aware of norms by direct communication, or (2) via (positive and negative) sanctions, where personal agents learn norms in the form of which actions are appropriate in a context [Andrighetto et al., 2013].

Under certain circumstances, we (as humans) may deviate from norms. When we deviate, we may offer an explanation typically revealing the context of the deviation. Revealing context may lessen the burden of deviation, and may help us avert sanctions resulting from the deviation. Deviations from norms often hint toward a different norm that is contextually more relevant. For instance, if Alice reveals to meeting attendees' that the call was from a sick friend who needs urgent care, the meeting attendees' (1) may not frown on Alice, and (2) may learn that although it is not appropriate to answer calls during meetings as it intrudes upon attendees' privacy, answering an emergency call is acceptable as it could ensure someone's well being or safety. An ability to reason about the deviation context, and an understanding of the values promoted or

demoted by different actions could assist SIPAs in providing a pleasing social experience to its stakeholders.

We recognize three key challenges. One, understanding what constitutes a social experience, and how SIPA's actions influence the social experience and privacy of its stakeholders? When SIPAs satisfy or violate norms, they might share certain contextual information related to satisfaction or violation. Social experience depends largely on how SIPAs' stakeholders perceive shared information. Two, how and what contextual information should a SIPA disclose? When norms conflict, SIPAs must perform actions that promote richer social experience. Three, how can we develop decision support to recommend actions?

1.3 Research Questions

Based on the aforementioned challenges and nuances illustrated by the above example, we seek to investigate the following research questions:

RQ_1 Social Intelligence. How can modeling social intelligence in a SIPA help deliver a social experience and respect its stakeholders' privacy?

RQ_2 Context. How can SIPAs share and adapt to deviation contexts, and learn contextually relevant norms?

RQ_3 Values. How can a SIPA reason about values promoted or demoted by its actions and understand preferences among these values?

1.4 Contributions

To address the research questions of modeling social intelligence and enabling ability to reason about deviation contexts and values, we develop (1) Arnor, a software engineering method; (2) Poros, a context reasoning approach; and (3) Ainur, a value-based decision-making framework.

1.4.1 Arnor: Modeling Social Intelligence via Norms

This work appears in Proceedings of the *16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2017* as a paper "Arnor: Modeling Social Intelligence via Norms to Engineer Privacy-Aware Personal Agents" [Ajmeri et al., 2017b].

To address the research question of modeling social intelligence, we develop Arnor (Chapter 2), a systematic software engineering method. It facilitates developers to model stakeholders' actions

and expectations, and how these influence each other. Arnor employs Singh's [2013] model of (social) norms to capture social requirements, and incorporates argumentation constructs [Bench-Capon and Dunne, 2007] for sharing decision rationale. Since, testing a SIPA's adaptability in all possible social contexts is logistically challenging and time consuming, Arnor also incorporates a SIPA simulation testbed. We rigorously evaluate Arnor via a developer study and a set of simulation experiments on the simulation testbed.

We hypothesize that the developers who follow Arnor (1) produce better models, (2) expend less time during application development, (3) feel it is easier to develop a SIPA, and (4) expend less effort, than those who follow Xipho Murukannaiah and Singh [2014], an existing software engineering methodology geared toward engineering personal agents. We find that developers using Arnor spend less time and effort, and overall feel it is easier to engineer a SIPA using Arnor. No significant difference is found in the model quality.

We hypothesize that SIPAs developed using Arnor (1) have better adaptability features, and (2) provide richer social experience, than SIPAs developed using Xipho. We measure social experience via norm compliance and sanction proportion measures. We find that SIPAs engineered using Arnor have greater adaptability correctness, similar norm compliance, and are prone to lesser sanctions.

1.4.2 Poros: Understanding and Reasoning about Social Context

This work appears in Proceedings of the *27th International Joint Conference on Artificial Intelligence (IJCAI), 2018* as a paper "Robust Norm Emergence by Revealing and Reasoning about Context: Socially Intelligent Agents for Enhancing Privacy" [Ajmeri et al., 2018d].

Norms describe the social architecture of a society and govern the interactions of its member agents. It may be appropriate for an agent to deviate from a norm; the deviation being indicative of a specialized norm applying under a specific context. Existing approaches for norm emergence assume simplified interactions wherein deviations are negatively sanctioned. To address the research question of understanding social context, we develop Poros (Chapter 3), an approach for building SIPAs that carry out enriched interactions where deviating SIPAs share selected elements of their context, and other SIPAs respond appropriately to the deviations in light of the received information.

We investigate via simulation the benefits of such enriched interactions. We find that as a result (1) the norms are learned better with fewer sanctions, indicating improved social cohesion; and (2) the agents are better able to satisfy their individual goals. These results are robust under societies of varying sizes and characteristics reflecting pragmatic, considerate, and selfish

agents.

1.4.3 Ainur: Reasoning about Values and Ethics

Parts of this work appears in *IEEE Internet Computing* magazine as a column "Designing Ethical Personal Agents" [Ajmeri et al., 2018b] and in Proceedings of the *5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security (HotSoS)* as a poster paper "Ethics, Values, and Personal Agents" [Ajmeri et al., 2018c].

Privacy, values, and ethics are closely intertwined. Preserving privacy presumes understanding of human values and acting ethically. If norms require agents to perform or not perform certain actions, values provide a reason to or not to pursue those actions [Dechesne et al., 2013]. Each action a Poros SIPA executes potentially promotes or demotes one or more *values*. For instance, a callee's action of answering an urgent phone call during a meeting may promote *safety* of the caller, but demote *privacy* of the meeting attendees. Being aware of these values and having an ability to reason about them helps a SIPA select ethical actions and yield pleasant experience.

To address the research question of reasoning about values, we propose Ainur, a framework to design such ethical SIPAs. We incorporate a multicriteria decision-making method in Ainur to aggregate value preferences of stakeholders and select an ethically appropriate action.

We empirically evaluate Ainur via multiple simulation experiments. We find that agents developed using Ainur produce ethical actions that yield a pleasant social experience to its stakeholders.

1.5 Organization

Chapter 2 details Arnor, and discusses its evaluation. Chapter 3 describes Poros and its evaluation via simulation experiments. Chapter 4 describes Ainur, and its empirical evaluation. Chapter 5 concludes with important future directions.

Chapter 2

Modeling Social Intelligence via Norms to Engineer Privacy-Aware Personal Agents

This chapter addresses our first research question, RQ_1 on social intelligence. The chapter describes Arnor, our methodology to model social intelligence in personal agents, and its empirical evaluation via a developer study and simulation experiments. It is based on a paper "Arnor: Modeling Social Intelligence via Norms to Engineer Privacy-Aware Personal Agents" that appears in *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

2.1 Introduction

Our actions and interactions in a society are not driven solely by individual needs. Instead, we adapt our behavior considering the needs of others, e.g., by being courteous and lending a helping hand. Such acts, even if inconvenient at times, deliver a pleasant social experience.

Privacy encompasses both social and technical aspects. However, most of the traditional works have approached privacy from a technical standpoint. We tackle the science of privacy from a sociotechnical viewpoint [Chopra and Singh, 2016; Kafalı et al., 2016].

Consider a society in which an agent acts and interacts on behalf of a *stakeholder* (human user). Our objective is to engineer the agents such that they deliver a *social experience* relative to that society, as opposed to individual user experiences. We refer to an agent delivering a social experience as a *socially intelligent personal agent* (SIPA). The *primary* stakeholder of a SIPA is the user who directly interacts with the SIPA, and on whose behalf the SIPA acts and

interacts. A *secondary* stakeholder of a SIPA may not directly interact with the SIPA, but the SIPA's actions affect the secondary stakeholder.

To understand the nuances in modeling social intelligence in SIPAs, let us revisit the example in Chapter 1.

Example 2 *Consider a ringer manager as a SIPA installed on Alice's phone. The ringer manager decides appropriate ringer modes (e.g., loud or silent) for incoming calls. Alice is the ringer manager's primary stakeholder. Bob, Alice's friend, calls her when Charlie and Dave, Alice's coworkers, are in her vicinity. Bob, Charlie, and Dave are the ringer manager's secondary stakeholders.*

We define social experience as the collective experience a SIPA delivers to each of its primary and secondary stakeholders. Respecting stakeholders' privacy is an important aspect of delivering social experience.

Example 3 *Bob calls Alice when she is in an important meeting with Charlie and Dave.*

In Example 3, should Alice's phone ring loud during the meeting, privacy implications may follow [Murukannaiah et al., 2016a; Solove, 2006]. A loud ring *intrudes* upon Alice's and other meeting attendees' privacy in that the call violates their reasonable expectation to be left alone. Further, Alice may receive nasty looks from other attendees (*disapprobation*). If Alice answers the call, those overhearing Alice and Bob's conversation can gain knowledge about her and her interlocutor (*information leak*).

Example 4 *Alice is in a meeting with Charlie and Dave. Bob is in a car accident and calls Alice for assistance. Bob's ringer manager communicates the urgency to Alice's ringer manager, which then sets her phone to ring loud. It also notifies Charlie and Dave about the situation.*

Should Alice's phone stay silent for Bob's urgent call, Bob's trust for Alice may reduce, affecting their social relationship. Instead, if the phone rings loud and Alice communicates a rationale to Dave and Charlie, presumably, they would not frown at her.

These examples demonstrate the nontrivial decisions a SIPA must make and the implications those decisions have on the stakeholders' social experience and privacy. These nuances prompt us to investigate the research question:

RQ. How can we engineer a SIPA such that it delivers a social experience but respects its stakeholders' privacy?

Three key challenges in engineering a SIPA to deliver a social experience are understanding (1) what constitutes social experience; (2) how a SIPA's actions influence the social experience and privacy for each stakeholder; and (3) how a SIPA's actions evolve when it is put to use in a variety of social contexts.

Existing agent-oriented software engineering (AOSE) methods provide a good starting point for addressing the first challenge. For example, Tropos [Bresciani et al., 2004] actor models and Gaia [Wooldridge et al., 2000] interaction models capture stakeholders and coarse dependencies between them. However, these methods provide little guidance on capturing how an agent's actions and interactions influence each stakeholder involved (second challenge). Also, these methods provide design-time constructs to model an agent, but fall short in modeling social interactions that support agents to adapt to evolving social contexts at run time (third challenge). Our formulation contrasts with Tropos where the stakeholders are characterized by their goals, as in caller, callee, and neighbor, but a single perspective is taken in the actor produced. We consider multiple perspectives where each agent corresponds to one user and has its loyalty to that user.

Norms have been widely studied with several works addressing norm conflicts, compliance, and emergence via either simulation or formalization [Alechina et al., 2016; Criado and Such, 2016]. Van Riemsdijk et al. [2015b] argue for a personal agent's need to explicitly represent norms. Social norms inform SIPAs about a set of reasonable actions in a social context. Norm compliance in a social context is achieved either by (1) establishment of norms, where SIPAs are made aware of norms by direct communication, or (2) via (positive and negative) sanctions, where SIPAs learn norms in the form of appropriate actions in a social context [Andrighetto et al., 2013]. Also, a SIPA's decision rationale for its action influences how other stakeholders perceive satisfaction or violation of a norm, and the nature of sanctions that they apply.

Contribution

To address the aforesaid challenges, we propose Arnor, a systematic method enabling the development of privacy-aware socially intelligent personal agents via social constructs. Arnor facilitates agent developers in modeling stakeholders' social expectations and, how an agent's actions influence those expectations, thereby enabling SIPAs that deliver a rich social experience. Arnor employs Singh's [2013] model of (social) norms to capture social requirements, and incorporates argumentation constructs [Bench-Capon and Dunne, 2007] for sharing a decision rationale.

Testing a SIPA's adaptability in all possible social contexts would be infeasible. To overcome

this challenge, Arnor incorporates a SIPA simulation testbed. Seeded with crowdsourced data, Arnor's testbed enables designers to test a SIPA's runtime adaptability. We rigorously evaluate Arnor via two studies: (1) a multiphase developer study in which developers engineer a SIPA, and (2) a set of adaptability studies in which we simulate the adaptability of SIPAs developed in the first study in a variety of social contexts.

Novelty

Arnor goes beyond existing AOSE methods by assisting developers to incorporate social norms and reason about how those norms influence social experience. In spirit, Arnor is a hybrid method in that it addresses the problem of engineering SIPAs combining top-down (via modeling) and bottom-up (via experience or social learning [Sen and Airiau, 2007]) styles.

Section 2.2 briefly describes the background works on which we build. Section 2.3 describes Arnor in detail. Section 2.4 describes our developer and simulation studies, and Section 2.5 presents our results and discusses threats to validity. Section 2.6 discusses related works and Section 2.7 concludes with important future directions.

2.2 Background

Arnor builds on the AOSE methods of Tropos and Xipho, and on the constructs of social norms and sanctions.

2.2.1 Tropos and Xipho

Tropos [Bresciani et al., 2004] is an end-to-end AOSE methodology spanning requirements modeling, design, and implementation. Tropos provides systematic steps to model and refine an application to be developed via high-level abstractions.

We adopt the following Tropos abstractions. An *actor* is a social, physical, or a software agent. An actor has *goals* (strategic interests) and *plans* (means of achieving a goal) within a system. Further, an actor's goals can be *hard* (having a specific satisfaction condition) or *soft* (not have a specific satisfaction condition). A *belief* is an actor's perspective of the environment and a *resource* is a physical or information entity. An actor may have *dependencies* with other actors to satisfy goals, execute plans, or acquire resources.

Figures 2.1 shows a Tropos system-as-is model (the as-is model captures the setting in which the agent to be developed, e.g., the ringer manager, operates). This model identifies the

stakeholders and dependencies between them as well as the goals and plans of the stakeholders.

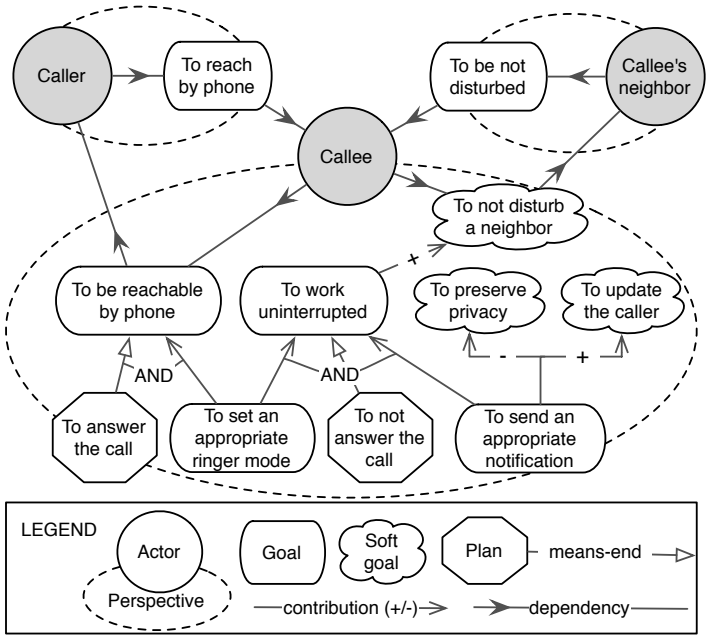

Figure 2.1 A Tropos system-as-is model of the ringer manager, expanding the callee's perspective [Murukannaiah and Singh, 2014].

Xipho [Murukannaiah and Singh, 2014] extends Tropos to engineer personal agents. Xipho introduces *context* as a high-level abstraction and treats an actor's goals, plans, and dependencies as inherently contextual. Xipho enables a developer to tailor a generic model of context to a specific application scenario via systematic steps through distinct development phases.

2.2.2 Norms and Sanctions

A norm as understood here [Singh, 2013] is directed from a subject to an object and is constructed as a conditional relationship involving an antecedent (which brings the norm in force) and a consequent (which brings the norm to satisfaction or violation). This representation yields clarity on who is accountable to whom. A norm can be formalized as:

$$\mathsf{N}(\text{subject}, \text{object}, \textit{antecedent}, \textit{consequent})$$

We employ the following types of norms in our approach.

- A *commitment* (C) means that its subject commits to its object to ensure the consequent if the antecedent holds. An example commitment is that, in a meeting room, the participants may be committed to each other to keep their phones silent: C(phone-user, coworker, place = *meeting*, ring = *silent*).

- A *prohibition* means that its subject is forbidden by its object from bringing about the consequent if the antecedent holds. An example prohibition (P) is that, in an examination hall, the students may be prohibited by a proctor from answering phone calls: P(phone-user, proctor, place = *examination*, ring = *silent*).

- A *sanction* specifies the consequences its subject faces from its object for satisfying or violating another norm, such as a commitment or a prohibition. A sanction can be positive, negative, or neutral [Nardin et al., 2016]. A sanction may be in the form of "feedback," e.g., a smile or a scowl, from one user to another. An example sanction (S) is that, in a meeting, if a participant's phone rings loud, he or she receives a scowl from other meeting participants: S(phone-user, coworker, place = $meeting \wedge ring = loud$, feedback = *scowl*).

2.3 Arnor

Arnor is a four-step method build on social constructs to systematically model the social experience provided by a SIPA. Arnor's steps include modeling of: (1) goals, (2) environmental contexts, (3) social expectations, and (4) social experience. Figure 2.2 shows a conceptual model of Arnor. Table 2.1 provides an overview.

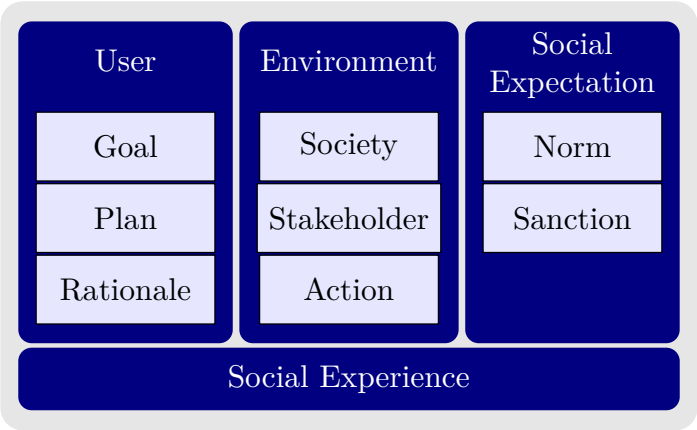

Figure 2.2 Arnor's conceptual model schematically.

Table 2.1 Overview of Arnor tasks and examples to engineer a SIPA.

Step	**Arnor Task**	**Example**
Goal Modeling	Identify all actors	Alice, Bob, Charlie, Dave, Erin, and strangers in the theater
	Abstract actors as primary and secondary stakeholders, as appropriate	Phone user is a primary stakeholder; friend, coworker, stranger in the vicinity of phone users are secondary stakeholders
	Identify goals of each actor	Phone user's goals *to be tele-reachable*, and *to be not disturbed*
	Identify all actions, and abstract them as appropriate	*Phone users do not answer phone calls during meetings*; *phone users answers their coworkers' urgent phone calls*
	Identify plans for abstract actions	*Set ringer mode as loud* for the action *phone user answers a phone call*
	Associate goals with plans	Phone user's goal of *tele-reachable* can be realized by the plan of *setting ringer mode as loud*
Context Modeling	Identify the contexts in which each actor's goals and plans are relevant	Coworker's goal *to be not disturbed* is relevant in the *meeting* context
	Identify conflicting goals (and inconsistent plans)	Phone user's goal of *tele-reachable* conflicts with the goal *to not disturb neighbors* in the *meeting* context
Social Expectation Modeling	Identify norms relevant to social and privacy expectations	*The phone user is committed to answering urgent phone calls from family*
	Identify possible conflicts between norms	*Phone user's commitment toward friend to answer phone calls* conflicts with *phone user's commitment to keep phone on silent during meeting*

Continue on the next page

Table 2.1 Overview of Arnor tasks and examples to engineer a SIPA (continued).

Step	Arnor Task	Example
	Resolve conflicts by capturing contextual preferences between norms	In the *meeting* context, prefer *phone user's commitment to keep phone on silent during meeting* over *phone user's commitment toward friend to answer phone calls*
Social Experience Modeling	Identify effects of stakeholders' actions on social expectations	A norm that is consistently being violated, e.g., *phone users always answering calls during meeting*
	Promote actions that enhance social experience	

2.3.1 Goal Modeling

For a SIPA to provide a social experience, it needs to be aware of the associated stakeholders, their goals and relevant plans. Goal modeling in Arnor uses Tropos constructs to elicit stakeholders, their goals, and relevant plans.

