Face-work for Human-Agent Joint Decision-Making

JiHyun Jeong
Information Science
Cornell University, Ithaca, NY
jihyun@infosci.cornell.edu

Guy Hoffman
Mechanical and Aerospace Engineering
Cornell University, Ithaca, NY
hoffman@cornell.edu

Abstract

We propose a method to integrate face-work, a common social ritual related to trust, into a decision-making agent that works collaboratively with a human. Face-work is a set of trust-building behaviors designed to "save face" or prevent others from "losing face." This paper describes the design of a decision-making process that explicitly considers face-work as part of its action selection. We also present a simulated robot arm deployed in an online environment that can be used to evaluate the proposed method.

Introduction

This paper describes a human-robot joint decision-making algorithm that integrates the social ritual of *face-work*. Face-work involves the actions taken to maintain one's own or another's "face", a positive self-image claimed in a social context (Goffman 1967). We first formalize a synchronous decision-making interaction between a human and a robot and present a simulation setup to elicit such interactions. We then illustrate the implementation of an algorithm that evaluates decisions while accounting for face threats in the action-selection process.

When people collaborate, especially on making decisions, the affect-based trust they have in each other is crucial to make good decisions together. The importance of affective trust is demonstrated in multiple studies; for a review, see Lee and See (2004). In particular, care and concern for another constitute a core basis for trust in peer relationships (McAllister 1995). Figure 1 illustrates this relationship between face-work, affect, and trust.

To maintain affective trust, people engage in social rituals that are designed to "save face" for themselves, or prevent others from "losing face". The loss of one's face, which can be induced by criticism or disagreement, is often accompanied by negative emotions such as shame (Goffman 1967). For example, if a collaborator publicly dismisses a peer, the loss of face of the peer can prevent them from raising future ideas. Given the importance of face in affective trust, humans commit to face-work as a social ritual to defend against and correct face threats.

Similar social and interaction aspects are also pertinent when humans and agents make joint decisions. Achieving productive decision-making with humans and algorithms can be challenging (Green and Chen 2019) and immediate performance gain is not always the best goal to strive for. For example, Elmalech et al. (2015) present cases when an algorithm that suggests a suboptimal option better matching a human's intuition leads to better performance over time, compared to suggesting optimal solutions that counter human intuitions. We posit that face-related factors can be another motivation for an algorithm or robot to prefer a suboptimal suggestion, i.e., when the optimal decision is a face-threat that can hamper social relationships.

Prior work in human-computer interaction already suggests that an agent can perform face-work to prevent humans from losing face, which can result in negative emotions that undermine affect-based trust. Reeves and Nass (1996) found that humans treat computers as social actors and can take offense to impolite computers. Similarly, Takayama, Groom, and Nass (2009) found that human subjects were sensitive to disagreements from robots. Jung (2017) emphasize the need for affective grounding in human-robot collaborations. In our prior work, humans ascribed social intentions to robot actions, and a robot's misaligned action was sometimes interpreted as contemptuous and resulted in mistrust (Law et al. 2019).

Politeness is one type of face-saving strategy, and in fact has been studied extensively in human-agent and human-robot interaction. Research has found that applying politeness strategies improved perception of a robot and increased its trustworthiness. Participants felt more positive towards a robot that applied a distancing politeness strat-

