Vision: I envision ethics-aware and privacy-respecting social machines that facilitate natural interactions among autonomous social entities (people and organizations). To develop the foundations for such social machines, I adopt a sociotechnical stance in which agents (as technical entities) help autonomous social entities or principals (people and organizations). This multiagent conception of a sociotechnical system (STS) captures how ethical and social concerns arise in the mutual interactions of multiple stakeholders. In pursuit of developing the foundations which would enable us to realize ethical STSs [9], I intend to make fundamental contributions in artificial intelligence and multiagent systems.

Research Plan

The challenge in realizing such ethical STSs is — how to understand social reality, i.e., how to understand social expectations, social context, values, and ethics. Specifically, how can we design STSs from the ground up to be secure, ethical, and fair? I approach this challenge of developing ethical STSs from the ground up via three thrusts:

Modeling Social Intelligence. 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 satisfactory experience. In a society where an agent acts on behalf of a stakeholder (a human user), it is important that the agent understands these nuances in social interactions.

To address the challenge of modeling social intelligence, I develop Arnor [5], a method which facilitates modeling stakeholders' actions and expectations, and how these influence each other. Arnor employs Singh's [10] conception of social norms to capture social expectations, and incorporates argumentation constructs for sharing decision rationale. Social expectation modeling via social norms in Arnor enables capturing *accountability*, and social experience modeling in Arnor helps incorporating *fairness* in decision-making.

In [8], we develop models to characterize how emotions influence norm outcomes. Modeling affect and incorporating emotions in personal agents are research directions — I intend to pursue — which will help in realizing agents which promote better social cohesion.

Understanding Social Context. Social 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. To address the challenge of understanding social context, I develop Poros [1], an approach for building intelligent agents that carry out enriched interactions where deviating agents share selected elements of their context as explanations, and other agents respond appropriately to the deviations in light of the received information. Revealing and reasoning about social contexts to infer contextually relevant norms yields both *transparency* and *accountability*.

I am interested in understanding the abstractions of shared context and modeling white lies. When agents share context, it is important for them to know *what to share* and *what not to share*. If agents can understand abstractions of the context being shared, they can better ensure privacy of their users while maintaining transparency.

Reasoning about Value Preferences. 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 pursue or not pursue those actions. Each action a Poros agent executes potentially promotes or demotes one or more values. Being aware of these values and having an ability to reason about them helps an agent select ethical actions and yield satisfactory experience. To address the challenge of reasoning about values, I develop Elessar [2], a framework to design such ethical agents. Elessar incorporates a multicriteria decision-making method to aggregate value preferences of stakeholders and select an ethically appropriate action — balancing *utilitarianism* and *egalitarianism*. Elessar provides agents with a decision-making ability to understand and reason about stakeholders' value preferences, and accordingly select ethically appropriate actions, thereby yields *fairness*.

I have an interest in formal verification and reasoning. Future directions in verification and reasoning are two-fold. First direction is adopting argumentation to model and to infer preferences among values [3]. Second direction is to generate optimal normative specification. Whereas Elessar promotes fairness, and both Poros and Elessar yield satisfactory social experience, these approaches do not formally verify if the norms that emerge in the society are optimal on metrics including liveness, safety, and robustness. I intend to develop formal approaches on lines of my other recent works [4, 6, 7] to compare normative specifications that emerge by computing tradeoffs and generating optimal normative specification.

Collaboration. As part of the AI group of researchers at Bristol, I will contribute toward developing new computing technologies and infrastructure that will enable realizing ethical STSs.

At Bristol, I will seek collaborations with (1) Professors Peter Flach, Nello Cristianini, and Andrew Charlesworth (from the Law School) on the broad theme of understanding ethical and privacy issues when designing STSs, (2) Professor Kerstin Eder on applying formal verification and reasoning techniques and developing multiagent simulations to evaluate STS specifications for ethicality, (3) Professor Weiru Liu and Dr. Iván Palomares on decision-making by understanding value preferences and social context, (4) Professor Awais Rashid on studying privacy and security as ethical-values and conducting large-scale user studies.

Outside Bristol, I will continue to collaborate with (1) Professor Munindar Singh at NCSU and Dr. Pradeep Murukannaiah at Delft University of Technology on the broader theme of developing ethical STSs, and (2) Dr. Özgür Kafalı at University of Kent on developing formal reasoning and verification techniques for STSs.

I will also seek industry collaborations with (1) Dr. Anup Kalia at IBM Research, New York on incorporating emotions and affect in agents and (2) Dr. Zhe Zhang at IBM Watson, Raleigh on developing NLP techniques for learning values (and value preferences) of users from text.

Expected Outcomes. I expect to publish my research findings in prestigious journals and conferences in Artificial Intelligence (e.g., JAIR, AIJ, JAAMAS, AAAI, IJCAI, AAMAS), in Software Engineering (e.g., TSE, TOSEM, ICSE, FSE, RE), and Security and Privacy (e.g., IEEE S&P, USENIX Security, CCS, SOUPS).

References

[1] 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 2018. IJCAI.

[2] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1–9, Auckland, May 2020. IFAAMAS.

[3] 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 2017.

[4] 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, 2016. AAAI Press.

[5] 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 2017. IFAAMAS.

[6] 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.

[7] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. Specification of sociotechnical systems via patterns of regulation and control. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 29(1):7:1–7:50, December 2019.

[8] Anup K. Kalia, Nirav Ajmeri, Kevin Chan, Jin-Hee Cho, Sibel Adalı, and Munindar P. Singh. The interplay of emotions and norms in multiagent systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 371–377, Macau, August 2019. IJCAI.

[9] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1–5, Auckland, May 2020. IFAAMAS. Blue Sky Ideas Track.

[10] 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.