A stakeholder is a user that participates in a society and interacts with or is affected by the SIPA. *Primary* stakeholders are the users that interact directly with the SIPA. *Secondary* stakeholders do not have direct interaction with the SIPA, but are affected by its interactions with the primary stakeholder.

A goal is a set of states of the environment that are preferred by the stakeholders.

A plan is a sequence of actions that can bring about a state in which a stakeholder's goal is satisfied. The SIPA acts on behalf of the stakeholders or assists stakeholders in bringing their goals.

Stakeholders in Arnor map to actors in Tropos or Xipho. Whereas Tropos and Xipho explicitly relate actors to the users that have goals, Arnor forces designers to additionally identify (secondary) stakeholders that do not necessarily have a goal, but are affected by the plans that (primary) stakeholders execute to achieve their goals. Capturing secondary stakeholders is necessary to providing a social experience. A stakeholder may adopt different roles.

Following Table 2.1, we create the goal model for the ringer manager SIPA described in Examples 2–4 and Figure 2.1.

Primary stakeholder. Alice, the phone user (S_1).

Secondary stakeholders. Bob (Alice's friend, S_2), Charlie and Dave (Alice's coworkers, S_3 and S_4), Erin (Alice's mother, S_5) and strangers (those in the theater who are in Alice's vicinity when the ringer manager SIPA is in use, S_6). Here Bob, Charlie, Dave and Erin could assume the roles of caller and neighbors in different contexts. Note that, although the ringer manager SIPA includes only one primary stakeholder, other settings could involve multiple primary stakeholders.

Goals. The phone user's goals are to be tele-reachable (G_1), to notify caller if not reachable (G_2), to work uninterrupted (G_3), and to avoid annoying neighbors (G_4). Bob, Alice's friend has goals to (1) tele-reach Alice (corresponds to G_1), and (2) be notified if Alice is not reachable (corresponds to G_2). Charlie and Dave's goals are to not be disturbed at work by anyone (same as G_4). Erin's mother has the same goals as Bob. Strangers in Alice's vicinity share the same goal as Charlie and Dave. When Charlie and Dave assume the caller role, they share Bob and Erin's goal of tele-reaching Alice.

Actions. Alice, the phone user, can answer a call if she is available, or can notify the caller otherwise. She could decide not to answer calls if she does not want to be disturbed or does not want to annoy her neighbors. Based on Alice's actions, Bob, Charlie, Dave, Erin, and other stakeholders act. For example, if Alice answers Bob's or Erin's call, they could give Alice a positive feedback. In social expectation modeling, we capture these feedback actions as sanctions.

Plans. The plan corresponding to the *answer call* action is to *set ringer mode on loud* (P_1). The other plans could be to *set ringer mode on vibrate* (P_2) or *set ringer mode on silent* (P_3).

Goal-plan association. The plan of setting the ringer on loud promotes the phone user's goal of being tele-reachable, and caller's goal of tele-reaching the callee. The plan of setting the ringer on silent promotes the phone user's goal to work uninterrupted, and the neighbors' goal of not being disturbed.

2.3.2 Social Context Modeling

Context modeling includes identifying social contexts in which the stakeholders of a SIPA interact. The social context could include the place where the interaction occurs, attributes

of the place, neighbors in the vicinity, the social relationship between primary and secondary stakeholders, the activities the stakeholders are involved in, and so on. The social context is decisive in identifying the goals to be brought about or plans to be executed in case of conflicts.

Some of the contexts associated with goals, G_1–G_4, and plans, P_1–P_3, are based on stakeholders' locations (meeting or theater), social relationship (colleagues, friends or family), reason associated with a phone call (urgent phone call or a casual phone call), and so on.

Goal G_1 of being tele-reachable conflicts with goals G_3 and G_4 for both the meeting and theater scenarios. In these scenarios, the SIPA must rely on social contexts to determine which goal to accomplish. Potentially, where multiple plans may help realize the same goals. For example, in a library, both the *phone on silent* plan and *phone on vibrate* plans serve the goal of not disturbing one's neighbors. The SIPA relies on social context to choose between multiple plans.

2.3.3 Social Expectation Modeling

Social expectations including the privacy ones influence the stakeholders' goals and plans. We model these expectations between stakeholders in terms of social norms and sanctions. The social norms of a society regulate how stakeholders act and conduct themselves. Some norms could be local to a stakeholder, for example, one's commitment toward family members to always answer their phone calls, and some norms could be specific to a social context, for example, in the context of a meeting, a phone user is committed to keep his or her phone silent.

We express social expectations for the ringer manager SIPA via norms, sanctions and conflicts.

Norms. We identify the following norms.

- A phone user is committed to answering phone calls from callers. This commitment is satisfied by the plan of setting the ringer mode on loud.

  C_{caller}: $\mathsf{C}(\text{phone-user}, \text{caller}, \text{call}, \text{ring} = \textit{loud})$

- A phone user is committed to notifying the caller if he or she does not answer. The commitment is satisfied by the plan of setting the ringer mode on silent and sending a notification to the caller.

  C_{notify}: $\mathsf{C}(\text{phone-user}, \text{caller}, \text{call},$
  $\text{ring} = \textit{silent} \wedge \textit{notify})$

- A phone user is committed to coworkers to not let the phone ring during meetings. This commitment is satisfied by the plan of setting the ringer mode on silent or vibrate.

$C_{meeting}$: C(phone-user, coworkers, call,
ring = *silent* $\vee$ ring = *vibrate*)

Sanctions. The associated sanctions are as below:

- A phone user is (negatively) sanctioned by coworkers for answering a phone call during a meeting.
  $S_{meeting}$: C(phone-user, coworkers, call
  $\wedge$ place = *meeting* $\wedge$ ring = *loud*, feedback = *negative*)

Conflicts. If a caller calls the phone user during a meeting, the phone user's commitment C_{caller} toward a caller conflicts with his or her commitment $C_{meeting}$ toward coworkers to not answer phone calls during meetings, i.e.,
conflict$(C_{caller}, C_{meeting})$.

Conflicts in social expectations can be resolved by capturing contextual preferences between conflicting norms. For example, a phone user can have a preference of $C_{meeting}$ (*keep phone on silent during meetings*) to C_{caller} (*answer calls from family members*).

2.3.4 Social Experience Modeling

Norms are satisfied or violated as stakeholders act and execute plans to achieve their goals. Norm satisfaction or violation provides positive or negative experience to the stakeholders. As agents derive social experience from norms, over time, certain norms are preferred over others, and some lose significance. If a certain phone user is always answering phone calls during meetings, the phone user could be banished from meetings. A SIPA should execute actions that promote yield social experience by choosing which plans to execute, which goal states to accomplish, and which norms to satisfy. To decide which actions to promote, SIPAs could employ argumentation [Bench-Capon and Dunne, 2007], and make use of argumentation schemes such as *arguments from consequences*, and *arguments from popular opinion* [Walton et al., 2008]. Additionally, a SIPA, depending upon its user's privacy attitude and information sharing preferences, can choose to share its decision rationale for choosing an action with the other stakeholders. The sharing of rationale could introduce nuances in social relationships of a SIPA's stakeholders such as increase of trust that we do not model.

2.4 Evaluation

We investigate our research question by evaluating Arnor via a developer study and a simulation experiment.

2.4.1 Developer Study

We begin with a multiphase developer study in which participants develop ringer manager SIPAs. Our study was approved by the Institutional Review Board (IRB). We obtained informed consent from each participant. The developer study lasted for six weeks.

Study Unit

The study unit is a ringer manager SIPA discussed in Examples 2–4 and Figure 2.1.

Participants

The developer study involved 30 participants, enrolled in a graduate-level computer science course. The participants earned points toward their course grades for completing the tasks described. However, participation in the study was not mandatory. Nonparticipants were offered an alternative task to earn points equivalent to what they would earn by participating in the study.

Study Mechanics

This developer study has two phases: learning and development. The study follows the one-factor design with two alternatives (Arnor and Xipho). We use Xipho as our baseline method because it is best suited among the existing AOSE methods to engineer personal agents.

We split participants into two groups (A that follows Arnor, and X that follows Xipho) balanced on skills indicated in a presurvey (detailed in Appendix A.1). All participants develop a ringer manager SIPA.

Learning Phase. During the learning phase of the study, participants proposed a SIPA, and created models of the proposed SIPA. This phase sought to help participants understand the nuances of a SIPA, and to teach them how to model requirements. The data collected in the learning phase is not used in the evaluation.

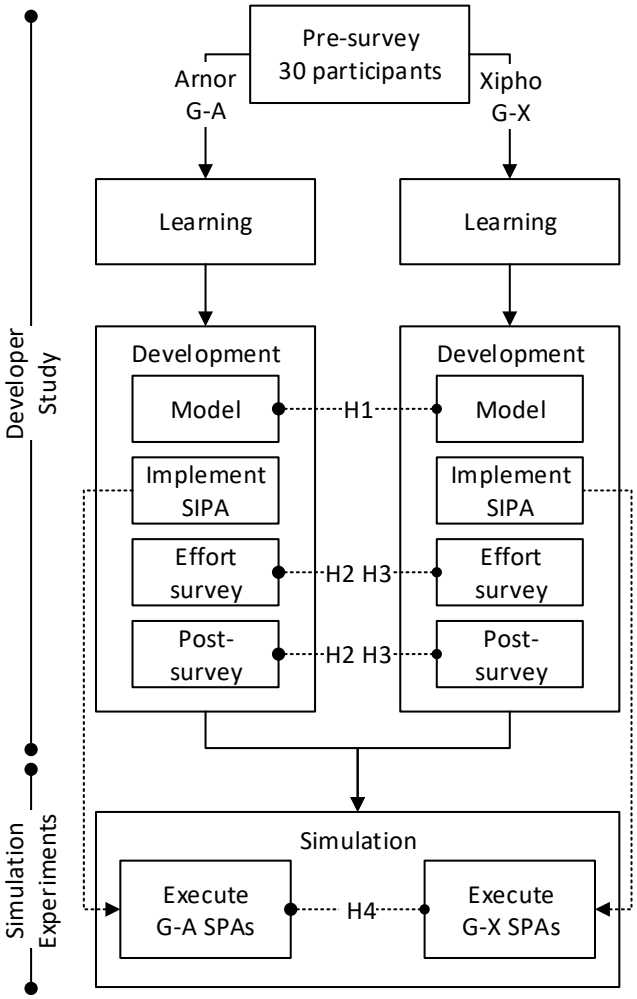

Figure 2.3 Experimental design.

Development Phase. In the development phase, participants modeled and implemented a ringer manager SIPA that adapts according to expectations of callers and neighbors, and sanctions received from callers and neighbors for each action.

In the two development phases, participants were provided with a testbed to verify the working of their SIPAs.

Deliverables

The participants submitted models and source code at the completion of the development phase. Additionally, the participants completed a time and effort survey (detailed in Appendix A.2) for each work session, and completed a post-phase survey (detailed in Appendix A.3) at the end of each phase.

Metrics

To measure the effectiveness of Arnor, we compute the following metrics.

Model coverage measures the completeness of the model. It is the ratio of the number of requirements identified correctly in the produced model to the total number of requirements of the SIPA. Higher is better.

Model correctness measures how correct the model is. It is the ratio of the number of correctly identified requirements to the total number of requirements of the SIPA identified. Higher is better.

Model quality is the product of model coverage and model correctness. Higher is better.

Time to develop is the actual time spent by participants in hours to develop the SIPA. Lower is better.

Difficulty of development is the subjective rating by participants on how easy it is to develop the SIPA on a Likert scale of 1 (very easy) to 7 (very difficult). Lower is better.

Effort to develop is the product of time spent in hours and ease of development rating for each work session. Lower is better.

Hypotheses

We consider the following hypotheses.

$\mathbf{H_1}$. Developers who follow Arnor produce better quality models than those who follow Xipho.

$\mathbf{H_2}$. Developers who follow Arnor spend less time to develop a SIPA, than those who follow Xipho.

$\mathbf{H_3}$. Developers who follow Arnor feel it is easier to develop a SIPA, than those who follow Xipho.

H_4. Developers who follow Arnor expend less effort to develop a SIPA, than those who follow Xipho.

Threats and Mitigation

We mitigated three main threats to our studies. Differences amongst participants' programming and modeling skills are inevitable. To handle the skill differences between participants, we surveyed participants about their educational backgrounds and prior experiences with programming and conceptual modeling. We balanced the two groups based on the survey. To mitigate the risk of participants' failing to report information, participants were instructed to complete a time and effort survey after each work session, while it was fresh in their minds. Communication between participants of different groups is yet another threat. To mitigate the risk of contamination, we created separate message boards for each participant group, and restricted participants to only posting clarification questions on the group message boards.

2.4.2 Simulation Experiments

We further investigate our research question via simulation experiments. We execute the ringer manager SIPAs implemented by third-party developers (as part of the aforementioned developer study) on a testbed fabricated to simulate different real-world environments.

Ringer adaptation scenarios

To test runtime adaptability, we test the applications for multiple iterations of incoming phone calls during a meeting.

Norms fixed. The meeting room participants are committed to keeping their phones silent.

Change in norms. The meeting room participants are initially committed to keeping their phones silent, but later the commitment expires.

Change in context. The meeting room participants are always committed to keeping their phones silent. Initially there are several participants in the meeting, but later all but two leave the meeting.

Change in sanction. The meeting room participants are always committed to keeping their phones silent. Initially they give negative feedbacks for loud ringing but later give more neutral feedbacks.

Metrics

To measure social experience, we compute the following social metrics in each of the above adaptation scenarios.

Adaptability coverage measures the completeness of code for adaptability requirements. It is the ratio of the number of adaptability requirements implemented correctly to the total number of adaptability requirements. Higher is better.

Adaptability correctness measures the correctness of the code for adaptability requirements. It is the ratio of the number of correctly implemented adaptability requirements to the total number of adaptability requirements implemented. Higher is better.

Norm compliance refers to the proportion of norm instances that are satisfied. Higher is better.

Sanction proportion measures the percentage of sanctions imposed. Lower is better.

Hypotheses

We consider these additional hypotheses:

H_5. SIPAs developed using Arnor yields better adaptability than SIPAs developed using Xipho.

H_6. SIPAs developed using Arnor provide a richer social experience than SIPAs developed using Xipho.

We use adaptability coverage and correctness to test hypothesis H_5, and use norm compliance and sanction proportion measures to test hypothesis H_6.

2.5 Results

We analyze deliverables produced by participants at the end of each phase, and compute the study parameters for each deliverable.

2.5.1 Developer Study

To test hypothesis H_1, we compare the models produced by Groups A and X. For hypothesis H_2, we compare the development time expended by Groups A and X during the two development phases. For hypothesis H_3, we compare the ease of development ratings reported by Groups A

and X during the two development phases, and for hypothesis H_4, we compare their expended effort.

Model quality. We evaluated models produced by the participants for correctness and coverage, and computed a quality metric. We found no significant difference in model quality.

Time and effort to develop. We found that average time (13.27 hours) and effort (61.54) expended by the participants using Arnor to be lower than average time (17.72 hours) and effort (96.6) expended by the participants using Xipho. Figures 2.4 and 2.5 show the boxplots for time and effort expended by participants using Arnor and Xipho to develop the social ringer SIPA.

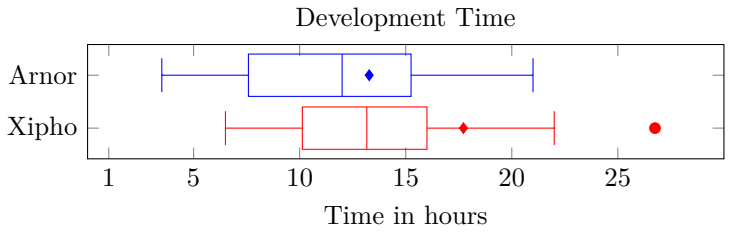

Figure 2.4 Arnor vs. Xipho's development time in hours as reported in the work session surveys.

Difficulty of development. The participants using Arnor found it easier to develop SIPAs with Arnor, compared to participants using Xipho. Figure 2.6 shows the difficulty of development boxplots.

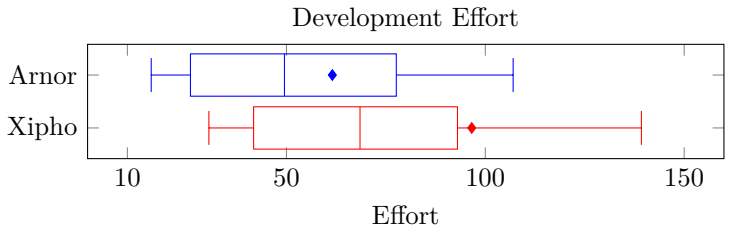

Figure 2.5 Arnor vs. Xipho's development effort as reported in the work session surveys.

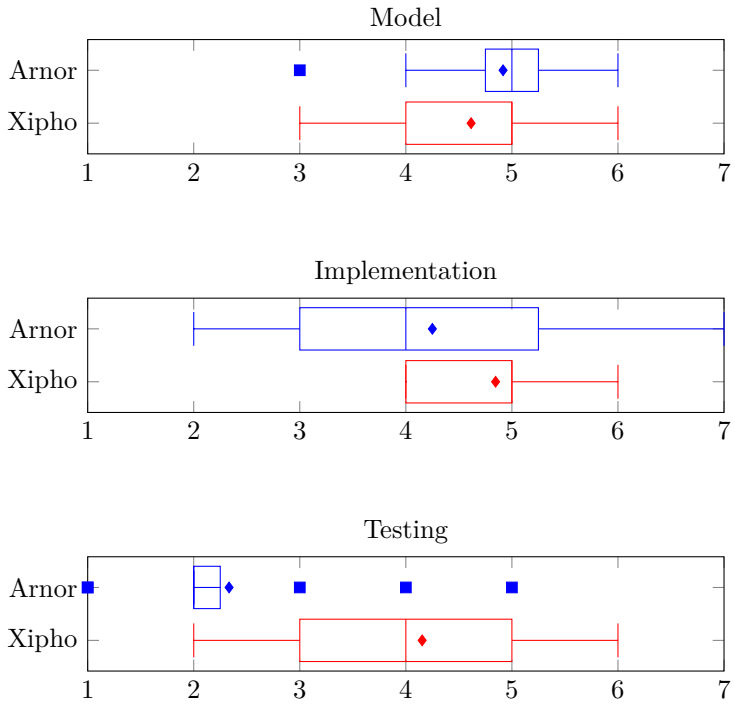

Figure 2.6 Arnor vs. Xipho's difficulty of development on a Likert scale of 1 (very easy) to 7 (very difficult).

2.5.2 Simulation Experiments

To evaluate H_5 and H_6, we analyzed the SIPA's implementation code and executed the SIPAs in diverse scenarios. We compare the execution results of Arnor and Xipho groups.

Adaptability features. We found average adaptability coverage (80%) to be the same for SIPAs developed by the Arnor and Xipho groups. This result could be attributed to the limited time we gave the participants to develop the SIPA. Average adaptability correctness was found to be higher for Arnor (100%) compared to the Xipho (95%). This gain could be attributed to the systematic steps provided by Arnor to engineer SIPAs.

Norm compliance. Figure 2.7 shows line plots for norm compliance in the four ringer adaptation scenarios. Though the average norm compliance values for SIPAs developed using Arnor and Xipho are mostly similar, Arnor performs slightly better in the fixed norms scenario.