Figure 1: Humans perform face-work to protect emotions attached to face (Goffman 1967). Affect-based elements such as care and concern for another constituted a core basis for trust in peer relationships (McAllister 1995).
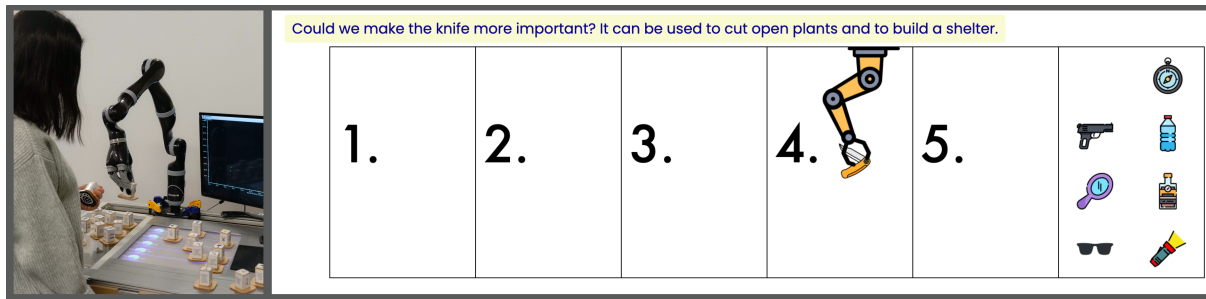
Figure 2: The user interface on the right allows both the human and the virtual robot to move objects to a box that corresponds to the preferred rank of the object. The interaction is designed to resemble a human-robot decision-making setup we used in prior work (left).

egy, having the disagreeing voice come from a box separate from the robot's body (Takayama, Groom, and Nass 2009). Robotic assistants that gave advice using hedges and polite discourse markers were perceived as more considerate and likeable, and less controlling (Torrey, Fussell, and Kiesler 2013). Similarly, politeness improved perceptions such as fairness and friendliness of a access-control robot (Inbar and Meyer 2019). Agents that politely engaged in small talk were viewed as more trustworthy by extroverts (Bickmore and Cassell 2001), while automation that had "poor etiquette" of interrupting humans, were trusted less (Parasuraman and Miller 2004; Miller 2005).

Our work extends this prior literature in two ways. First, the robot or agent in the above studies was almost always in an assistive role, and the human had the sole agency to accept or reject the robot's suggestions. In contrast, we consider an interaction that resembles a process between equal partners with shared agency. This means that both human and the robot can accept or reject each other's suggestions. They continually negotiate their ideas and preferences back and forth, which provides opportunities for social breakdowns.

Second, prior works study politeness as a possible predictive construct in human-robot and human-agent interaction. In our work, we propose to systematically include facework in the decision-making process of the robot. We do so by presenting a computational framework for joint decision-making that integrates face-work in the agent's algorithm. We also describe a simulated environment that can be used to evaluate this framework.

Joint Decision-Making Context

To demonstrate the integration of face-work in human-robot interaction, we implement a simulator with a virtual robot. We chose this method for ease of user testing while keeping some physical affordances of robots (Figure 2). This way we can use spatial cues as embodied elements (e.g., for epistemic actions) and use the robot's nonverbal behavior for face-saving gestures.

Adapted Desert Survival Task

Our task is an adapted version of the Desert Survival Problem (Lafferty and Pond 1974). A participant first ranks a set of given items in the order of importance for survival, then collaborates with a teammate to arrive at a group solution. The problem and its variations have been used widely in teamwork and group decision-making studies (Burke and Barron 2015; Hall and Watson 1970).

We adapt the task to be applicable to more general group decisions where a group not only rank, but also accept and reject different options. Instead of having the participants rank all items, we ask them to choose five out of eight items and rank only those they have selected. Selecting the top five presents a clear bound between accepted and rejected items, which is the case in many group decisions. For instance, when a group engages in brainstorming, some ideas will be accepted, some rejected, and some may be preferred over others.

Interface and Procedure

Decisions to rank an item are expressed by moving an object to a location that corresponds to the desired rank. Both parties can add an object to any rank, remove an object, or swap the location of two objects. The human ranks an object by dragging the icon representing the object to a box with the preferred rank. The virtually embodied robot moves an object through a series of animations: moving to the location of an object, picking it up, moving to the designated rank, and dropping the object. The series of animations take around 7 seconds to move one object, which is slower than the speed at which a human is capable of dragging and dropping an object. This design preserves the difference in speed between a human and a real-world collaborative robot.

The human and the robot take turns to negotiate the ranking. To simulate a real-world interaction more accurately, we implemented a flexible turn-taking paradigm: The human can choose to move as many objects as they want for a prolonged period of time in one turn. During that time, the robot does not move any objects. When a human wants to yield their floor to the robot, they can pause for the robot to start moving. The robot will then move objects until the human takes the floor back, which can only happen during a robot pause. Manipulating the length of robot pause times allows for the robot to also take multiple actions in one turn.