Sanction proportion. Figure 2.8 shows the plots for sanction proportion in the four adaptation scenarios. For the first three scenarios (norms fixed, norms change, and context change), the

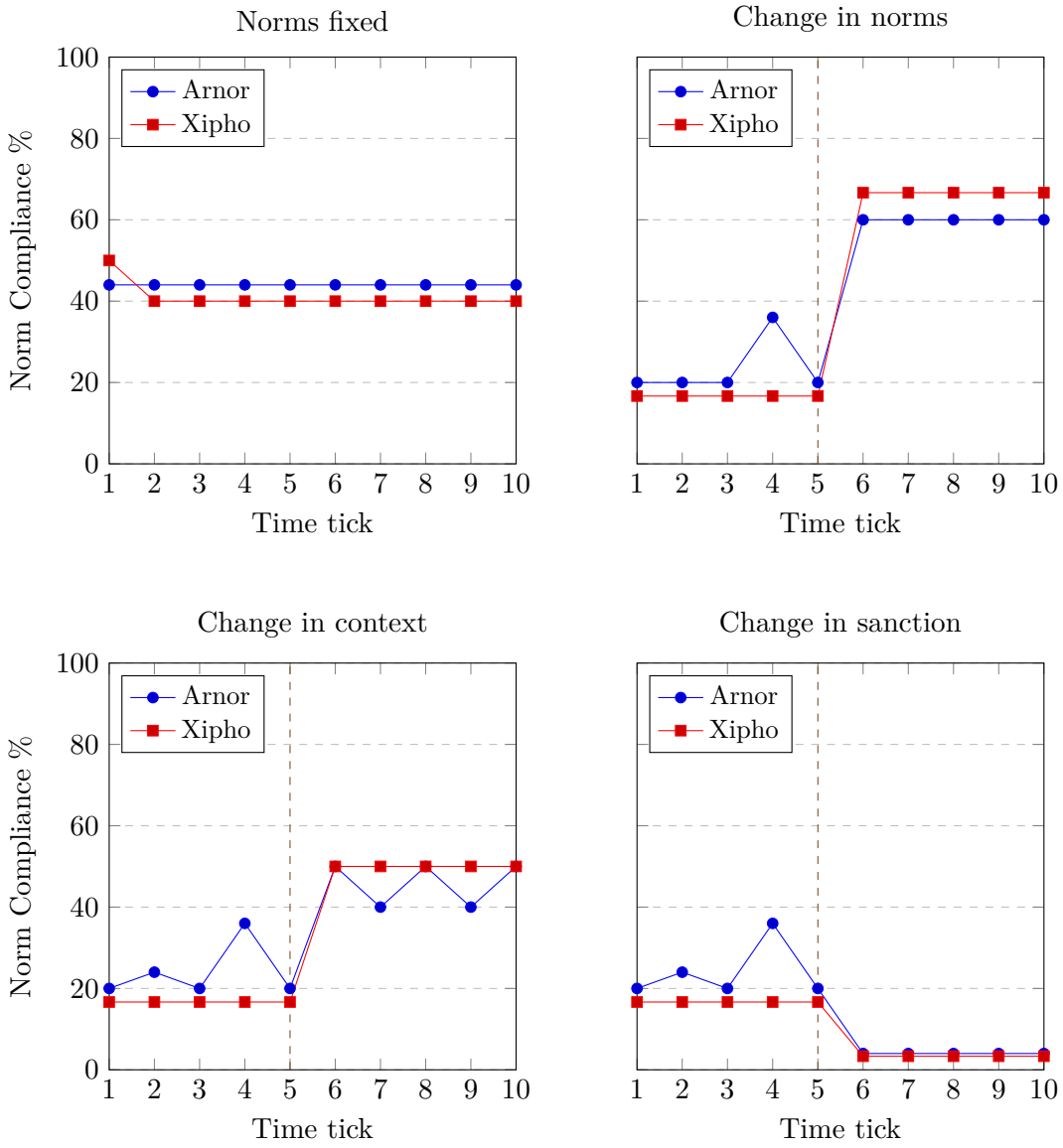

Figure 2.7 Arnor vs. Xipho's norm compliance.

SIPAs developed using Arnor have a lower sanction proportion. For the sanction change adaptation scenario, the SIPAs developed using Arnor take slightly longer to adapt, and only have a slightly higher sanction proportion than the SIPAs developed using Xipho.

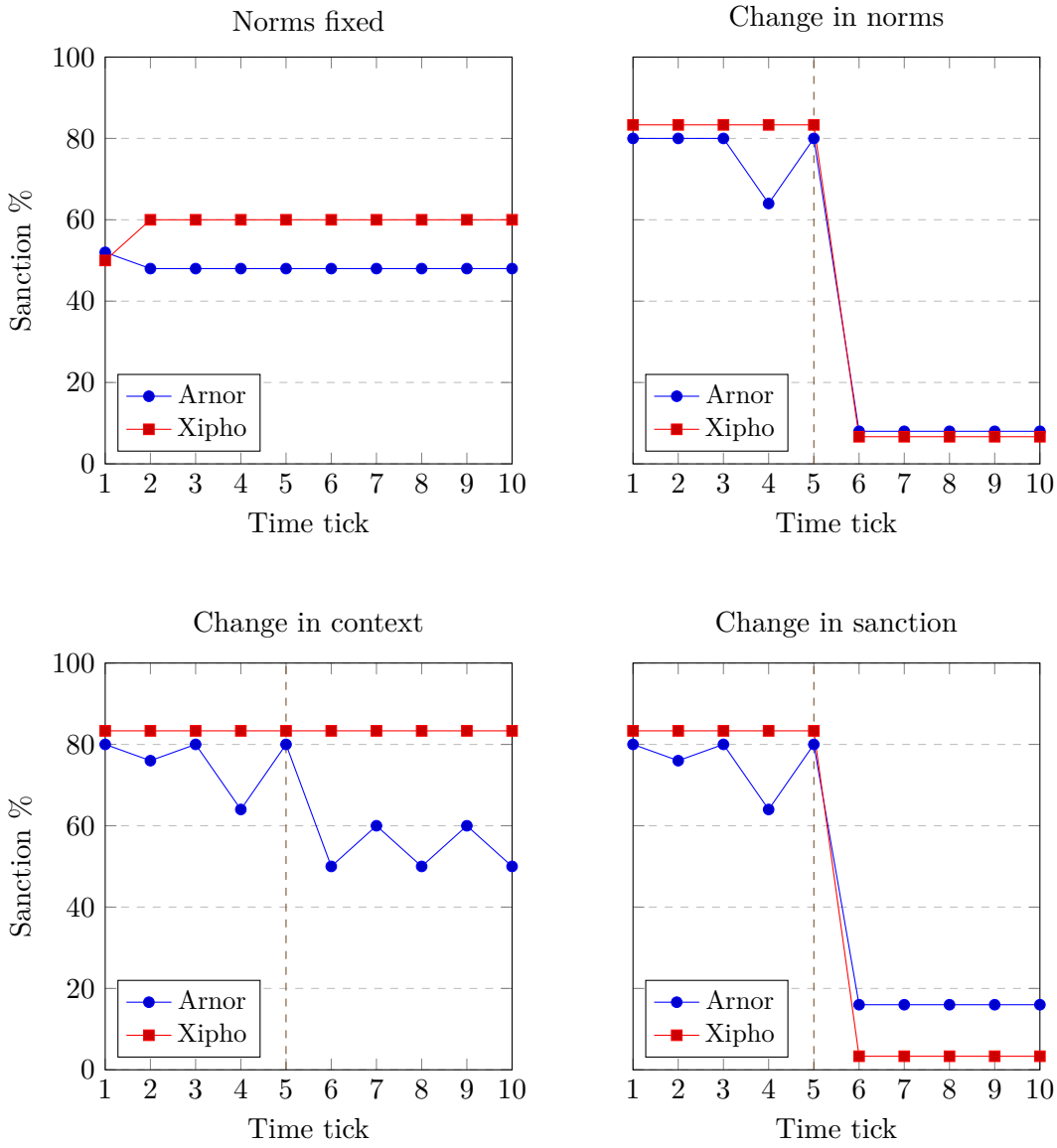

Figure 2.8 Arnor vs. Xipho's sanction proportion.

2.5.3 Threats to Validity

In the developer study, we mitigated the threats of skills difference, participants' failure to report information, and the risk of contamination. However, some threats remain.

First, our results are based only on the development of a single SIPA (ringer). For conclusive

results on the effectiveness of Arnor, future studies may require participants to develop more than one kind of SIPA.

Second, the SIPAs developed by the study participants mostly reflect the participants' (developers) privacy attitudes and information sharing preferences. To generalize our results, it is required to collect real data on SIPA users' privacy attitudes and information sharing preferences.

Third, in simulation experiments, we tested runtime adaptability of SIPAs under diverse, but a limited set of scenarios. The scenarios we incorporated may not represent all real world scenarios in which a ringer SIPA would be employed.

Collecting real data about users' attitudes, preferences, and contexts is essential, though nontrivial, to mitigate the second and third threat. Crowdsourcing is a promising avenue for future studies to collect such data at a large scale.

2.6 Related Works

Ali et al. [2013] propose an AOSE-based contextual requirements engineering framework, with a focus on consistency and conflict analysis. Arnor goes beyond conflict analysis, and promotes goals, plans, and norms that promote greater social experience. Rahwan et al. [2006] propose a framework to integrate goal models and social models. Arnor models subsume social models, and provide richer abstractions to capture agents' interactions and affects on experience.

Sugawara [2011] attempt to resolve conflicts through reinforcement learning. Mashayekhi et al. [2016] propose a hybrid mechanism to monitor interactions and recommend norms to resolve conflicts. Mihaylov et al. [2014] study convergence and propose a decentralized approach based on strategies in game theory. Villatoro et al. [2013] introduce social instruments to facilitate norm emergence via social learning. Yu et al. [2013] study norm emergence through collective learning from local interactions, and find that collective learning is superior to pairwise learning. Arnor provides constructs to engineer socially adaptable SIPAs that can make use of these approaches for norm emergence.

Hao et al. [2016] propose a lightweight formal method to design normative systems, which uses Alloy modeling language and analyzer to synthesize and refine norms. Van Riemsdijk et al. [2015a] propose a semantic norm compliance framework for socially adaptive agents. They use LTL to express norms. Agents in van Riemsdijk et al.'s framework identify and adopt new norms, and determine execution mechanisms to comply to these norms. Aldewereld et al. [2016] present a formalism for group norms, and provide mechanisms to reason about these norms. Ajmeri

et al. [2016a] propose Coco, a formalism to express and reason about conflicting commitment instances at runtime, and dominance among them. Coco employs Answer Set Programming to compute the nondominated commitment instances and determines compliance of actions with nondominated commitment instances. These formalisms could use Arnor's social constructs to assist SIPAs in compliance, adoption of new norm, and resolution of conflicts amongst norms at runtime.

2.7 Conclusion and Future Directions

We advance the science of privacy by tackling nuanced notions of privacy, including intrusion, disapprobation, and information leakage, in personal agents. We treat respecting stakeholders' privacy as an inherent aspect of delivering a social experience. We envision socially intelligent personal agents that (1) adapt to the social contexts of their stakeholders; and (2) act and interact in their best interest (not just the primary stakeholder).

We develop Arnor, a method that provides social constructs to engineer privacy-aware social agents. We demonstrate the method via a ringer manager SIPA. We evaluate Arnor using a developer study and simulation experiments. Compared to Xipho, we find that Arnor (1) facilitates faster development of SIPAs; and (2) yields SIPAs of higher quality, higher adaptability correctness, lower sanction proportion, and similar adaptability coverage and norm compliance. These observations suggest that Arnor promotes SIPAs to deliver a rich social experience.

Future Directions

Ferreira et al. [2013] propose a computational model for emotional agents that considers norms, social relations, roles and socially acceptable behaviors in a given context. Sollenberger and Singh [2011] introduce Kokomo to develop affective applications, and provide a middleware for building such applications. Incorporating an affective [Sollenberger and Singh, 2011] and emotional basis of norms in social agents is an interesting future direction. Modeling affect could assist SIPAs learn contextually relevant norms. A middleware implementation of Arnor could facilitate development.

Fogués et al. [2017b] study how context, users' preferences, and arguments influence a sharing decision in a multiuser privacy scenario. They collect data about appropriate sharing policies for a variety of multiuser scenarios from human participants in a large scale study. We conjecture that such data can be used to seed SIPAs with an initial set of norms, which the SIPAs can

evolve once put to use.

Chapter 3

Understanding and Reasoning about Social Context

This chapter addresses our second research question, RQ_2 on social context. The chapter describes Poros, our approach for building personal agents that carry out enriched interactions where deviating agents share selected elements of their contexts, and others respond appropriately, and its empirical evaluation via simulation experiments. The chapter is based on a paper "Robust Norm Emergence by Revealing and Reasoning about Context: Socially Intelligent Agents for Enhancing Privacy" that appears in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

3.1 Introduction

Social *norms* provide a robust means to regulate interactions in human society. Our everyday actions tend to *comply* with social norms. For example, *ignoring a phone call during a meeting* and *remaining silent in a public library* are expected behaviors that accord with social norms. However, we often *deviate* from the applicable social norms, for instance, when *stepping out of a meeting to answer a phone call*.

The ability to deviate from norms is crucial for autonomy. We may *sanction* each other based on how we are interacting. In particular, negative sanctions in response to deviations are a means for establishing norms [Andrighetto et al., 2013]. For example, when a meeting attendee's phone rings, a *scowl* on other attendees' faces hints at a norm of *keeping one's phone silent during meetings*.

Existing approaches for norms provide simplified interactions: a deviation or not, followed by a sanction or not. But real-life interactions are more complex. Whether a deviation leads to a positive or negative sanction depends on how others perceive its *context* or circumstances of occurrence. When we deviate from a norm, we may offer an apology, describing the context. One, revealing context may soften a deviation and help avert negative sanctions. Suppose, upon receiving a call during a meeting, Alice says that the call was from her sick father. As a result, the meeting attendees may excuse Alice for taking the call. A deviation may result in a positive sanction. For instance, a physician who reveals a patient's private data to save the patient's life would receive a positive sanction despite violating a norm. Even in the phone call setting, a positive sanction may ensue for deviating from a norm. For example, a user who hesitantly takes a call from his nine-month pregnant wife during a lab meeting would generally receive positive comments from coworkers. Two, context helps refine the relevant norms. For example, Alice's revelation may help refine the norm from *ignoring a phone call during a meeting* to *ignoring a phone call during a meeting, unless the call is urgent*. In essence, deviation context and any ensuing sanction help characterize the boundaries of a norm in play.

Accordingly, we propose Poros, an approach for building agents that carry out enriched interactions where deviating agents share selected elements of their contexts, and others respond appropriately. A socially intelligent personal agent (SIPA) is an agent who acts in accordance with (but may deviate from) social norms [Ajmeri et al., 2017b]. We imagine an artificial agent society in which SIPAs of three main types act and interact on behalf of (human) users, as a basis for empirically investigating the emergence and quality of norms.

This research applies in developing privacy-supporting SIPAs. Norms provide a basis for understanding privacy [Nissenbaum, 2011]. Regulations about information disclosure, as in healthcare, are context-dependent norms [Ajmeri et al., 2016a], as are social practices. Privacy involves control over when and what information to disclose [Westin, 1967]. In some construals, actions that intrude upon one's solitude or bring disapprobation are privacy violations. In essence, all privacy-relevant interactions are modulated by norms. Therefore, social intelligence in making decisions cognizant of norms while preserving social cohesion is crucial.

Our main contribution is to study two research questions in light of a specific decision by a SIPA, namely, whether to reveal its context to others when it deviates from a norm:

Q_1 Norm: Does revealing context and reasoning about revealed context promote emergence of robust social norms?

Q_2 Goal: Does acting in accordance to such robust norms result in an improved goal satisfaction?

Our results show that (1) norms that emerge in Poros are robust, implying improved social cohesion and (2) SIPAs yield higher goal satisfaction to their users when acting in Poros than when acting in a conventional setting (just sanctions).

3.2 Related Work

Research on normative systems has addressed the problems of conflict, compliance, and emergence of norms. We sample some of the literature from the following themes.

Social norms regulate agent interactions by characterizing what behavior one agent may legitimately expect from another in a particular setting [Kafalı et al., 2016; Singh, 2013]. We adopt Singh's [2013] computational representation of social norms. A norm is directed from a subject (stakeholder) to an object (stakeholder), and is constructed as a conditional relationship involving an antecedent (which brings the norm into force) and a consequent (which brings the norm to satisfaction or violation). Ajmeri et al. [2017b] introduce Arnor, a method to model social intelligence in personal agents. They argue that personal agents who understand the intricacies of social norms, deviations, and associated arguments can provide a privacy-preserving social experience to their users.

Works on designing *context-aware* agents emphasize modeling [Murukannaiah and Singh, 2014] and sharing [Ajmeri et al., 2017b]. Poros is novel in the way it helps SIPAs infer social norms by revealing deviation context and reasoning about context revealed by others. Poros examines the effect of revealing context by agents after norm deviations. Kökciyan and Yolum [2017] propose an argumentation-based approach to enable agents to reason about context and reveal information based on it. Whereas their focus is on understanding the context to make a privacy decision, we demonstrate the benefits of revealing context. Naively revealing context could violate user privacy. However, a SIPA would reveal selectively by evaluating the tradeoff between privacy lost by revealing and sanctioning faced by not revealing. (For simplicity, in our experiments, the context model is simple and the SIPAs always reveal—to demonstrate the benefit of revelation.)

The study of *norm conflicts and compliance* has drawn much interest. An agent may face conflicts between multiple applicable norms [Ajmeri et al., 2016a], or between norms and its own goals. Van Riemsdijk et al. [2015a] develop a norm compliance framework to design socially adaptive agents in which agents identify and adopt new norms, and determine execution mechanisms to comply with those norms. Van Riemsdijk et al. argue that a personal agent needs explicit norms. Aldewereld et al. [2016] present a formalism and mechanism to comply with

group norms. Ajmeri et al. [2016a] present a formalism to represent normative conflicts and dominance relationships among conflicting norms. Sugawara [2011] uses reinforcement learning to resolve norm conflicts and shows how social conventions for resolving conflicts emerge. However, the efficiency and stability of the results differ across agents. These works give us insights into defining agents' decision-making processes.

Agent interactions lead to dynamic *norm emergence and evolution* [Savarimuthu et al., 2009]. Boella et al. [2009a] propose a normative framework to evaluate and classify normative system change. Mashayekhi et al. [2016] propose a hybrid mechanism for norm emergence and conflict resolution in sociotechnical systems. Villatoro et al. [2013] present social instruments such as "rewiring" and "observation" to assist norm emergence. Yu et al. [2013] suggest using collective, instead of pairwise, learning for norm emergence. Poros is novel in that it supports revealing and reasoning about contextual information to facilitate understanding of contextually relevant norms.

Sanctions are mechanisms to achieve social coherence. An agent decides whether to comply with or deviate from a norm. A sanction, negative or positive, is associated with the reaction of other agents to this decision. Previous works adopt sanctions as a way to promote norm compliance [Egas and Riedl, 2008; Noussair and Tucker, 2005]. Alechina et al. [2012] present a programming language for norm-aware agents who might deviate from norms and expect sanctions. Nardin et al. [2016] develop a sanction typology and introduce a conceptual sanctioning process model to promote governance in sociotechnical systems. Recent works explore combining norm communication with sanctions to promote cooperation [Andrighetto et al., 2013]. Van Riemsdijk et al. [2015a] emphasize understanding norm violations as a basis for designing socially adaptive agents. Poros differs from these works in addressing the problem of understanding a deviation by modeling the context in which a deviation occurs.

3.3 Interaction in a SIPA Society

A SIPA society we seek to engineer consists of stakeholders, a social architecture, and SIPAs acting on behalf of stakeholders. Figure 3.1 shows a conceptual model of a SIPA society.

The stakeholders are users, *primary* or *secondary*, depending on the context (defined later). The *primary* stakeholder of a SIPA is the user who directly interacts with it, and on whose behalf the SIPA acts and interacts. A *secondary* stakeholder is the user who may not directly interact with the SIPA, but is affected by the SIPA's actions [Ajmeri et al., 2017b]. Each stakeholder has goals and plans.

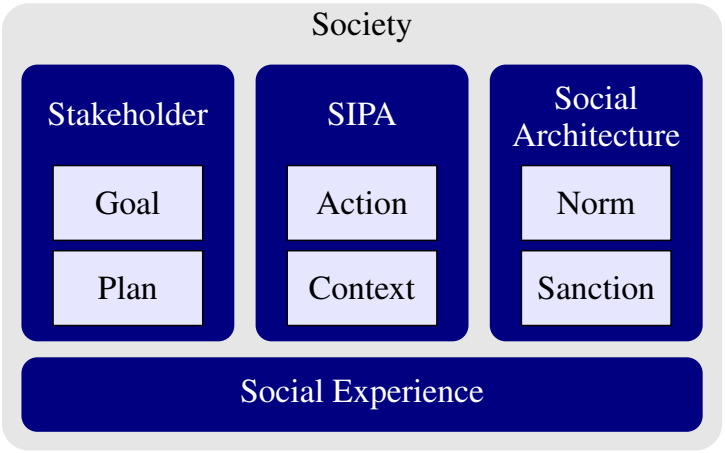

Figure 3.1 A society of SIPAs and stakeholders.

- A *goal* of a stakeholder describes a state the stakeholder would prefer; a stakeholder may have multiple goals.
- A *plan* of a stakeholder is a set of actions that can bring about one or more goals.

The social architecture of a society captures its structure; it comprises social norms and the sanctions that promote or ensure compliance with norms.

- A *norm* is a tuple of ⟨subject, object, antecedent, consequent, context⟩ [Singh, 2013]. Norms characterize the social architecture that promotes prosocial behavior.
- A *deviation* from a norm occurs when a stakeholder, or *deviant*, performs an action that does not comply with it.
- A *sanction* is a set of actions a stakeholder may take toward a deviant on observing a deviation. A sanction may be positive or negative [Nardin et al., 2016].

A SIPA acts and interacts on behalf of a stakeholder and is aware of the social architecture of the society.

- An *action* is a step a SIPA takes to execute its stakeholder's plan, thereby bringing about the corresponding goal. An action may satisfy or violate a norm. SIPAs in a society can observe each other's actions.
- A *context* captures the circumstances under which a SIPA acts [Dey, 2001]. In our approach, the context is social and incorporates whether a norm is satisfied or violated. Context includes social relationships between stakeholders and spatiotemporal parameters relevant to describing interactions between a SIPA and its stakeholders. We adopt Murukannaiah and Singh's [2012] notion of *place* as a location such as home, library, meeting, or party

understood in conceptual terms. Parameters describing a place may include physical conditions (e.g., noise level), expected activities (e.g., reading a book), social interactions (e.g., having a discussion), and temporal information (e.g., during office hours on a weekday).

The social experience a SIPA delivers reflects the extent to which the SIPA promotes its primary and secondary stakeholders' goals. It relates to how a SIPA's stakeholders perceive a norm deviation, and the sanctions they apply. Our objective is to promote each SIPA to act toward maximizing the overall social experience, despite competing interests.

We define social experience (E) as the weighted aggregation of payoffs perceived by a SIPA's stakeholders for each action executed by the SIPA. That is, for each potential action, a SIPA determines the payoffs for its primary and secondary stakeholders, and computes an aggregation as a weighted sum of the payoffs. A SIPA's aggregation method reflects its primary user's preferences and privacy attitudes. For instance, a pragmatic user's SIPA may aggregate payoffs by giving equal weight to all stakeholders, whereas a selfish user's SIPA may give a smaller weight to secondary stakeholders.

3.3.1 Poros Explained with an Example SIPA

We now describe Poros, a framework to build SIPAs, using Ringer, an example SIPA who answers or ignores phone calls on behalf of its primary stakeholder by ringing the phone or keeping it silent. Ringer is a privacy-enhancing technology that acts on behalf of its primary stakeholder; it determines when to allow intrusions, and when to risk being overheard in a phone call (and thus when to intrude on others' solitude).

Ringer's primary stakeholder is the *callee* with privacy goals of *being reachable by phone*, *to work uninterrupted*, and *to not disturb neighbors*. Ringer's secondary stakeholders are (1) a *caller* with the goal *to reach the callee*; and (2) a *neighbor* with a privacy goal *to not be disturbed*. Ringer observes other SIPAs' actions and potentially sanctions them based on their actions and the context as revealed by them.

Each SIPA in Poros maintains a history of interactions and the associated experience. The actual experience is determined after each interaction based on the revealed context and any resulting sanctions. The history helps a SIPA determine the action that would maximize its stakeholder's predicted social experience.

We define a SIPA's history (H) as a set of tuples $h_i = \langle c_i, g, p, N, s_i \rangle$, each of which describes an interaction i, including context c_i describing the circumstances in which goal g is brought about via plan p under a set of applicable norms N, and all resulting sanctions $\{s_i\}$. For Ringer,

c_i includes the places where the stakeholders are, their social relationships, and urgency of the incoming call.

Each SIPA maintains its history locally, and scans it when selecting a plan. In a conflict situation, SIPAs look up their history to predict social experience and decide which norms or goals to prefer over which others in a given context; thus infer contextually relevant norms.

A SIPA's behaviors include acting on behalf of its stakeholder, deciding whether to reveal its context, reasoning about the contexts revealed by others, and issuing sanctions to others. It does so based on knowledge of its context, its stakeholder's goals, associated plan, and applicable norms.

- *Plan selection.* A SIPA selects a plan (and its associated actions) that would achieve its primary stakeholder's goals. In the Ringer example, it selects to ring or keep silent for an incoming phone call. If more than one plan are available, from the history (if available) it identifies the one that maximizes the social experience, or chooses a random plan from the applicable plans with a small probability α.

- *Revealing context.* When a SIPA chooses and executes a plan, it might deviate from some applicable norms. It decides which norms to prefer in the current context and whether to reveal unobserved context to other SIPAs. For instance, if Ringer decides to prefer the *family norm—always answer calls from family* over the *meeting norm—never answer calls during meetings* by ringing during a meeting for an urgent phone call from a sick family member, it reveals the unobserved context, i.e., urgency of the call and the caller's sickness to other meeting attendees. Ideally, a SIPA should selectively reveal context to others according to its stakeholder's goals and privacy attitude.

- *Sanctions.* A SIPA observes other SIPAs' actions, and sanctions them when its stakeholder is affected by their actions. On receiving the context revealed by a deviating SIPA, the SIPA of an affected stakeholder evaluates whether the observed action would be norm compliant in the revealed context. In the Ringer example, *neighbors'* and *caller's* SIPAs decide whether they would ring for an urgent phone call from a sick family member during a meeting and accordingly sanction the *callee's* SIPA.

The complete interaction, including the selected plan and executed actions, observed and revealed context, applicable norms, and sanctions, is recorded in SIPAs history. As SIPAs interact by acting and evaluating actions for norm compliance from interaction history, they understand the boundaries of applicable norms in different contexts, and thus promote emergence of robust social norms.

Figure 3.2 summarizes the interaction, learning and inferring norms by revealing context in Poros.

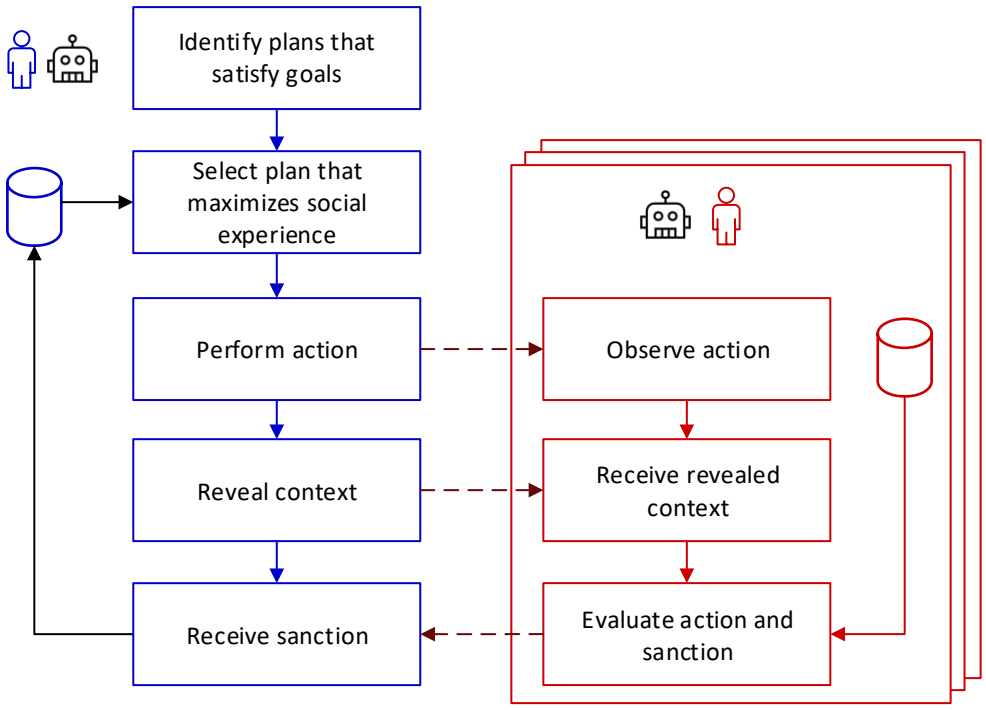

Figure 3.2 Interaction and inferring norms in Poros. Primary SIPA (in blue) performs action and reveals context, while secondary SIPAs (in red) observe action and receive context

3.4 Simulation Model

We evaluate Poros via a simulated *ringer environment* built using MASON [Luke et al., 2005].

3.4.1 The Ringer Environment

The ringer environment contains shared places (home, party, meeting, library, and emergency room). Corresponding to each place, we define social circles such as family, friends, and coworkers.

Each agent belongs to a family circle, a friend circle, and a coworker circle. Agents who do not share any of these circles are considered strangers. We define the social network or place network topology in a way such that there is only one type of relationship, i.e., family, friends, coworkers, or strangers, between any pair of agents. In the ringer environment, there are (1) several homes, each corresponding to a family circle, (2) several parties, corresponding to multiple friend circles, and (3) multiple meetings, corresponding to multiple colleague circles. There is one library and one emergency room (ER). The numbers of homes, parties, and meetings follow the network setups specified in Table 3.6.

In the simulation, agents stay at each place for a random number of steps (averaging 60 steps) and then move. If an agent enters home, party, or meeting, it is more likely to enter the place that is associated with its own social circle than entering a place with strangers. For example, if an agent chooses to enter home, it is likelier to enter its own family's home than to enter a stranger's home. Therefore, when it is at home, an agent is usually surrounded by its family members with only a few strangers.

The agents in the ringer environment perform the following actions depending upon their roles:

- A caller initiates an urgent or a casual phone call.
- A callee answers or ignores a phone call.
- A callee shares context for answering or ignoring a call.
- A caller and neighbors respectively reason about context.
- A caller and neighbors respectively sanction a callee for answering or ignoring a phone call.

Each place and each circle has predefined norms, as defined in Table 3.1. For example, emergency room (ER) is conceptualized as a place where the default norm is to always answer calls, whereas the norm in a library is to ignore calls. Norms could conflict. For example, the norm to *answer an urgent phone call from a family member* conflicts with *ignore during a meeting*. We let the agents figure out contextually relevant norms in case of conflict.

For each phone call, based on the callee's response of answering or ignoring, the caller, callee, and neighbors perceive a fixed payoff, as shown in Tables 3.2–3.4.

Table 3.1 Norms for answering calls based on (top) place and (bottom) caller's social circle and casual or urgent call types.

Norms by place

Place	Response
Emergency (ER)	Answer
Home (H)	Answer
Library (L)	Ignore
Meeting (M)	Ignore
Party (P)	Answer

Norms by circle and call type

Circle	Casual	Urgent
Coworker	Answer	Answer
Family	Answer	Answer
Friend	Answer	Answer
Stranger	Ignore	Answer

Table 3.2 Payoff for callee for casual or urgent call types.

Caller's Relationship	Callee's Response	Casual	Urgent
Family, Friend, or Coworker	Answer	0.50	1.00
	Ignore	0.00	–0.50
Stranger	Answer	0.00	0.50
	Ignore	0.25	–0.25

Table 3.3 Payoff for caller for casual or urgent call types.

Callee's Response	Casual	Urgent
Answer	0.50	1.00
Ignore	–0.50	–1.00

3.4.2 Agent Types

To evaluate effectiveness of Poros, we define two baseline agent types—*Fixed* and *Sanctioning*, other than Poros agents.

Fixed agents act according to the fixed set of norms listed in Table 3.1. If the norms conflict,

Table 3.4 Payoff for neighbors by place (ER, H, L, M, P).

Callee's Response	Emergency	Home	Library	Meeting	Party
Answer	1.00	0.67	–1.00	–1.00	–0.33
Ignore	–1.00	–0.33	1.00	1.00	0.67

Table 3.5 Payoff for a neighbor based on how callee acts and what the neighbor expects in the context revealed by callee.

Callee Action	Neighbor Expects	Emergency	Home	Library	Meeting	Party
Answer	Answer	1.00	0.67	1.00	1.00	0.67
Answer	Ignore	–1.00	–0.33	–1.00	–1.00	–0.33
Ignore	Answer	–1.00	–0.33	–1.00	–1.00	–0.33
Ignore	Ignore	1.00	0.67	1.00	1.00	0.67

the agents toss a fair coin to choose between alternative actions. If Fixed agents perceive an action as a deviation, they sanction the deviant.

Sanctioning agents infer social norms from sanctions [Andrighetto et al., 2013]. These agents start with the same strategy as Fixed agents. They continue to record the interaction history. Once they have gained enough number of records in their history of sanctions, they decide their subsequent actions based on history. In our simulation, this number is empirically selected so that an agent visits each scenario at least once. As callees, when norms conflict, they select the action that provides a higher payoff, computed according to Tables 3.2–3.4. As callers and neighbors, these agents sanction callees as per fixed norms listed in Table 3.1.

Poros agents infer social norms by revealing and reasoning about context. They start with the same strategy as Fixed agents following norms listed in Table 3.1. As callees, they reveal context, i.e., reveal the caller's relationship and the call's urgency to their neighbors, and reveal their place and neighbors' relationships to the caller. As neighbors or callers, they understand the callee's revealed context and decide what action they would have performed were they in that context, and sanction accordingly. Poros agents use Table 3.5's payoffs.

We employ a linear regression model over interaction history to choose actions based on sanctions by stakeholders.

3.5 Experiments and Results

We evaluate our research questions via multiple experiments on the ringer environment in which we simulate 1,000 or 250 Fixed, Sanctioning, and Poros agents in pragmatic, considerate, and selfish agent societies. The agents in societies use different schemes to aggregate payoffs. We run each simulation for 3,000 steps and compute the following metrics.

Social cohesion measures the proportion of agents that perceive actions as norm compliant. Higher the social cohesion, lower is the number of negative sanctions.

Social experience measures the goal satisfaction delivered by an agent, computed by aggregating payoffs for all stakeholders according to the payoff Tables 3.2, 3.3, 3.4, and 3.5.

To answer Q_1 on norms, we consider the following hypotheses pertaining to specified agent types. For brevity, we omit the corresponding null hypotheses indicating no gain. We test significance via the two-tailed paired t-test.

H$_1$ Poros yields greater *social cohesion* than Fixed.

H$_2$ Poros yields greater *social cohesion* than Sanctioning.

To answer Q_2 on goals, we consider these hypotheses:

H$_3$ Poros yields greater *social experience* than Fixed.

H$_4$ Poros yields greater *social experience* than Sanctioning.

3.5.1 Experiments with Pragmatic Agent Society and Varying Network Types

We simulate Fixed, Sanctioning, and Poros agents on four network types—large or small network with dense or sparse connectivity—as Table 3.6 describes. The society in this experiment is pragmatic in that the agents perceive social experience as the average payoff (equally weighted) for all stakeholders in an interaction. We summarize our results next.

Fixed agents. The average social experience was found to be between 0.53 and 0.56, and the social cohesion to be about 52% for the four network types.

Sanctioning agents. As expected, at around step 1,000 we see Sanctioning agents offer a rise in social experience over Fixed agents. The rise is gradual as the agents start to infer from

Table 3.6 Characteristics of network types studied.

Network Type	Agents	Circles		
		Family	Coworker	Friend
Large Dense	1,000	20	20	20
Large Sparse	1,000	100	100	100
Small Dense	250	5	5	5
Small Sparse	250	25	25	25

history. For the first 1,000 steps, the average social experience is the same as Fixed agents. It later stabilizes between 1.11 and 1.21 for all four networks. The social cohesion values were between 61.2% and 63.7%.

Poros agents. At around step 1,000, as agents acquire confidence, we see a significant increase in social experience offered by Poros agents. It stabilizes between 2.14 and 2.19 for the different networks. Social cohesion was found be significantly higher between 82.0% and 83.2%. For the first 1,000 steps, Poros agents yield the same average social experience as Fixed and Sanctioning agents.

Social cohesion and experience offered by Poros agents are significantly greater than those offered by Fixed and Sanctioning agents; thus the null hypotheses corresponding to H_1, H_2, H_3, and H_4 are rejected. Figure 3.3 shows the social experience plots indicating the results are consistent across the four network types. Table 3.7 summarizes the findings of the experiment with pragmatic agents. It shows stabilized values for social experience and social cohesion, and p-values from the two-tailed paired t-tests.

3.5.2 Experiment with Considerate Agent Society

We experiment with a considerate agent society where agents give a larger weight to their neighbors' payoffs than to their own payoffs when computing social experience and deciding the actions to perform when norms conflict. These agents continue to sanction based on their history.

Figure 3.4 shows the social experience for considerate Sanctioning and Poros agents in a Small-Dense network. The average social experience drops for Sanctioning and Poros agents after they have gained enough confidence. We attribute this decline to the fact that these agents value the neighbors' experience more than their own, and thus ignore calls they should have

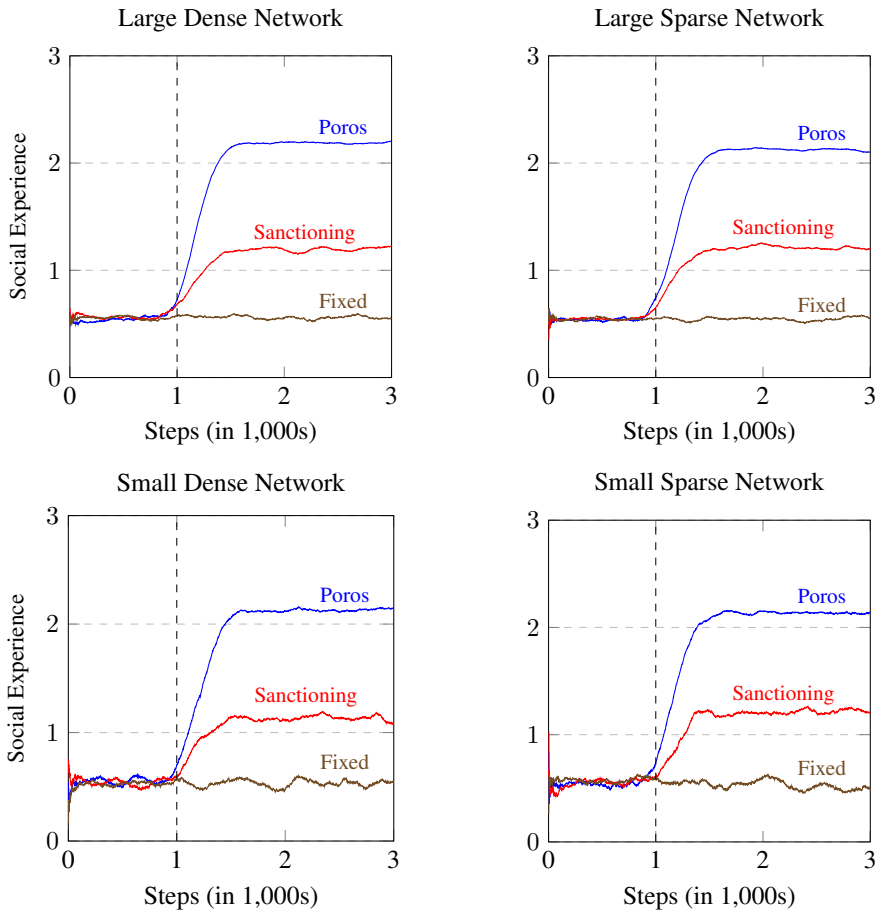

Figure 3.3 Social experience yielded by Poros, Sanctioning, and Fixed agents (per phone call for a window size of 200 steps) in pragmatic agent societies of different network sizes and densities.

answered. Poros agents offer higher social cohesion and experience than Sanctioning agents because the secondary stakeholders give smaller negative sanctions when they reason about context. The results for the other three network types are similar. Table 3.8 summarizes these results.

3.5.3 Experiment with Selfish Agent Society

In a selfish agent society, agents give a very large weight to their own payoffs when computing social experience. Agents here may not always negatively sanction others who disturb them. As in other societies, agents in a selfish society sanction a deviant based on their history.

Table 3.7 Effectiveness of Poros in a pragmatic society.

	Agent Type	Experience	Cohesion	p
Large Dense	Fixed	0.56	52.7%	< 0.01
	Sanctioning	1.21	63.5%	< 0.01
	Poros	2.19	83.2%	–
Large Sparse	Fixed	0.55	52.5%	< 0.01
	Sanctioning	1.21	63.5%	< 0.01
	Poros	2.19	83.2%	–
Small Dense	Fixed	0.53	52.1%	< 0.01
	Sanctioning	1.11	61.2%	< 0.01
	Poros	2.14	82.0%	–
Small Sparse	Fixed	0.54	52.5%	< 0.01
	Sanctioning	1.22	63.7%	< 0.01
	Poros	2.14	82.1%	–

Table 3.8 Effectiveness of Poros in considerate and selfish societies.