In each move, the robot provides a reason for ranking an object higher or lower. The final team rank can be submitted

when the human and the robot both agree on a solution.

Agent Design

We adapt a formalization of the survival task from Bergner et al. (2016). Each object has an integer identifier i, and a ranking at time step j is represented as an array R_j of eight numbers corresponding to the location (rank) of each object. Different objects can share the rank 6 when they are not ranked in the top five. A robot move at turn j is represented by a tuple $m_j = (i, orig, dest)$ including the identifier of the object in the array, its current (origin) rank and its destination rank.

The robot considers each possible move candidate m_j and the ranking it produces, R_j, by applying m_j to its previous state of ranking R_{j-1}. It then evaluates how far each candidate ranking is to the robot's preferred ranking. We use the sum of two widely used evaluation metrics for this task, minimizing distance ($D(R_j)$) from the current ranking to the desired ranking (Burgoon et al. 2000) and maximizing the number of concordant pairs ($C(R_j)$) between the two (Bergner et al. 2016).

Integrating Machine Face-Work

We define an additional measure of decorum based on face-work, $F(m_j)$, for a move m_j. Thus, each move m_j that achieves ranking R_j, will be evaluated using a combination of metric that evaluates the ranking produced by a move, together with an assessment of how socially appropriate the move is. The robot will execute the action m_j that leads to the maximum value of $(C(R_j) - D(R_j)) * F(m_j)$.

To determine a valid model for F, we incorporate the following politeness strategies (adapted from Brown and Levinson (1987)) to mitigate face threatening acts: seek agreement, avoid disagreements, and make indirect requests. First, for all candidate moves m_j, F will be initialized to $F(m_j) = 1$. Next, we illustrate how we compute $F(m_j)$ for each candidate.

Seek agreement: The robot looks for items that it and the human both agree on to rank first. Let us represent the robot's preferred ranking[1] as $R^A = [r_1^A, \ldots, r_8^A]$ and the human's preferred ranking, expressed in the initial stage of the task when working on their own, as $R^H = [r_1^H, \ldots, r_8^H]$. For any move $m_j = (i, orig, dest)$ where $dest = r_i^A = r_i^H$, has the value of $F(m_j) = 2$.

Avoid disagreements: We avoid any moves that directly reverse any actions that the human has done in their previous turn. A human's last turn can be represented as an array consisting of tuples of human moves. For convenience, we represent all human moves on their last turn $m_{H_1}, m_{H_2}, \ldots, m_{H_n}$ as $m_H = (i_H, orig_H, dest_H)$. Any candidate robot move $m_j = (i, orig, dest)$ that meets $i = i_H$ and $dest = orig_H$ is considered a reversal, and will have the value of $F(m_j) = 0$.

We also avoid any moves that repeat any of the robot's actions, since the need to do so implies that the human

---

[1]We use A for "agent" to avoid confusion with the notation for ranking

had disagreed with the choice before and might be viewed as a dismissing insistence. We represent all previous robot moves $m_{A_1}, m_{A_2}, \ldots, m_{A_k}$, as $m_A = (i_A, orig_A, dest_A)$. Any candidate robot move $m_j = (i, orig, dest)$ that meets $i = i_A$ and $dest = dest_A$ is considered a repeat, and will have the value of $F(m_j) = 0$.

A useful side-effect of using the above face-work evaluation function is that it can also implicitly signal an ostensible compromise. For instance, if the human places an item on the third rank after the robot placed it on first, the robot will then try placing it on second instead to avoid reversing the human's previous decision.

Make indirect requests: Lastly, we also incorporate verbal politeness markers and make indirect requests using questions and hedges. For example, the robot asks "Could we make the knife more important?" instead of saying "Make the knife more important."