	Agent Type	Experience	Cohesion	p
Considerate	Sanctioning	–0.33	41.3%	< 0.01
	Poros	–0.14	48.4%	–
Selfish	Sanctioning	1.22	63.5%	< 0.01
	Poros	2.13	82.0%	–

Figure 3.4 shows the social experience plot for selfish Sanctioning and Poros agents in a Small-Dense network. The plots resemble those in the experiment with pragmatic agents, but with slightly lower stabilized values. Here, agents tend to answer all calls, which benefits both caller and callee most of the time. We observe similar results for the other three networks. Table 3.8 summarizes these results.

3.5.4 Threats to Validity

We identified and mitigated two threats. The first concerns a differences in how users perceive experience. In reality, not all users perceive social experience the same way, and thus aggregating with only one scheme introduces the threat of difference in perceiving social experience. To mitigate this threat, we conduct experiments with three agent societies with different experience

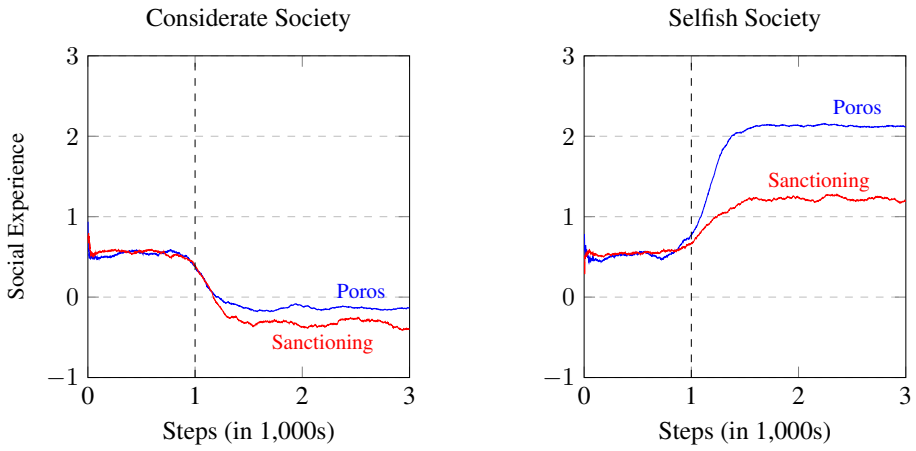

Figure 3.4 Social experience (averaged over a window size of 200 steps) yielded by Poros and Sanctioning agents in considerate and selfish agent societies simulated in a Small-Dense network.

aggregation schemes. The second threat concerns scalability. Since we simulate agent actions and interactions, a threat is whether our results scale to a large number of agents. To mitigate this threat, we evaluate Poros considering varying network sizes and types.

However, some threats remain. In particular, first, our results are based on simulation. Testing a SIPA's adaptability with end-users across contexts is challenging, as is reliably eliciting user attitudes and preferences.

Second, Poros agents always reveal context, which may pose a privacy threat. Ideally SIPAs should reveal context selectively. We leave this reasoning for future studies.

3.6 Conclusion and Future Directions

In Poros, SIPAs reveal and reason about context to understand the boundary of applicable norms and infer contextually relevant social norms. We find that Poros agents deliver significantly higher (1) social cohesion and (2) social experience than other agents. These findings are stable under changes to network size and characteristics of agents.

Being sensitive to norms, Poros SIPAs can naturally address challenges in engineering software tools for privacy. A SIPA would need data about its user's sharing preferences, privacy attitudes, and values and ethics [Ajmeri et al., 2018b] to make effective recommendations. A SIPA can learn its user's preferences and attitudes, but it would be helpful to bootstrap a SIPA via crowdsourced data about diverse user classes [Fogués et al., 2017a,b]. To better support

privacy-respecting SIPAs, Poros could incorporate characteristics suggested by Such [2017] and adopt argumentation as in Kökciyan and Yolum's [2017] work when deciding the subset of context to reveal.

Other future directions are incorporating affect in relation to norms [Ferreira et al., 2013] and supporting white lies to promote privacy (and social cohesion). For example, Bob may say his son is in hospital, instead of drug rehab. It would be instructive to study how such deception modulates effects on norms and goals.

Chapter 4

Reasoning about Values and Ethics

This chapter addresses our third research question, RQ_3 on values. It describes Ainur, our framework for designing ethical socially intelligent personal agents (SIPAs) that understand value preferences and reason about them, and its empirical evaluation.

Chapter 2 provides a method to assists software developers in SIPAs who promote privacy. We extend Chapter 2 by adding constructs to model stakeholders' values.

Chapter 3 enable SIPAs to infer contextually-relevant norms and apply those norms to provide privacy assistance to their stakeholders. The notion of privacy implicitly incorporates values by considering not only confidentiality but higher-level concerns such as disapprobation and avoiding infringing into others' space. Although Poros agents seek to maximize the social experience of their respective users, maximizing social experience may not translate to fairness. Ainur's focus is to balance the needs of a SIPA's primary and secondary stakeholders by understanding their preferences over values. Understanding of value preferences to aid group decision-making and its evaluation are novel.

4.1 Introduction

A *social machine* is characterized as a sociotechnical system comprising of social entities such as humans and technical components such as software jointly involved in the physical realization of a process [Chopra and Singh, 2016; Smart et al., 2014]. In the original conception of social machines, humans engage in the creative elements of this realization process, whereas technical components administer the process [Berners-Lee, 1999].

The conception of social machines has been gaining significant attention from academia

and industry in recent years. Prominent social web applications such as Wikipedia, Facebook, Twitter, Instagram, and Snapchat are examples of social machines. The very nature of social machines dictates that privacy be a key concern in their design. Whereas current methods for designing privacy into (social) web applications emphasize technology-centric notions such as authentication [Ruoti et al., 2015] and access control [Paci et al., 2018], they struggle to incorporate social mechanisms to control the threats to privacy [Hendler and Berners-Lee, 2010]. Such social mechanisms are of paramount importance in realizing privacy respecting social machines.

As envisioned, social machines facilitate natural interactions among autonomous parties (humans and organizations) as opposed to the predominant style of supporting social interactions via centralized servers (ranging from email servers to social network sites).

On the backdrop of privacy concerns and a rising number of privacy breach incidents, new regulations and standards such as the General Data Protection Regulation (GDPR) [GDPR, 2018] are being introduced. Spiekermann and Cranor [2009] suggest that to ensure effective implementation of privacy standards, it is necessary for engineers to design privacy-preserving systems that enable their users to control access to their private information. However, giving control to users raises two key concerns. First, does the information the users share on the web accord with their *values*? Second, does this sharing of information promote or demote any *values* for any other users concerned with the information? More often than not, these concerns are not addressed when users make sharing decisions today, because there is an excessive burden of decision making on them. A *SIPA* can help a user in decision making and sometimes automate the decision making.

We seek to design a social machine in which each user is supported by an (artificial) personal agent [Murukannaiah and Singh, 2014]. These agents interact with each other to facilitate creativity in social machines. We refer to such an agent as a socially intelligent personal agent (SIPA). Values are broad motivational goals or ideals worth pursuing for humans [Dechesne et al., 2013; Schwartz, 2012]. Ethicists subsume ethics in the theory of values [Friedman et al., 2008b], and regard privacy as a value with an ethical import [Langheinrich, 2001; Taylor, 2002]. Importantly, SIPAs in our setting understand values and act ethically. Whereas much of the existing literature on artificial intelligence focuses on rational decision-making by agents, we consider the problem of designing agents that act according to users' values and preferences among those values.

4.1.1 Values and Social Norms

Representing and reasoning about social norms (in context) is essential to producing an ethical SIPA. That is, an ethical SIPA acts in compliance with contextually relevant social norms (but it may choose to break some norms intentionally, e.g., when the norms conflict) (Chapter 2 explains). Even in the case of privacy, social norms are the centerpiece of privacy according to Nissenbaum's theory of *contextual integrity* [Nissenbaum, 2004, 2011], where privacy violations occur when information flows do not respect contextual norms.

In general terms, social norms describe interactions between a subject and an object in terms of what they ought to be, or as reactions to behaviors, including attempts to apply sanctions. We adopt Singh's [2013] representation of social norms, in which a norm is directed from a subject to an object and is constructed 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 norm generates a new instance each time it applies. This representation yields clarity on who is accountable to whom, when, and for what. We consider two main norm types in the present study: commitment and prohibition. A commitment norm means its subject is committed to its object to bring about a consequent if an antecedent holds, and a prohibition norm means its subject is forbidden by its object to bring about a consequent if an antecedent holds. For instance, *Frank* (subject), a high school student is *committed* (norm) to *Grace* (object), his mother, that he *will keep Grace updated about his location* (consequent) when he is *away from home* (antecedent).

Whereas norms require agents to perform or not perform certain actions, values provide a reason to pursue or not pursue those actions [Dechesne et al., 2013]. In general, each action by a SIPA promote or demote one or more values. For instance, in the phone ringer SIPA example described in Chapter 2, a callee's action of answering an urgent phone call during a meeting may promote the value of safety (for the caller), but demote the value of privacy (of the meeting attendees).

Only a few previous works have attempted to relate values with norms. Murukannaiah et al. [2016a] model actors, context, and social expectations via norms to engineer privacy respecting agents. Da Silva Figueiredo and Da Silva [2013] propose an algorithm to identify conflicts between norms based on values. A conflict occurs when (1) a consequent action of a commitment norm demotes a value, or (2) a consequent action of a prohibition norm promotes a value important to a SIPA user. Dechesne et al. [2013] develop a model of norms and culture, represented by values, to study compliance of norms. They concur that values are important in deciding whether or not a norm should be introduced. Kayal et al. [2014] present a model in

which norms and context are centered on values. Such a model could be employed to govern a SIPA by identifying value preferences of the SIPA's users.

4.1.2 Contribution

A SIPA's actions may promote or demote certain values of its users. Performing actions that promote values preferred by users is essential to providing a satisfactory experience. If a SIPA understands its users' value preferences and reasons about the values promoted or demoted by each of its actions, it could select ethically appropriate actions, such as setting a phone to ring loud for an urgent phone call during a meeting, that provide a satisfactory social experience to its stakeholders. Accordingly, we consider the following research question:

RQ Does an ability to reason about values promoted or demoted by actions and an understanding of preferences among these values help a SIPA deliver a value-driven social experience to all its users?

To investigate the research question above, we develop Ainur, a framework to design ethical personal agents that can reason about values. Importantly, Ainur considers multiparty privacy (1) in reference to users having distinct value preferences, and (2) based on ethical decision-making in light of other user's preferences [Fogués et al., 2017b].

Unlike earlier works [Fogués et al., 2017b] that consider the majority opinion for decision making or select actions considering negative or positive consequences, Ainur adapts a multicriteria decision-making approach [Opricovic and Tzeng, 2004] to identify a consensus action. The actions identified by Ainur adhere to Rawl's moral theory of justice that suggests Maximin as a basis of fairness [Rawls, 1985].

We evaluate Ainur via multiple simulation experiments with agent societies varying in privacy attitudes. Our simulation experiments are grounded in data from an immersive survey wherein participants select a location check-in policy for a given context.

We find that Ainur SIPAs that understand the value-preferences of their stakeholders act ethically. That is, a SIPA selects fair actions—actions that maximize the minimum (i.e., worst-case) experience for each stakeholder involved in interactions with the SIPA, and yields a better overall social experience, i.e., higher mean experience for all stakeholders.

4.1.3 Organization

The chapter is structured as follows. Section 4.2 provides a motivating example from the domain of mobile social applications. Section 4.3 describes our approach, including a conceptual model

to design ethical SIPAs that understand value preferences and reason about them. Section 4.4 details our simulation setup and the human-subject study we conduct to collect data about real users' attitudes and value preferences. We use this data to seed our simulation. Section 4.5 describes the simulation experiments we conduct to evaluate Ainur, and their results. Section 4.6 concludes with a discussion of relevant related works and future directions.

4.2 Motivating Example

For concreteness, we consider the domain of mobile social applications where privacy is an important value [Spiekermann and Cranor, 2009; Taylor, 2002], and present an example SIPA to demonstrate our ideas. Consider Pichu, a location sharing application, as a SIPA that enables its user to stay connected with his or her friends and family. A Pichu user can share his or her location publicly, with common friends, with companions, with specific people, or with no one. Here, the common friends situation arises when the user is accompanied by someone, and revealing the user's own location would indirectly reveal the accompanying person's location. Pichu suggests a sharing policy to its user. To produce a policy, it relies upon multiple contextual attributes, such as the place where the user is, the user's companions, the activity the user and companions are engaged in, and so on. Additionally, if the user has companions, Pichu must understand their preferences and act ethically.

Example 5 (Olympiad) *Frank, a Pichu user, is a high school student in New York who values pleasure and social recognition. Also, he is committed (a norm) to his mother Grace that he will share his location with her when he is not at home. Sharing location promotes security but demotes privacy. Frank travels to University of Illinois at Urbana-Champaign to participate in the National Science Olympiad. Pichu shares publicly that Frank is at University of Illinois at Urbana-Champaign participating in the Science Olympiad, and thus satisfies Frank's commitment to his mother, and promotes pleasure and social recognition for him.*

Example 6 (Pizza at Giordano's) *When returning from Urbana-Champaign, Frank visits his uncle Harold in Chicago. Harold is an Intelligence analyst with the National Security Agency and values privacy. He and Frank visit Giordano's, a famous pizzeria for lunch. Pichu prefers Harold's privacy over Frank's pleasure and social recognition, and shares only with Grace that Frank is at Giordano's with Harold. Doing so, also satisfies Frank's commitment to his mother without harming Harold.*

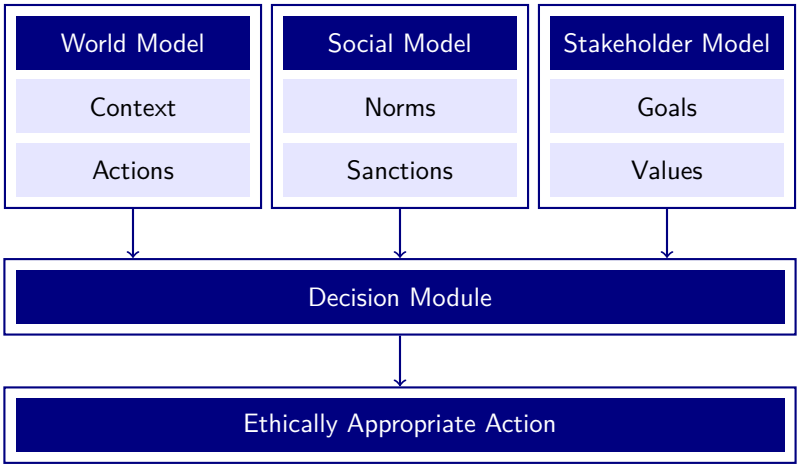

Figure 4.1 A conceptual model of a Ainur SIPA.

The Pichu examples illustrate some of the opportunities for SIPAs to reason about values and act ethically. Note that Pichu is an example SIPA, and not the only application of Ainur. We use Pichu as a running example to explain Ainur.

4.3 Ainur

We propose Ainur to design ethical SIPAs that understand and reason about preferences among values to make policy decisions, as explained in Section 4.2.

A SIPA should be aware of its users, their goals, and relevant actions to bring about the goals, which may vary with the social context. A SIPA should choose and execute actions, especially when there are conflicts among goals and social expectations, based on its users' contextual preferences of the applicable social norms [Ajmeri et al., 2017b]. Users' preferences among values provide a strong basis for choosing which goal to bring about or which norm to satisfy. In Ainur, a SIPA selects ethically appropriate actions by learning its users' preferences across the various values.

4.3.1 Conceptual Model

Figure 4.1 shows a conceptual model of a Ainur SIPA. Each SIPA maintains an instance of this model internally. The conceptual model includes a world model, a social model, a stakeholder model, and a decision module to assist in ethical decision-making.

4.3.1.1 Stakeholder Model

A SIPA's *stakeholder model* describes a SIPA's stakeholders, and their goals and values.

- *Stakeholders* are users who either interact with a SIPA directly—*primary stakeholders*, or are affected by a SIPA's actions—*secondary stakeholders* [Friedman et al., 2008b]. In Examples 5 and 6, Frank is the primary stakeholder, and Grace and Harold are the secondary stakeholders of Frank's Pichu SIPA.
- A *goal* defines the preferable states of the world for a SIPA's stakeholder. For example, Frank's goal is to *be connected with his family and friends*.
- A *value* for a stakeholder is an ideal worth pursuing in a given context. For instance, in Examples 5 and 6, Frank has a preference for values of pleasure, recognition, and security over other values, and Harold values privacy over other values.

4.3.1.2 World model

A SIPA's *world model* describes the context in which a SIPA acts.

- A *context* is the circumstance in which a SIPA takes an action [Murukannaiah and Singh, 2014]. For instance, Frank's context in Example 5 is *participating in the National Science Olympiad at the University of Urbana Champaign*.
- An *action* represents the steps a SIPA takes to bring about its stakeholders' goals. For instance, in Example 5, Frank's Pichu SIPA's action of publicly sharing Frank's context *participating in the National Science Olympiad at the University of Illinois at Urbana Champaign*, helps Frank achieve his goal of *being connected with his mother*.

4.3.1.3 Social model

A SIPA's *social model* specifies the norms governing a SIPA's interactions in a society and the associated sanctions.

- *Norms* characterize the social architecture that promotes prosocial behavior. In the Pichu examples, Frank's commitment for sharing his location with Grace, his mother, is one such norm that Frank's Pichu SIPA should adhere to.
- A *sanction* is an action that one stakeholder may take against another stakeholder when he or she satisfies (positive sanction) or violates (negative sanction) a norm [Nardin

et al., 2016]. For instance, in the Pichu example, some resulting sanctions could be Grace appreciating Frank on keeping her informed, and Harold scolding Frank if he publicly shared his location tagging Harold.

4.3.1.4 Decision Module, Ethically Appropriate Action, and Social Experience

A SIPA's *decision module* is responsible for producing an *ethically appropriate action* that yields a (fair) social experience to the SIPA's stakeholders, especially in scenarios where either the norms conflict or the value preferences of stakeholders are not aligned.

A SIPA's stakeholders perceive a utility for each alternative action available to a SIPA in a given context. The decision module adapts VIKOR [Opricovic and Tzeng, 2004], a multicriteria decision-making approach, to aggregate these utilities and produce a consensus action. VIKOR is based on closeness to the ideal solution. It prioritizes social utility over individual utility. Our experiments, which we describe in Section 4.5, show that solutions obtained by Ainur adapting VIKOR exhibit the Rawlsian property of justice in terms of maximizing the minimum experience across a SIPA's stakeholders [Leben, 2017; Rawls, 1985].

Social experience is the aggregated utility perceived by a SIPA's stakeholders. It characterizes the extent to which an action taken by a SIPA is perceived as ethical by its stakeholders in a given context. For instance, understanding Frank's and Harold's value preferences, and Frank's commitment to Grace, Frank's Pichu SIPA deciding to share only with Grace that Frank is with Harold at Giordano's is ethically more appropriate than its sharing that information with no one or sharing that information with public.

4.3.2 Ainur SIPA Society

We now describe a SIPA society consisting of Pichu SIPAs as introduced in Section 4.2.

4.3.2.1 Society

A SIPA society in Ainur is defined as a tuple $\mathbb{S} = (S, P, R, RT, f)$, where $S = \{s_1, \ldots\}$ is the set of SIPAs in the society; $P = \{p_1 \ldots\}$ is a set of their primary stakeholders; R is the set of relationships between the stakeholders such that $R \subseteq S \times P \times RT$; RT is the set of type of relationships; and $f : S \to P$ is the $\mathbb{S}$'s function that maps a SIPA s_i to its primary stakeholder p_i.

Each stakeholder is a tuple $p_i = (G_i, V_i)$ where G_i are set of p_i's goals and V_i are set of values preferred by p_i.

Each SIPA is a tuple $s_i = (p_i, A_i, f_i')$ where p_i is s_i's primary stakeholder, A_i is a set of s_i's actions, and $f_i' : G \rightarrow A$ is s_i's planning function that maps p_i's goals to s_i's actions.

$C = \{c_i, \ldots\}$ is the set of social contexts in which the SIPAs interact; and N is the set of norms such that $c_i \subseteq N$ governing the SIPAs' interactions in these contexts.

In a society of Pichu SIPAs, when a stakeholder moves to a new place or meets new people, his or her SIPA may share the context in which the stakeholder is, to bring about the stakeholder's goal g of *staying connected*. The stakeholder's (and the SIPA's) context includes the place where a stakeholder is and the people who he or she is with. A SIPA selects one of the following three actions, $A = \{$share with all, share with common friends, share only with companions$\}$ in each context.

4.3.2.2 Relationship

Each SIPA stakeholder in a SIPA society is connected to another stakeholder via a relationship edge $r \in R$. Each relationship edge r_{ij} in R is a tuple $r_{ij} = (s_i, s_j, rt \mid s_i, s_j \in A, rt \in RT)$, where RT is set of relationship types, RT $= \{rt_1, rt_2, \ldots, rt_n\}$.

In the Pichu society, RT includes co-worker, family, friend, and so on.

4.3.2.3 Context

At any given instant t, a SIPA s with its stakeholders is in a context c, which is defined by a tuple $(l, \hat{S} \mid l \in L, \hat{S} \subseteq S \setminus s)$, where l is a place from $L = \{l_1, l_2, \ldots, l_n\}$. A place is a location such as home, office, meeting, or restaurant as understood in conceptual terms. L in Pichu includes conference, hiking, restaurant, and so on. Each place l is defined by attributes such as physical conditions (e.g., rainy), expected activities (e.g., hiking), social interactions (e.g., having a discussion), and temporal information (e.g., at late night). $\hat{S}$ is a set of other SIPAs at l such that in the current context c, the primary stakeholders of SIPAs in $\hat{S}$ are the secondary stakeholders of s—who could be affected by s's actions in context c.

When a SIPA's stakeholder moves between places, or when new people (also stakeholders) join a SIPA's stakeholder, the context changes. For instance, the context changes when Harold joins Frank at Giordano's from when Frank is alone at Giordano's.

Each context c includes a set of contextually relevant norms $Nc \subseteq N$ that govern the interaction of SIPAs in that context. For example, Frank's commitment to Grace may be relevant only when he is traveling.

4.3.2.4 Values and Contextual Preference

V_c is a set of values that are influenced by a SIPA's actions in the current context c. For example, when Frank is at Giordano's, V_c = {pleasure, privacy, recognition, security}.

Each agent s has a preference q over values that depends on context c, represented by a set of tuples $\{(v_j, v_k, c) \mid v_j, v_k \in V, c \in C\}$ such that s prefers v_j over v_k in c. Frank's preference for values of pleasure and recognition over privacy during the Olympiad can be represented as {(pleasure, privacy, olympiad), (recognition, privacy, olympiad)}

In a decision-making episode, a SIPA determines (1) the context it is in through the sensors the SIPA is equipped with, (2) the future state of the world for each action it can perform, (3) the value preferences of its stakeholders, and (4) the social experience its stakeholders will derive for each action it can perform. Then, based on the applicable norms in a given context and its stakeholders goals, a SIPA identifies an action to perform.

4.3.3 Value Preferences

A SIPA's stakeholders may have inconsistent preferences in some context. Thus, a SIPA's actions based solely on one (e.g., primary) stakeholder's preference may conflict with its other stakeholders' preferences. For instance, in Example 6, if Frank's SIPA shares publicly that Frank and Harold are having a pizza considering Frank's preference for *pleasure*, the selected action conflicts with Harold's preference for *privacy*.

Sotala [2016] proposes using a reward function for a human's values, which a value-respecting AI system can learn and maximize. If a SIPA maintains numeric representations of its stakeholder's preferences over different values, it can aggregate the gain of values promoted when choosing an action.

4.3.3.1 The VIKOR Method

In Ainur, we use the VIKOR method [Opricovic and Tzeng, 2004], a multicriteria decision-making (MCDM) method to identify which actions to perform in situations where (1) actions prescribed by the norms conflict with actions that promote the values preferred by a SIPA's stakeholders, or (2) the stakeholders of a SIPA have different value preferences and thus prefer different actions. VIKOR's ranking method is based on closeness to the ideal solution, and provides an ethically appropriate solution that yields high social utility as against high individual utility.

We now summarize the VIKOR method [Opricovic and Tzeng, 2004] below. VIKOR relies on numeric payoffs. We can map preferences to numeric payoffs by adopting techniques such

as *cumulative voting*—distributing a fixed number of points to each value preference over each available alternative action, or *cardinal voting*—giving numeric payoff (ratings) on a fixed scale to each value for all available alternative actions [Pacuit, 2017].

1. Determine the best and worst numeric payoffs, f_x^* and f_x^- for each value preference x over the alternative actions y to bring about a goal. That is, $f_x^* = \max_y f_{xy}$, $f_x^- = \min_y f_{xy}$.

2. For each alternative action y, compute the weighted and normalized Manhattan distance [Krause, 1973]:

   $S_y = \sum_{x=1}^{n} w_x(f_x^* - f_{xy})/(f_x^* - f_x^-)$, where w_x is the weight for value preference x, which is subject to a stakeholder context and preferences over values. In particular, $S_y = 0$ when $f_x^* = f_x^-$.

3. Compute the weighted and normalized Chebyshev distance [Cantrell, 2000]:

   $R_y = \max_x[w_x(f_x^* - f_{xy})/(f_x^* - f_x^-)]$, where w_x is the weight for value preference x.

4. Compute $Q_y = k(S_y - S^*)/(S^- - S^*) + (1-k)(R_y - R^*)/(R^- - R^*)$, where

   - $S^* = \min_y S_y$,
   - $S^- = \max_y S_y$,
   - $R^* = \min_y R_y$,
   - $R^- = \max_y R_y$, and
   - k is a weight of the strategy to maximize either group or individual experience.

   We set $k = 0.5$ to select a consensus policy.

5. Rank alternative actions, sorting by the values S, R, and Q, in increasing order. The results are three ranked lists of actions.

6. Choose the alternative based on $\min Q$ as the compromise solution if it is better than the second-best alternative by a certain threshold or also the best ranked as per S and R.

Table 4.1 demonstrates possible numeric values of the value preferences and the calculated ranking of three alternative actions (share with all, share with common friends, and share only with Grace) that Pichu can take when Frank is with Harold at Giordano's, as in Example 6. Since Harold is highly cautious about his privacy, we give a higher weight to Harold's privacy (3) and a lower but equal weight to other seven criteria including Harold's other values and Frank's values. We assume $k = 0.5$ in this case, and find the alternative y_3, *share only with Grace* as the best solution.

Table 4.1 Computing rankings for policy alternatives using VIKOR for context *Pizza at Giordano's* in Example 6. Bold indicates the best alternative.

Policy Alternatives	Frank's Values				Harold's Values				S_y	R_y	Q_y
	Pleasure	Privacy	Recognition	Safety	Pleasure	Privacy	Recognition	Safety			
y_1 All	10	5	10	5	5	0	5	5	3.5	3	0.75
y_2 Common	5	5	5	10	5	0	5	5	0.4	3	1
y_3 Grace	0	5	0	0	5	15	5	5	**0.3**	**1**	**0**
w_x	1	1	1	1	1	3	1	1			
f_x^*	1	0	1	1	0	1	0	0			
f_x^-	0	0	0	0	0	0	0	0			

4.4 Simulations

We adopt MASON [Luke et al., 2005], a multiagent simulation toolkit, to develop a simulation environment containing a society of Pichu SIPAs.

4.4.1 Simulation Society Setup

The society contains a social network of Pichu SIPAs.

SIPAs. The society contains several SIPAs. Each SIPA has a primary stakeholder on whose behalf the SIPA acts.

Relationships. SIPA stakeholders are either socially related to other stakeholders—co-worker, family or friend, or are strangers.

Goal. Stakeholders have a goal to stay connected with their connections—co-workers, family and friends.

Actions. To help bring about their stakeholders' goal of staying connected, as stakeholders move, their SIPAs either *share their context publicly*, *share only among common friends of all companions*, or *share only with companions*. Context-sharing actions are selected based on norms and stakeholders' value preferences.

Norm. The society is governed by a privacy norm—*preserve privacy of all stakeholders*, i.e., stakeholders are committed to act in a privacy-preserving manner when sharing context while accompanying others. This commitment is directed from primary stakeholders to secondary stakeholders.

Value preference. For simplicity, we consider only four values—pleasure, privacy, recognition, and security, relevant to our problem domain.

Places and contexts. Table 4.2 lists the places we represent. Each place has two attributes—*how safe it is* and *how sensitive it is*. In the simulation, the combination of places where SIPA stakeholders are, and who accompanies them, define their context.

Stakeholders along with their SIPAs move between places. At each place in the context, a stakeholder is either alone, with companions—co-workers, family, or friends, or with crowd (several people are around but they are strangers).

Table 4.2 List of places in the simulation environment, each marked safe or sensitive.

Place	Safe	Sensitive
Attending graduation ceremony	–	No
Presenting a conference paper	–	No
Studying in library	Yes	–
Visiting airport	Yes	–
Hiking at night	No	–
Being stuck in a hurricane	No	–
Visiting a bar with fake ID	–	Yes
Visiting a drug rehab center	–	Yes

Stakeholder types. Stakeholders are of three types based on their privacy attitudes [Westin, 2003]. We do not use Westin's questionnaire but bucket stakeholders based on Schnorf et al.'s [2014] privacy attitude survey.

A **privacy cautious** stakeholder is most protective about his or her privacy.

A **privacy casual** stakeholder is inclined to share information about himself or herself. He or she perceives more benefit from sharing information than holding it.

A **privacy conscientious** stakeholder exhibits a pragmatic behavior, and weighs pro and cons of sharing information based on a given context.

4.4.2 Human-Subject Study to Seed Simulation

Naeini et al. [2017] conducted a human-subject study on privacy expectations in which 1,007 participants stated their preferences in the contexts of 380 IoT data collection and use scenarios. They suggest that users' preferences can be accurately predicted after observing their decisions in a few scenarios. We take insights from their findings in conducting our human-subject study.

To seed the simulation environment with value preferences of real users, we conducted a survey of students enrolled in a mixed graduate and undergraduate-level computer science course. The study was approved by our university's Institutional Review Board (IRB). We obtained informed consent from each of 58 participants.

First, the participants completed a privacy attitude survey [Schnorf et al., 2014] in which they answered questions on their level of comfort in sharing personal information on the Internet on a Likert scale of 1 (very comfortable) to 5 (very uncomfortable), and the

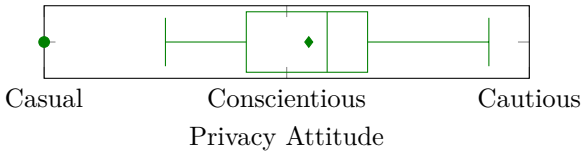

Figure 4.2 Distribution of privacy attitudes of the human-subject study participants.

extent sharing personal information causes (or could cause) them negative experience, again on a Likert scale of 1 (not at all) to 5 (to a very great extent). Based on a participant's response, we bucket him or her into one of the three privacy attitude buckets—*casual* to represent privacy unconcerned people, *conscientious* to represent privacy careful people who take decisions on a case-to-case basis, or *cautious* to represent privacy concerned people. Figure 4.2 shows the distribution of privacy attitudes of the study participants.

Next, the participants completed two context-sharing surveys. In the first survey, they were given a list of contexts, as listed in Table 4.2, and their companions (alone, co-worker, family, friend, or crowd) in the given context, and were asked to select a sharing policy (share with all, share only with common friends, or share only with companions). In the second survey, participants were additionally informed of the values that are promoted or demoted on sharing and on not sharing the context, and were asked to select a context-sharing policy accordingly. We use the first survey to engage and immerse the participants in various contextual scenarios, and the second to help them make informed decisions according to the values promoted or demoted in each context.

We use the privacy attitudes of the participants and the context-sharing policies selected by the participants to create multiple artificial societies of stakeholders and to seed the simulation experiments described in Section 4.5.

4.5 Experiments and Results

We evaluate our research question via two experiments in which we simulate societies of Pichu SIPAs who visit different places and may share their context. First, we experiment with a society of stakeholders with mixed privacy attitudes representing the attitudes of participants, collected in the study described in Section 4.4.2. Second, we experiment with three societies with a majority of privacy casual, conscientious, or cautious stakeholders, respectively.

Our results are stable with respect to changes in the network size and the connectedness

of a SIPA society.

4.5.1 Decision-Making Strategies

As Pichu SIPAs move between places and interact with each other, they make policy decisions that affect their stakeholders. To evaluate SIPAs designed via Ainur, we define four (Ainur and three baseline) policy decision-making strategies.

S_{Ainur}: Ainur. The SIPA, based on its stakeholders' value preferences, computes a context-sharing policy using the VIKOR method.

$S_{primary}$: Primary user's preference. The SIPA produces a context-sharing policy based only on its primary stakeholder's value preferences. This strategy is representative of how location sharing works today in social networking websites such as Facebook.

$S_{conservative}$: Most privacy conservative policy. The SIPA produces the least privacy violating, i.e., the most restrictive context-sharing policy among the available alternatives based on its stakeholders' value preferences. This strategy represents policy selection based on the least negative consequence.

$S_{majority}$: Majority policy. The SIPA produces the most common context-sharing policy based on its stakeholders' value preferences. This strategy represents policy selection based on majority voting.

4.5.2 Metrics

For each SIPA interaction, we compute these measures:

Mean social experience, the mean utility obtained by the society as a whole based on context-sharing policy decisions. Higher is better.

Best individual experience, the maximum utility obtained by any of the SIPA's stakeholders during a single interaction. Higher is better.

Worst individual experience, the minimum utility obtained by any of the SIPA's stakeholders during a single interaction. The intuition behind choosing this measure is to verify if a society supports Maximin [Leben, 2017]. Higher is better.

Fairness, the reciprocal of the difference between the best and the worst individual experience obtained by the SIPA's stakeholders during a single interaction. This measure is based on the dispersion of the experience yielded by SIPAs [Rawls, 1985]. Higher is better.

Computing Utility.

The utility that a SIPA obtains from a sharing policy in a certain context, whether to a primary or a secondary stakeholder, is a weighted sum of four numeric utility payoffs that the stakeholder perceives with respect to the four types of values considered in our example. We preset these numbers in a utility matrix such that they reflect a human-subject's preferences over the corresponding values. We assume that a stakeholder receives the maximum utility when the chosen sharing policy is the preferred one, and the utility decreases linearly when the policy chosen by a SIPA deviates from it. Table 4.3 lists the preferred policies and utility numbers for each value of one human-subject in different contexts.

Table 4.3 Example numeric utility matrix for a stakeholder.

Place	Companion	Policy	Value			
			Pleasure	Privacy	Recognition	Security
Graduation	Family	All	1	0	1	0
Conference	Co-workers	None	0	1	0	0
Library	Friends	All	1	0	0	0
Airport	Friends	Common	0	1	0	0
Hiking	Alone	All	1	0	0	1
Hurricane	Family	All	1	0	0	1
Bar	Alone	None	0	2	0	0
Rehab	Friends	None	0	2	0	0

4.5.3 Hypotheses

We propose the following hypotheses to evaluate our research question. We omit the corresponding null hypotheses for brevity.

H_{social}. Ainur yields better mean social experience than baseline strategies.

H_{best}. Ainur yields higher best individual experience than baseline strategies.

H_{worst}. Ainur yields higher worst individual experience than baseline strategies.

$H_{fairness}$. Ainur yields higher fairness than baseline strategies.

4.5.4 Experimental Setup

We run simulations on a society of Pichu SIPAs. All parameters described below are set empirically based on the human-subject study we conducted.

Specifically, we experiment on a society of 580 SIPAs, ten per study participant, each of which assumes the properties, including preferred choices and privacy attitude, of an actual study participant. In the default setting, which we use in the experiment with a mixed agent society, the SIPAs are mapped evenly to the participants. For each pair of SIPAs, their relationship is co-worker, friend, family (with equal probability), or strangers. Relationships are assigned at the beginning of the simulations such that they exhibit small world properties (degree: 10, rewiring prob: 0.05, edges: 3,445, clustering coefficient: 0.56, density: 0.014, average distance: 4.71) [Watts and Strogatz, 1998].

At each step in the simulation, each SIPA is at one of the eight places listed in Table 4.2. The SIPA moves after one step to another place with equal probability. A SIPA decides a context-sharing policy based on the current place and the SIPA's stakeholders' privacy attitudes, value preferences, and decision-making strategy in Section 4.5.1.

For each setting, we run the simulation 2,000 steps three times and report the mean social experience, the best individual experience, the worst individual experience, and fairness. We plot the mean social experience after every 100 steps in one run.

4.5.5 Experiment with Mixed Agent Society

First, we experiment using the default settings as described above. The privacy attitude distribution of the mixed agent society mimics the privacy attitude distribution of our study participants shown in Figure 4.2.

To evaluate hypothesis H_{social}, we compare the *mean social experience* obtained by SIPAs built according to the four decision-making strategies—S_{Ainur}, $S_{primary}$, $S_{conservative}$, and $S_{majority}$. Similarly, for H_{best}, H_{worst}, and $H_{fairness}$, we compare the *best individual experience*, *worst individual experience*, and *fairness*, respectively, as yielded by these decision-making strategies.

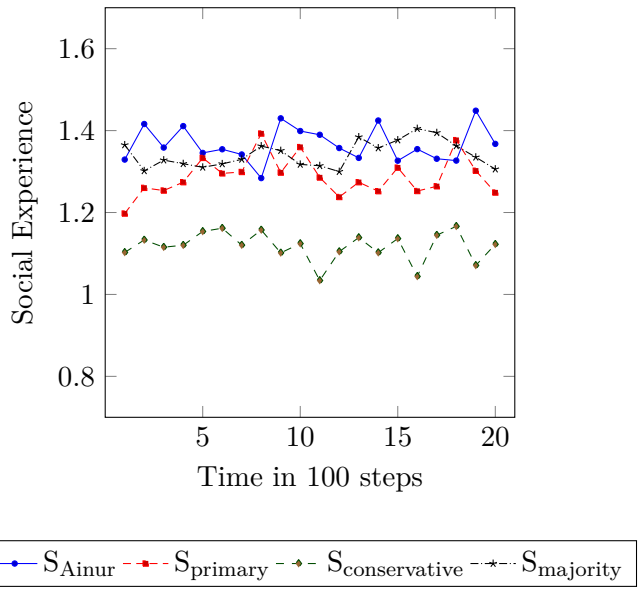

Figure 4.3 Ainur vs. other strategies: Social experience in a mixed society.

Table 4.4 summarizes the results for the experiments with a mixed agent society. It shows values for mean, best, and worst social experience, fairness, and p-values from the two-tailed paired t-tests comparing the mean social experience yielded by Ainur and by other strategies. Figure 4.3 shows the mean social experience plots.

Table 4.4 Comparing social experience, best and worst individual experience, and fairness yielded by Ainur SIPAs using VIKOR vs. other decision-making strategies in a society with mixed privacy attitudes.

Strategy	Mean	Best	Worst	Fairness	p
S_{Ainur}	**1.361**	1.715	**0.767**	**1.05**	–
$S_{primary}$	1.286	1.789	0.579	0.83	<0.01
$S_{conservative}$	1.106	1.721	0.472	0.80	<0.01
$S_{majority}$	1.339	**1.836**	0.570	0.78	<0.01

We observe that Ainur yields better mean social experience than other decision-making strategies. Although the mean best individual experience obtained by Ainur SIPA stakeholders is not the largest, they yield the highest mean worst individual experience and

fairness. These results indicate that Ainur yields solutions such that each companion is treated fairly, and thus Ainur SIPAs act ethically. Thus, the null hypotheses corresponding to H_{social}, H_{best}, $H_{fairness}$ are rejected.

4.5.6 Experiments with Majority Privacy Attitudes

Next, since our study sample may not be representative of privacy attitudes of the general population, we create three artificial societies with stakeholders having different distributions of privacy attitudes from the study data. We experiment with societies that are dominated by privacy casual, conscientious, and cautious stakeholders. Boxplots in Figure 4.4 show the distributions of privacy attitudes of the stakeholders in these artificial societies.

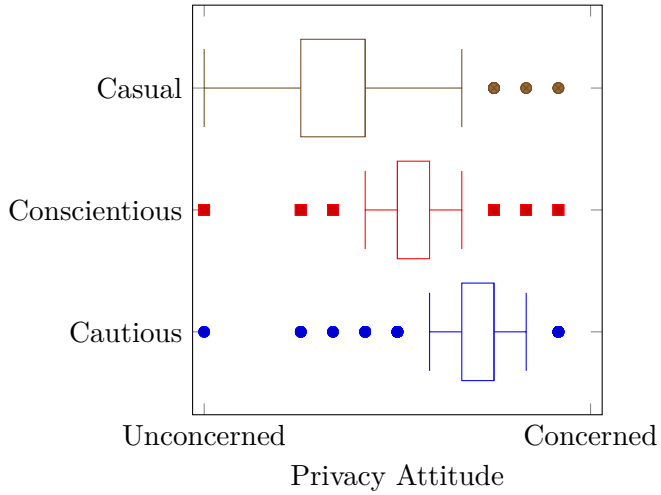

Figure 4.4 Privacy attitude distributions for artificial societies of cautious, conscientious, and casual stakeholders.