Evaluation Plan

We plan to conduct experiments to test whether these strategies are effective at maintaining human's trust. As a first step, we designed a within-subjects study, where participants interact with both agents, one that performs face-work as indicated above and one that does not. The order in which these agents are presented is randomized. There are two variations of the survival task, for which the order is also randomized. This measure is taken to avoid preferences carrying over between conditions. We will perform a manipulation check to confirm that our agent can reliably mitigate face-threats using scales such as the Revised Instructional Face-Support (RIFS) scale (Kerssen-Griep, Trees, and Hess 2008). Lastly, human's trust towards the agent will be measured using validated questionnaires such as those proposed by Madsen and Gregor (2000).

Conclusion

This work is part of an on-going work of developing socially intelligent agents that can maintain human's trust in group interactions. We introduce a joint decision-making setting between a human and a virtual agent, and implement face-work into the algorithm of the agent's decision.

In future work, this platform can host more intelligent agents to interact with humans that considers more of the social contexts and relationships. We make a crass distinction of face-threats in this work and only deal with one mitigation strategy: to avoid it. However, there are degrees of impositions and a spectrum of mitigation strategies. We can also incorporate different social factors from literature that might determine the acceptable degree of face threats, such as power or social distance. Additionally, future work will also examine ethical challenges of developing a socially intelligent agent that learns to negotiate, as there can be negative implications for socially deceptive behaviors that can manipulate a person's affect and trust.

Acknowledgements

References

Bergner, Y.; Andrews, J. J.; Zhu, M.; and Gonzales, J. E. 2016. Agent-based modeling of collaborative problem solving. *ETS Research Report Series* 2016(2):1–14.

Bickmore, T., and Cassell, J. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 396–403. ACM.

Brown, P., and Levinson, S. C. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Burgoon, J. K.; Bonito, J. A.; Bengtsson, B.; Cederberg, C.; Lundeberg, M.; and Allspach, L. 2000. Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in human behavior* 16(6):553–574.

Burke, R., and Barron, S. 2015. Lost at sea. In *Project Management Leadership*. John Wiley & Sons, Inc. 351–354.

Elmalech, A.; Sarne, D.; Rosenfeld, A.; and Erez, E. 2015. When suboptimal rules. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 1313–1319. AAAI Press.

Goffman, E. 1967. On face-work. *Interaction ritual* 5–45.

Green, B., and Chen, Y. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–24.

Hall, J., and Watson, W. H. 1970. The effects of a normative intervention on group decision-making performance. *Human Relations* 23(4):299–317.

Inbar, O., and Meyer, J. 2019. Politeness counts: Perceptions of peacekeeping robots. *IEEE Transactions on Human-Machine Systems* 49(3):232–240.

Jung, M. F. 2017. Affective grounding in human-robot interaction. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 263–273. IEEE.

Kerssen-Griep, J.; Trees, A. R.; and Hess, J. A. 2008. Attentive facework during instructional feedback: Key to perceiving mentorship and an optimal learning environment. *Communication Education* 57(3):312–332.

Lafferty, J. C., and Pond, A. W. 1974. *The desert survival situation: A group decision making experience for examining and increasing individual and team effectiveness*. Human Synergistics.

Law, M. V.; Jeong, J.; Kwatra, A.; Jung, M. F.; and Hoffman, G. 2019. Negotiating the creative space in human-robot collaborative design. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. ACM.

Lee, J. D., and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46(1):50–80.

Madsen, M., and Gregor, S. 2000. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, volume 53, 6–8. Citeseer.

McAllister, D. J. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal* 38(1):24–59.

Miller, C. A. 2005. Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods. In *Proceedings of the 1st international conference on augmented cognition*, 22–27. Citeseer.

Parasuraman, R., and Miller, C. A. 2004. Trust and etiquette in high-criticality automated systems. *Communications of the ACM* 47(4):51–55.

Reeves, B., and Nass, C. I. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

Takayama, L.; Groom, V.; and Nass, C. 2009. I'm sorry, dave: i'm afraid i won't do that: social aspects of human-agent conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099–2108. ACM.

Torrey, C.; Fussell, S.; and Kiesler, S. 2013. How a robot should give advice. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, 275–282. IEEE Press.