Table 4.5 summarizes the results for the experiments with privacy cautious, conscientious, and casual societies. Figure 4.5 shows the social experience plots for these experiments.

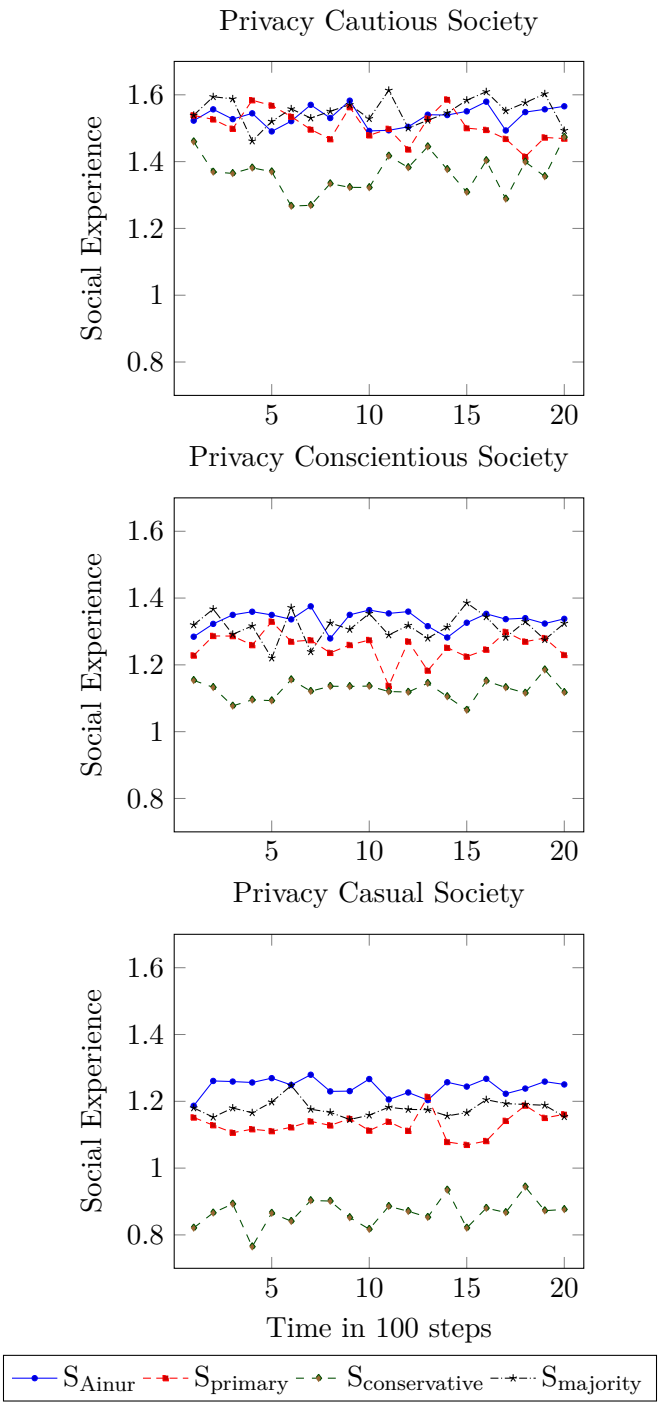

Figure 4.5 Comparing Ainur with other strategies with respect to social experience in societies based on privacy attitudes.

Table 4.5 Comparing social experience, best and worst individual experience, and fairness yielded by Ainur SIPAs using VIKOR with other decision-making strategies in societies based on majority privacy attitudes.

Strategy	Cautious				Conscientious				Casual			
	Mean	Best	Worst	Fairness	Mean	Best	Worst	Fairness	Mean	Best	Worst	Fairness
S_{Ainur}	1.535	1.664	**1.233**	**2.27**	**1.329**	1.531	**0.867**	**1.51**	**1.242**	1.457	**0.768**	**1.45**
$S_{primary}$	1.506	1.766	1.082	1.46	1.253	1.592	0.679	1.10	1.129	1.466	0.584	1.13
$S_{conservative}$	1.366	1.745	1.059	1.46	1.093	1.519	0.608	1.10	0.870	1.338	0.454	1.34
$S_{majority}$	**1.551**	**1.858**	1.007	1.18	1.318	**1.699**	0.575	0.89	1.176	**1.534**	0.518	0.98

4.5.6.1 Privacy Cautious Society

In the experiment with a privacy cautious society, Ainur yields the second-best mean social experience, next only to the majority strategy. Ainur yields the highest worst individual experience, i.e., the minimum utility that SIPA stakeholders obtain is higher compared to other decision making strategies, and hence supports Maximin criteria. For fairness, Ainur has the highest outcome. Thus, the null hypotheses related to H_{worst} and $H_{fairness}$ are rejected.

4.5.6.2 Privacy Conscientious Society

Here, Ainur yields the best mean social experience and maximizes the worst individual experience while giving the fairest solutions. Hence, the null hypotheses related to H_{worst} and $H_{fairness}$ are rejected.

4.5.6.3 Privacy Casual Society

Again, Ainur yields the best mean social experience while giving the fairest solutions with the highest worst individual experience; thus, the null hypotheses related to H_{worst} and $H_{fairness}$ are rejected.

4.5.7 Threats to Validity and Mitigation

We identify and mitigate three threats. The first threat concerns simulation as an evaluation methodology. Obtaining data with actual human users' preferences and attitudes is challenging. And more so, testing a SIPA's adaptation in all possible social contexts would be infeasible. Simulation provides an excellent avenue to overcome that challenge. Our simulation results are grounded in data obtained from users.

Second, in reality users may perceive social experience differently than when completing a survey. Self-reported attitudes are unreliable and indirect methods may yield better quality data. To mitigate the threat associated with self-reported attitudes, we employ context check-in scenarios in which participants immerse themselves and accordingly provide check-in policies.

Third, the privacy attitudes in our survey sample may not be representative of the broader population. Even with a survey on a larger scale, imagining all possible contexts is challenging. To mitigate this threat, we conduct multiple experiments with societies having different privacy attitudes.

4.6 Discussion

We advance the science of privacy by tackling a nuanced notion of privacy—understood as an ethical human value. We envision a society of socially intelligent agents participating in creative processes while they serve as technical components of a social machine. We borrow principles from operational research to guide these agents in ethical decision-making.

Specifically, we address the problem of designing an ethical personal agent that understands the value (including privacy and security) preferences of all its stakeholders (and not just its primary human user), and reasons about those preferences when acting on behalf of or suggesting actions to its user.

4.6.1 Other Related Works

We now discuss some other relevant related works.

4.6.1.1 Norms, Values, and Privacy in Multiagent Systems

The contributions of this paper relate to the extensive work on norms and values, as well as privacy in the multiagent systems community.

Cranefield et al. [2016] provide a machine learning approach through which agents can identify its societal norms based on their observations of behavior and sanction in a society. In contrast, the present work is focused on values. It is empirically grounded in data from users, and it takes into account group decision-making.

Borning and Muller [2012] address Value Sensitive Design and suggest that researchers respect differences in user's views on human values widely across cultures and contexts.

Anderson and Anderson [2015], however, assert that autonomous systems should be guided by ethical consensus determined by ethicists in different areas. They propose a case-supported paradigm to help ensure autonomous systems that make decisions only when there is a consensus on what is ethically correct. Ainur SIPAs select ethically appropriate actions by aggregating value preferences of all its users.

Bench-Capon and Modgil [2017] argue for the need of value-based reasoning in agents, especially when norms should be violated. They propose that agents keep track of the preference ordering of values. However, they posit that rules are made to be broken, and consider only the circumstances where norms should be violated. We argue that agents should model and resolve conflicts among norms based on stakeholders' views on values as

well as social contexts. Under different social contexts, stakeholders' preferences of values may also vary.

Cranefield et al. [2017] describe a mechanism of value-based reasoning for BDI (Belief-Desire-Intention) agents. They argue that decision-making by agents in normative systems, such as the selection of norms, is indirectly influenced by the value system, and therefore do not model norms in their approach. However, without norms, agents would need a complete understanding of human values to make morally correct decisions, which is difficult to realize.

Dignum [2017] argues for the need for AI reasoning to take into account societal values because autonomous AI systems increasingly affect our lives. Dignum proposes several approaches to responsibility in AI design considering human values.

Serramia et al. [2018] show how to incorporate values along with norms in a heuristic decision-making framework. Kayal et al. [2018] propose an automatic value-based model for resolving conflicts between norms, especially social commitments, in multiagent systems. Their results from a user study provide evidence that values can influence, and therefore could be used to predict, users' preferences when resolving conflicts. Ainur supplements Kayal et al.'s model by providing constructs and mechanisms to develop value-driven ethical SIPAs and thus, goes beyond conflict resolution.

4.6.1.2 Value Alignment in Artificial Intelligence

Some recent works focus on value alignment of artificial intelligence systems and how agents can learn human values correctly. Riedl and Harrison [2016] argue that it is not easy for developers to exhaustively enumerate human values, and propose that agents use sociocultural knowledge embedded in stories, such as crowdsourced narratives, to learn human values. Arnold et al. [2017] address how inverse reinforcement learning (IRL) can be used in value alignment, and propose a hybrid approach for reasoning about moral norms combining IRL and logical representations of norms. Ainur SIPAs can adopt such approaches to learn value preferences of their users.

4.6.1.3 Understanding Privacy Preservation

Many researchers focus on understanding the growing aspects of privacy, especially with new information technologies being constantly developed. Smith and Shao [2007] survey privacy in e-commerce from point of view of both a consumer and a business. They find consumers either view privacy as a human value or relate it to economics of information.

In either case, the consumers prefer control of their personal information (or economic property). Businesses have lost market share or income because of privacy concerns, and thus are leaning toward giving consumers control over their personal information to alleviate privacy concerns and gain trust. Smith and Shao also survey privacy enhancing technologies and suggest anonymising technologies, while effective in several cases, may not be always viable or desirable. Acquisti et al. [2017] review research on assisting individuals' online privacy and security choices. They discuss the merits and limits of interventions that nudge users toward making the "right" choices, balancing revelation and protection of data.

4.6.1.4 Privacy-Preserving Applications

Other recent studies focus on the designing privacy-preserving applications. Campagna et al. [2017] present an architecture of a privacy-preserving virtual assistant for online services and the Internet of Things (IoT) that follows natural language commands for trigger-action tasks. They state that virtual assistants need to handle all of the users' personal information in order to comprehensively serve them. To preserve privacy, they propose a system that can run locally to store the information. Whereas Campagna et al.'s [2017] treatment of privacy is limited to information disclosure, we tackle nuanced notions of privacy understood as an ethical value.

Barry et al. [2017] propose a framework that adopts an Aristotelian virtue ethics concept, phronesis, for developing ethical mobile health apps. Phronesis describes the practical wisdom of gathering experience in a specific context. Barry et al. [2017] claim that applications with phronesis learn contextual client knowledge, and therefore make the right choices that inherently involve ethical reflection. However, their design does not address conflicts of different choices and priorities, which are common in social settings.

Chapter 5

Conclusions and Directions

5.1 Conclusions

This dissertation tackles nuanced notion of privacy, understood as an ethical value, from a sociotechnical viewpoint. Specifically we address the challenges of understanding social reality, i.e., understanding social expectations, social context, values, and ethics. We develop multiagent system techniques for privacy-respecting and ethics-aware social computing.

Arnor, a software engineering method, assists software developers to engineer personal agents by capturing stakeholders' social expectations, goals, and plans, and how these influence each other. Social expectation modeling via social norms in Arnor enables capturing accountability, and social experience modeling in Arnor helps incorporating fairness in decision-making.

Poros, a context reasoning approach, enables personal agents to understand social context, and infer contextually relevant social norms that respect stakeholders' privacy. Revealing and reasoning about social contexts to infer contextually relevant norms yields both transparency and accountability.

Ainur, a decision-making framework, provides personal agents with a decision-making ability to understand and reason about stakeholders' value preferences, and accordingly select ethically appropriate actions, thereby yields fairness.

5.2 Possible Directions for Future Dissertations

Future dissertations can be pursued in three dimensions—artificial intelligence, software engineering, and privacy.

5.2.1 Artificial Intelligence

In the artificial intelligence dimension, modeling white lies when revealing context, and incorporating affect in personal agents to promote social cohesion and privacy are promising future directions [Ajmeri et al., 2018d; Kalia et al., 2014].

Adopting argumentation and value-based reasoning to model and to infer preferences among values is another future direction [Ajmeri et al., 2016a, 2017a]. Value preferences can further help in inferring dominance relationship between norm instances, which methods such as Coco [Ajmeri et al., 2016b] can utilize in identifying and dealing with normative conflicts.

Whereas Ainur promotes fairness, and both Poros and Ainur yield satisfactory social experience, these approaches do not formally verify if the norms that emerge in the society are optimal. Formal approaches can be adopted here to compare normative specifications that emerge by computing normative tradeoffs and generating optimal normative specification [Kafalı et al., 2016, 2017].

5.2.2 Software Engineering

CrowdRE is a promising avenue for engaging crowd in human-intensive tasks such as capturing requirements for a personal agent like the Ringer SIPA described in Chapters 2 and 3, and the Pichu SIPA described in Chapter 4.

A first direction in the software engineering dimension is developing new techniques that incorporate creativity in the CrowdRE process to capture privacy requirements from stakeholders [Dhinakaran et al., 2018; Murukannaiah et al., 2016b].

A second direction is to design a requirements engineering method to assist software developers in developing ethical social applications [Ajmeri et al., 2017b, 2018b].

A third direction is to develop a knowledge evolution framework with constructs from the SIPA conceptual model [Ghaisas and Ajmeri, 2013]. The knowledge base captured in such a framework can enable a software developer engineering a new SIPA to jump start the development process by providing a SIPA's requirements skeleton.

5.2.3 Privacy

A first direction in the privacy dimension is developing a privacy-enhancing middleware based on Poros and Ainur to support ethical decision-making in social applications [Ajmeri et al., 2018a; Murukannaiah and Singh, 2015]. This middleware can absorb the abstractions in the world model, the social model, and the stakeholder model that are common across several SIPAs of a stakeholder for a coherent experience.

A second direction is to develop recommendation systems around Poros and Ainur techniques to tackle usability issues in privacy, security, and ethics.

BIBLIOGRAPHY

Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys*, 50(3), October 2017.

Nirav Ajmeri, Jiaming Jiang, Rada Chirkova, Jon Doyle, and Munindar P. Singh. Coco: Runtime reasoning about conflicting commitments. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 17–23, New York, 2016a. AAAI Press.

Nirav Ajmeri, Jiaming Jiang, Rada Y. Chirkova, Jon Doyle, and Munindar P. Singh. Coco: Runtime reasoning about conflicting commitments. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 17–23, New York, July 2016b. IJCAI.

Nirav Ajmeri, Chung-Wei Hang, Simon D. Parsons, and Munindar P. Singh. Aragorn: Eliciting and maintaining secure service policies. *IEEE Computer*, 50(12):50–58, December 2017a.

Nirav Ajmeri, Pradeep K. Murukannaiah, Hui Guo, and Munindar P. Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 230–238, São Paulo, May 2017b. IFAAMAS.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Ethics, values, and personal agents: Poster. In *Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security (HotSoS)*, page 17:1, Raleigh, April 2018a. ACM.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Designing ethical personal agents. *IEEE Internet Computing*, 22(2):16–22, 2018b.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Ethics, values, and personal agents: Poster. In *Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security (HotSoS)*, page 17:1, Raleigh, April 2018c. ACM.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 28–34, Stockholm, July 2018d. IJCAI.

Huib Aldewereld, Virginia Dignum, and Wamberto W. Vasconcelos. Group norms for multi-agent organisations. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 11(2):15:1–15:31, June 2016.

Natasha Alechina, Mehdi Dastani, and Brian Logan. Programming norm-aware agents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1057–1064, Valencia, 2012. IFAAMAS.

Natasha Alechina, Joseph Y. Halpern, Ian A. Kash, and Brian Logan. Decentralised norm monitoring in open multi-agent systems: (extended abstract). In *Proceedings of the 15th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1399–1400, Singapore, 2016. IFAAMAS.

Raian Ali, Fabiano Dalpiaz, and Paolo Giorgini. Reasoning with contextual requirements: Detecting inconsistency and conflicts. *Information and Software Technology*, 55(1):35–57, 2013.

Michael Anderson and Susan Leigh Anderson. Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. *Industrial Robot: An International Journal*, 42(4):324–331, 2015.

Giulia Andrighetto, Jordi Brandts, Rosaria Conte, Jordi Sabater-Mir, Hector Solaz, and Daniel Villatoro. Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS ONE*, 8(6):1–8, 06 2013.

Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. Value alignment or misalignment– what will keep systems accountable? In *Proceedings of the AAAI Workshop on AI, Ethics and Society*, pages 81–88, San Francisco, 2017. AAAI Press.

Marguerite Barry, Kevin Doherty, Jose Marcano Belisario, Josip Car, Cecily Morrison, and Gavin Doherty. mHealth for maternal mental health: Everyday wisdom in ethical design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, pages 2708–2756, Denver, 2017. ACM.

Trevor Bench-Capon and Sanjay Modgil. Norms and value based reasoning: Justifying compliance and violation. *Artificial Intelligence and Law*, 25(1):29–64, March 2017.

Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10–15):619–641, July 2007.

Tim Berners-Lee. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. Harper Business, New York, 1999.

Guido Boella, Leendert Van Der Torre, and Harko Verhagen. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2):71–79, 2006.

Guido Boella, Gabriella Pigozzi, and Leendert van der Torre. Normative framework for normative system change. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 169–176, Budapest, 2009a. IFAAMAS.

Guido Boella, Gabriella Pigozzi, and Leendert van der Torre. Normative systems in computer science - ten guidelines for normative multiagent systems. In Guido Boella, Pablo Noriega, Gabriella Pigozzi, and Harko Verhagen, editors, *Normative Multi-Agent Systems*, number 09121 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2009b. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany. URL `http://drops.dagstuhl.de/opus/volltexte/2009/1902`.

Alan Borning and Michael Muller. Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1125–1134, Austin, 2012. ACM.

Paolo Bresciani, Anna Perini, Paolo Giorgini, Fausto Giunchiglia, and John Mylopoulos. Tropos: An agent-oriented software development methodology. *Autonomous Agents and Multiagent Systems*, 8(3):203–236, May 2004.

Giovanni Campagna, Rakesh Ramesh, Silei Xu, Michael Fischer, and Monica S. Lam. Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant. In *Proceedings of the 26th International Conference on World Wide Web*, pages 341–350, Perth, Australia, 2017. International World Wide Web Conferences Steering Committee.

Cyrus D Cantrell. *Modern Mathematical Methods for Physicists and Engineers*. Cambridge University Press, Cambridge, 2000.

Amit K. Chopra and Munindar P. Singh. From social machines to social protocols: Software engineering foundations for sociotechnical systems. In *Proceedings of the 25th International World Wide Web Conference*, pages 903–914, Montréal, April 2016. ACM.

Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. A Bayesian approach to norm identification. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)*, pages 622–629, Amsterdam, August 2016. IOS Press.

Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in BDI agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 178–184, Melbourne, 2017. IJCAI.

Natalia Criado and Jose M. Such. Selective norm monitoring. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 208–214, New York, 2016. AAAI Press.

Karen Da Silva Figueiredo and Viviane Torres Da Silva. An algorithm to identify conflicts between norms and values. In *Proceedings of the 9th International Conference on Coordination, Organizations, Institutions, and Norms in Agent Systems (COIN)*, pages 259–274, St. Paul, MN, 2013. Springer. ISBN 978-3-319-07313-2.

Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: Values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1):79–107, Mar 2013.

Anind K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5 (1):4–7, January 2001. ISSN 1617-4909.

Venkatesh T. Dhinakaran, Raseshwari Pulle, Nirav Ajmeri, and Pradeep K. Murukannaiah. App review analysis via active learning: Reducing supervision effort without compromising classification accuracy. In *Proceedings of the 26th IEEE International Requirements Engineering Conference (RE)*, pages 170–181, Banff, 2018. IEEE.

Virginia Dignum. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4698–4704, Melbourne, 2017. IJCAI.

Martijn Egas and Arno Riedl. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society of London B: Biological Sciences*, 275 (1637):871–878, 2008. ISSN 0962-8452. doi: 10.1098/rspb.2007.1558.

Nuno Ferreira, Samuel Mascarenhas, Ana Paiva, Gennaro Di Tosto, Frank Dignum, John McBreen, Nick Degens, Gert Jan Hofstede, Giulia Andrighetto, and Rosaria Conte. An agent model for the appraisal of normative events based in in-group and out-group relations. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1220–1226, Bellevue, 2013. AAAI Press.

Ricard López Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. SoSharP: Recommending sharing policies in multiuser privacy scenarios. *IEEE Internet Computing*, 21(6):28–36, November 2017a.

Ricard López Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):5:1–5:29, March 2017b.

Batya Friedman, Peter H. Kahn Jr., and Alan Borning. Value sensitive design and information systems. In Kenneth Einar Himma and Herman T. Tavani, editors, *The Handbook of Information and Computer Ethics*, chapter 4, pages 69–101. John Wiley & Sons, Hoboken, New Jersey, 2008a.

Batya Friedman, Peter H. Kahn Jr., and Alan Borning. Value sensitive design and information systems. In Kenneth Einar Himma and Herman T. Tavani, editors, *The Handbook of*

Information and Computer Ethics, chapter 4, pages 69–101. John Wiley & Sons, Hoboken, New Jersey, 2008b.

GDPR, 2018. Regulation (EU) 2016/679 of the European Parliament and of the Council. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, 2018. Accessed: 2018-12-12.

Smita Ghaisas and Nirav Ajmeri. Knowledge-assisted ontology-based requirements evolution. In Walid Maalej and Anil Kumar Thurimella, editors, *Managing Requirements Knowledge (MaRK)*, pages 143–167. Springer Berlin Heidelberg, 2013.

Jianye Hao, Eunsuk Kang, Jun Sun, and Daniel Jackson. Designing minimal effective normative systems with the help of lightweight formal methods. In *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE)*, pages 50–60, Seattle, 2016. ACM.

Jim Hendler and Tim Berners-Lee. From the semantic web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence*, 174(2):156–161, 2010.

Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. Revani: Revision and verification of normative specifications for privacy. *IEEE Intelligent Systems*, 31(5):8–15, Sep 2016.

Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. Kont: Computing tradeoffs in normative multiagent systems. In *Proceedings of the 31st Conference on Artificial Intelligence (AAAI)*, pages 3006–3012, San Francisco, February 2017. AAAI.

Anup K. Kalia, Nirav Ajmeri, Kevin Chan, Jin-Hee Cho, Sibel Adalı, and Munindar P. Singh. A model of trust, moods, and emotions in multiagent systems, and its empirical evaluation. In *Proceedings of the 16th AAMAS Workshop on Trust in Agent Societies (Trust)*, Paris, May 2014.

Alex Kayal, Willem-Paul Brinkman, Rianne Gouman, Mark A. Neerincx, and M. Birna van Riemsdijk. A value-centric model to ground norms and requirements for epartners of children. In *Proceedings of the 9th Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 329–345, Cham, 2014. Springer.

Alex Kayal, Willem-Paul Brinkman, Mark A. Neerincx, and M. Birna van Riemsdijk. Automatic resolution of normative conflicts in supportive technology based on user values. *ACM Transactions on Internet Technology (TOIT)*, 18(4):41:1–41:21, May 2018.

Nadin Kökciyan and Pınar Yolum. Context-based reasoning on privacy in internet of things. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4738–4744, Melbourne, 2017. IJCAI.

Eugene F Krause. Taxicab geometry. *The Mathematics Teacher*, 66(8):695–706, 1973.

Marc Langheinrich. Privacy by design – principles of privacy-aware ubiquitous systems. In *Proceedings of the 3rd International Conference on Ubiquitous Computing (Ubicomp)*, volume 2201 of *Lecture Notes in Computer Science*, pages 273–291, Atlanta, 2001. Springer.

Derek Leben. A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, 19(2):107–115, June 2017.

Sean Luke, Claudio Cioffi-Revilla, Liviu Panait, Keith Sullivan, and Gabriel Balan. MASON: A multiagent simulation environment. *Simulation: Transactions of the Society for Modeling and Simulation International*, 81(7):517–527, July 2005.

Mehdi Mashayekhi, Hongying Du, George F. List, and Munindar P. Singh. Silk: A simulation study of regulating open normative multiagent systems. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 373–379, New York, 2016. AAAI Press.

John-Jules Ch. Meyer and Roel J. Wieringa, editors. *Deontic Logic in Computer Science: Normative System Specification*. Wiley, Chichester, United Kingdom, 1993.

Mihail Mihaylov, Karl Tuyls, and Ann Nowé. A decentralized approach for convention emergence in multi-agent systems. *Autonomous Agents and Multiagent Systems*, 28(5): 749–778, 2014.

Pradeep K. Murukannaiah and Munindar P. Singh. Platys Social: Relating shared places and private social circles. *IEEE Internet Computing*, 16(3):53–59, May 2012.

Pradeep K. Murukannaiah and Munindar P. Singh. Xipho: Extending Tropos to engineer context-aware personal agents. In *Proceedings of the 14th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 309–316, Paris, May 2014. IFAAMAS.

Pradeep K. Murukannaiah and Munindar P. Singh. Platys: An active learning framework for place-aware application development and its evaluation. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 24(3):19:1–19:32, May 2015.

Pradeep K. Murukannaiah, Nirav Ajmeri, and Munindar P. Singh. Engineering privacy in social applications. *IEEE Internet Computing*, 20(2):72–76, Mar 2016a. ISSN 1089-7801. doi: 10.1109/MIC.2016.30.

Pradeep K. Murukannaiah, Nirav Ajmeri, and Munindar P. Singh. Acquiring creative requirements from the crowd: Understanding the influences of personality and creative potential in crowd re. In *Proceedings of the 24th IEEE International Requirements Engineering Conference (RE)*, pages 176–185, Beijing, 2016b. IEEE.

Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy expectations and preferences in an IoT world. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS)*, pages 399–412, Santa Clara, 2017. USENIX Association.

Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review (KER)*, 31: 142–166, March 2016.

Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, February 2004.

Helen Nissenbaum. A contextual approach to privacy online. *Dædalus, the Journal of the American Academy of Arts & Sciences*, 140(4):32–48, Fall 2011.

Charles Noussair and Steven Tucker. Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, 43(3):649–660, 2005.

Serafim Opricovic and Gwo-Hshiung Tzeng. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, 156(2):445–455, 2004.

Federica Paci, Anna Squicciarini, and Nicola Zannone. Survey on access control for community-centered collaborative systems. *ACM Computing Surveys*, 51(1):6:1–6:38, January 2018.

Eric Pacuit. Voting methods. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition, 2017.

William L. Prosser. Privacy. *California Law Review*, 48(3):383–423, August 1960.

Iyad Rahwan, Thomas Juan, and Leon Sterling. Integrating social modelling and agent interaction through goal-oriented analysis. *Computer Systems Science and Engineering*, 21(2):87–98, 2006.

John Rawls. Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs*, 14(3):223–251, 1985.

Mark O. Riedl and Brent Harrison. Using stories to teach human values to artificial agents. In *Proceedings of the Workshops of the 30th AAAI Conference on Artificial Intelligence: AI, Ethics, and Society*, pages 105–112, Phoenix, 2016. AAAI Press.

Milton Rokeach. *The Nature of Human Values*. Free Press, New York, 1973.

Scott Ruoti, Brent Roberts, and Kent Seamons. Authentication melee: A usability analysis of seven web authentication systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 916–926, Florence, 2015. International World Wide Web Conferences Steering Committee.

Bastin Tony Roy Savarimuthu, Stephen Cranefield, Martin K. Purvis, and Maryam A. Purvis. Norm emergence in agent societies formed by dynamically changing networks. *Web Intelligence and Agent Systems: An International Journal*, 7(3):223–232, 2009.

Sebastian Schnorf, Aaron Sedley, Martin Ortlieb, and Allison Woodruff. A comparison of six sample providers regarding online privacy benchmarks. In *In Proceedings of the SOUPS Workshop on Privacy Personas and Segmentation*, pages 1–7, Menlo Park, 2014.

Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

Sandip Sen and Stéphane Airiau. Emergence of norms through social learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1507–1512, Hyderabad, 2007. Morgan Kaufmann Publishers Inc.

Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael Wooldridge, Javier Morales, and Carlos Ansótegui. Moral values in norm decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1294–1302, Stockholm, July 2018. IFAAMAS.

Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013.

Paul Smart, Elena Simperl, and Nigel Shadbolt. A taxonomic framework for social machines. In Daniele Miorandi, Vincenzo Maltese, Michael Rovatsos, Anton Nijholt, and James Stewart, editors, *Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society*, Computational Social Sciences, pages 51–85. Springer, Cham, Switzerland, 2014.

Rhys Smith and Jianhua Shao. Privacy and e-commerce: A consumer-centric perspective. *Electronic Commerce Research*, 7(2):89–116, 2007.

Derek J. Sollenberger and Munindar P. Singh. Kokomo: An empirically evaluated methodology for affective applications. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 293–300, Taipei, May 2011. IFAAMAS.

Daniel J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3): 477–564, 2006.

Kaj Sotala. Defining human values for value learners. In *Proceedings of the Workshops of the 30th AAAI Conference on Artificial Intelligence: AI, Ethics, and Society*, pages 113–123, Phoenix, 2016. AAAI Press.

Sarah Spiekermann. The challenges of privacy by design. *Communications of the ACM (CACM)*, 55(7):38–40, July 2012.

Sarah Spiekermann and Lorrie Faith Cranor. Engineering privacy. *IEEE Transactions on Software Engineering (TSE)*, 35(1):67–82, January 2009.

Jose M. Such. Privacy and autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1–7, Melbourne, 2017. IJCAI.

Toshiharu Sugawara. Emergence and stability of social conventions in conflict situations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 371–378, Barcelona, 2011. AAAI Press.

James Stacey Taylor. Privacy and autonomy: A reappraisal. *The Southern Journal of Philosophy*, 40(4):587–604, 2002.

M. Birna van Riemsdijk, Louise Dennis, Michael Fisher, and Koen V. Hindriks. A semantic framework for socially adaptive agents: Towards strong norm compliance. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 423–432, Istanbul, May 2015a. IFAAMAS.

M. Birna van Riemsdijk, Catholijn M. Jonker, and Victor Lesser. Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1201–1206, Istanbul, 2015b. IFAAMAS.

Daniel Villatoro, Jordi Sabater-Mir, and Sandip Sen. Robust convention emergence in social networks through self-reinforcing structures dissolution. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 8(1):2:1–2:21, April 2013.

Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

Fei-Yue Wang, Kathleen M. Carley, Daniel Zeng, and Wenji Mao. Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems*, 22(2), 2007.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

Alan F. Westin. *Privacy and Freedom*. Atheneum, New York, 1967.

Alan F. Westin. Social and political dimensions of privacy. *Journal of Social Issues*, 59(2): 431–453, 2003.

Michael Winikoff and Lin Padgham. *Developing Intelligent Agent Systems: A Practical Guide*. Wiley, Chichester, UK, 2004. ISBN 0470861207.

Michael Wooldridge, Nicholas R. Jennings, and David Kinny. The Gaia methodology for agent-oriented analysis and design. *Autonomous Agents and Multi-Agent Systems*, 3(3): 285–312, September 2000.

Chao Yu, Minjie Zhang, Fenghui Ren, and Xudong Luo. Emergence of social norms through collective learning in networked agent societies. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 475–482, Saint Paul, 2013. IFAAMAS.

APPENDICES

Appendix A

Arnor: Surveys

A.1 Pre-participation Survey

1. How long (in years) is your academic programming experience? Count a semester as six months or less depending on your programming effort that semester. An approximation is fine.
   - <1
   - 1–2
   - 2–6
   - 6+
2. How long (in years) is your industry programming experience? Count full-time experience and approximate part-time experience.
   - <1
   - 1–2
   - 2–6
   - 6+
3. How familiar are you with conceptual modeling? Conceptual modeling includes requirements engineering, system specification, architectural design, and so on.
   - Not familiar

- ○ Familiar with the concepts, but no practical experience
- ○ Familiar and some practical experience from an earlier academic project
- ○ Familiar and some practical experience from the industry
- ○ Expert (e.g., system architect, requirements engineer, researcher in this area, etc.)
- ○ Other

A.2 Time and Effort Survey

Answer this survey after each work session. It is best to answer the survey right after a work session. A work session is typically a sitting.

Please be as accurate as possible. If you make a mistake such as submitting a survey twice or incorrectly reporting something, please email the investigator so that we can fix the mistake.

1. How long was this work session? In hours:minutes (e.g., 03:30). Please exclude interruptions from the session duration. If an interruption is long enough, it might be better to treat a session as two.

2. How do you rate the difficulty of the work your performed in this session? Note that session duration and difficulty are not necessarily related. Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.

3. What did you do in this work session?

   - □ Read project description
   - □ Learning and understanding methodology to develop socially-aware application
   - □ Learning Lucidchart
   - □ Reading other material (specify in comments)
   - □ Modeling
   - □ Workout Social Benefit Function (specify in comments)
   - □ Implementation (specify in comments)
   - □ Documentation (specify in comments)
   - □ Other

4. What is an approximate breakdown of this work session? For example, [reading project description (10%) + coding (65%) + documentation (15%)]. You can breakdown however you want (not necessary to breakdown three way like in the example). Also, briefly describe what you did in this session.

5. Any additional comments?

6. Anything else you may want to mention about this work session.

A.3 Post Survey

Overall Time and Difficulty

1. Give an estimate of the overall time you spent in hours:minutes (e.g., 03:30) to understand the project requirements and to prepare the requirement specification (models)?

2. Give an estimate of the overall time you spent in hours:minutes (e.g., 03:30) to design the social benefit function?

3. Give an estimate of the overall time you spent in hours:minutes (e.g., 03:30) to implement the project?

4. Give an estimate of the overall time you spent in hours:minutes (e.g., 03:30) to test the project?

5. Give an estimate of the overall time you spent in hours:minutes (e.g., 03:30) to document the project?

6. How easy were the following phases? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.

    - Understanding Requirements
    - Preparing requirement specification (models)
    - Implementation
    - Testing
    - Documentation

7. In what aspects do you think the application violates your privacy?

Usability

Consider yourself as a user of your application.

1. List all actions that you can perform using the application.
2. To what extent do you think the application preserves your privacy? Answer on a numeric scale 1–7 where 1 is *application is privacy-preserving*, and 7 is *application is privacy violating*.
3. How usable is your application considering the following aspects? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
    - How easy is it to accomplish the actions the first time you use the application?
    - Once you have learned the application, how quickly can you perform the actions?
    - When you return to the application after a period of not using it, how easily can you re-establish proficiency?
    - How many errors do you make when you use the application? (1: not many, 7: quite a lot)
    - How severe are the errors? (1: not very severe, 7: very severe)
    - How easily can it recover from the errors? (1: quickly, 7: takes a long time)
    - How pleasant is it to use the application? (1: very pleasant, 7: not at all pleasant)
    - How easy is it to accomplish the actions the first time you use the application?
    - Once you have learned the application, how quickly can you perform the actions?
    - When you return to the application after a period of not using it, how easily can you re-establish proficiency?
    - How many errors do you make when you use the application? (1: not many, 7: quite a lot)
    - How severe are the errors? (1: not very severe, 7: very severe)
    - How easily can it recover from the errors? (1: quickly, 7: takes a long time)
    - How pleasant is it to use the application? (1: very pleasant, 7: not at all pleasant)

Methodology and Project Development

Understanding requirements, implementation, testing and documentation

1. How clear were (or are) the requirements of this project for you? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - When you started the project
   - Now, at the end of the project
2. To what extent did the methodology help you in the following aspects of the project? Answer on a numeric scale 1–7 where 1 is *didn't help me at all*, and 7 is *helped me a lot*.
   - Understanding the project requirements
   - Implementing the project requirements
   - Testing
   - Documentation
3. How easy is it to understand your implementation for someone else? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - Who knows the methodology
   - Who does not know the methodology
4. How easy is it to understand your documentation for someone else? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - Who knows the methodology
   - Who does not know the methodology
5. How easy was (or is) it to identify or incorporate actors? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - (For you to) Identify all actors in the scenario description?
   - (For someone else to) Identify all actors in your implementation?
   - (For you to) Incorporate a new actor in your implementation?
6. How easy was (or is) it to identify or incorporate actions? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.

- (For you to) Identify all actions in the scenario description?
- (For someone else to) Identify all actions in your implementation?
- (For you to) Incorporate a new action in your implementation?

7. How easy was (or is) it to identify or incorporate contexts in the actions? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - (For you to) Identify contexts in the scenario description?
   - (For someone else to) Identify contexts in your implementation?
   - (For you to) Incorporate a new context in your implementation?
8. How easy was (or is) it to identify or incorporate norms? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - (For you to) Identify all norms in the scenario description?
   - (For someone else to) Identify all norms in your implementation?
   - (For you to) Incorporate a new norms in your implementation?
9. How easy was (or is) it to identify norm conflicts and inconsistencies? Answer on a numeric scale 1–7 where 1 is *very easy*, and 7 is *very difficult*.
   - (For you to) Identify norm conflicts in the scenario description?
   - (For someone else to) Identify conflict resolution in your implementation
10. In your implementation, how did you resolve conflicts between norms? If you answered this in the report, please paste it here.
11. Do you think the methodology missed some crucial step that could have helped in understanding the project requirements, implementation, testing, or documentation?
12. Any other comments or feedback.

Appendix B

Ainur: Surveys

B.1 Privacy Attitude Survey

We measure privacy attitudes of the survey participants using [Schnorf et al., 2014]'s survey instrument.

1. For personal purposes, how often do you normally use the Internet?
    - Every hour or more often
    - Every few hours
    - Once or twice a day
    - Multiple times per week
    - Once per week or less often
2. Which of the following best describes when you buy or try out new technology?
    - Among the first people
    - Before most people, but not among the first
    - Once many people are using it
    - Once most people are using it
    - I don't usually buy or try out new technology
3. In general, how would you rate technology's impact on people's lives?

- Very positive
- Somewhat positive
- Neither positive nor negative
- Somewhat negative
- Very negative

4. How comfortable or uncomfortable are you with information about yourself on the Internet that anyone can find and see?

   - Very comfortable
   - Somewhat comfortable
   - Neither comfortable nor uncomfortable
   - Somewhat uncomfortable
   - Very uncomfortable

5. How comfortable or uncomfortable are you providing information about yourself online to a business or organization?

   - Very comfortable
   - Somewhat comfortable
   - Neither comfortable nor uncomfortable
   - Somewhat uncomfortable
   - Very uncomfortable

6. To what extent do you think information about yourself on the Internet, that is available to another person, business or organization, might cause you negative experiences?

   - Not at all
   - To a small extent
   - To a moderate extent
   - To a large extent
   - To a very great extent

7. Have you had any negative experiences because information about yourself on the Internet was available to another person,business or organization?

    - Yes
    - No

8. If your answer for the previous question was "Yes", recall the most negative experience you had due to information about yourself on the Internet. What consequences were there?

    - Unwanted commercial offers or spam
    - Reputation damage or embarrassing situation
    - Stalking or harassment
    - Financial loss
    - Identity theft

9. If your answer for the question on negative experience was "Yes", recall again the negative experience you had due to information about yourself on the Internet. How severe were the consequences?

    - Not at all severe
    - Slightly severe
    - Moderately severe
    - Very severe
    - Extremely severe

B.2 Policy Survey

Deciding the correct privacy policy for location check-in is non-trivial. It involves weighing several contextual factors. To overcome this challenge, Aron, a graduate student, decides to develop a policy recommender application for himself that would suggest him an appropriate policy for location check-in on social media.

Consider you are Aron. For each context, select the correct policy based on the companion. Note that some combinations may not seem realistic; make a guess in those cases.

1. Attending a graduation ceremony

2. Presenting a research paper at an international conference
3. Studying at a library during the day
4. Visiting an airport at night
5. Hiking a mountain at night
6. Being stuck in a hurricane
7. Going to a bar with a fake ID
8. Going to a drug rehabilitation center

Companion	Check-in Policy			
	Share with all	Common friends	Companions	No one
Alone	○	○	○	○
Colleague	○	○	○	○
Friend	○	○	○	○
Family member	○	○	○	○
Crowd	○	○	○	○