The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
AI, Ethics, and Society: Technical Report WS-16-02

Relational Enhancement: A Framework for Evaluating and Designing Human-Robot Relationships

Jason R. Wilson, Thomas Arnold, Matthias Scheutz
Human-Robot Interaction Lab
Tufts University
200 Boston Ave
Medford, MA 02155

Abstract

Much existing work examining the ethical behaviors of robots does not consider the impact and effects of long-term human-robot interactions. A robot teammate, collaborator or helper is often expected to increase task performance, individually or of the team, but little discussion is usually devoted to how such a robot should balance the task requirements with building and maintaining a "working relationship" with a human partner, much less appropriate social relations outside that team. We propose the "Relational Enhancement" framework for the design and evaluation of long-term interactions, which composed of interrelated concepts of efficiency, solidarity, and prosocial concern. We discuss how this framework can be used to evaluate common existing approaches in cognitive architectures for robots and then examine how social norms and mental simulation may contribute to each of the components of the framework.

Like many other autonomous systems, social robots promise to occupy an increasing range of social setting to fulfill an ever-expanding set of work roles. Human beings will be interacting with and working with them on a more regular basis. In terms of the effects of that increased presence, however, it is clear that evaluating robots as autonomous systems in such interaction contexts is not as straightforward as measuring the success of a single task or the success of tasks not involving any human-robot interactions (HRIs). Only relatively recently in the Google car's course of testing have emerged serious ethical questions about its decisions in dangerous circumstances (where to go and whom to harm in a complex crash scenario). In contrast to the *DARPA Robot Challenge*, which makes the task goals clear and efficient completion paramount, social robotics will typically have to track multiple concurrent objectives and values in order to integrate robots into society.

There are two facets of human-robot sociality and collaboration that we find deserve more explicit consideration for evaluating collaborative work. First, the working relationship between human and robot partners is itself an object of evaluation and cultivation, devoted to a particular task but judged as a team for completing tasks. As a dynamic and interactive entity, this relationship may have interests in conflict with short-term efficiencies and task completion. Secondly, the human-robot team may find itself in a socially interactive environment, where interactions ripple outward beyond the team and immediate task, engaging in various ways with the rest of society. Evaluating a decision-making architecture for robots as autonomous systems, then, must incorporate more than productivity on a primary short-term task– it may well need to gauge different scales of social benefit that the team's work can sustain in its environments.

We propose a framework to guide in the evaluation and design of robots working in a heterogenous team of humans and robots, in an environment with various levels of social interaction. We integrate two considerations – 1) working relationships developed over extended periods of time, and the personal interest of the human agent that long-term bonds may depend on recognizing and 2) the wider ethical lens of how human-robot teamwork contributes to social goods – into an overall framework of evaluating human-robot teamwork, which we call *Relational Enhancement*. That framework, we submit, has three interacting components, which we designate *efficiency*, *solidarity*, and *prosocial concern*. We use a concrete scenario of a joint task to consider two computational approaches for generating *Relational Enhancement*: social norms and mental simulation. Each approach, we demonstrate, has strengths and weaknesses in how it orchestrates efficiency, solidarity, and prosocial concern, suggesting a hybrid approach in design will ultimately be necessary.

Background and Motivation

Social working relationships between humans and robots (or computers) is a multi-faceted research area. For some, this topic involves the machine's ability to engage a human user through life-like means of representations – such as face, voice, and affect – in order to accompany, encourage, and plan for the human partner/client (Bickmore and Picard 2005; Salem et al. 2011). This research has empirically examined topics such as mood and overall stated willingness to work or play with an artificial companion.

As for HRI studies of teamwork, a good deal of scholarship has moved toward recognizing different themes of teamwork between humans and robots (Atkinson, Clancey, and Clark 2014). Scholars have attempted to integrate critical notions like "trust" with detailed mapping of how hu-

man and robot agents should exchange information and execute plans (Bradshaw et al. 2009; 2012; Woods et al. 2004). There have been several decision-making architectures proposed that aim to provide anticipatory capability (increasing "fluency" of collaboration), predictability (to help with efficiency and trust), and a "common ground" of knowledge (that facilitates more coordinated planning and execution) (Alami et al. 2005; Bradshaw et al. 2009; 2012; Shah and Breazeal 2010; Shah et al. 2011). The idea that not just tasks, but the human-robot relationships that sustainably executes tasks, matter, has undergone some empirically testing through various human-robot tasks (Alili et al. 2009; Alili, Alami, and Montreuil 2009). It has been found that the better the robot can communicate its own decision-making process, the better "fluency" might result (Hoffman and Breazeal 2004; Johnson et al. 2008). Advances of this work have increasingly suggested that the human-robot working relationship itself is more than a secondary function of the primary purpose of their work– it can be understood as increasing in usefulness and range in and of itself as it develops (Johnson et al. 2008; St Clair and Mataric 2015).

Our aim in this paper is to present a template for evaluating the performance of human-robot teams while accounting for interests and effects of various social spheres (team, organization, community). This framework, by illustrating social interests distinct from a task's explicit purpose, better enables a human-robot working relationship to be assessed as a long-term, dynamic process. Most computational theory and testing thus far has helpfully acknowledged the practical necessity of exchanging information throughout a complex task, but communication and trust in a human-robot team could mature and bear full fruit over a range of tasks and actions (Johnson et al. 2008). Human agents whose actions are not acknowledged, or are interrupted, or are disregarded for no reason, could stop engaging in the environment, quit learning important information and circulating it with robotic agents with whom they work. The working relationship between human and robot worker may form a higher-level priority depending on how a task fits within a larger mission of coordination. By emphasizing personal considerations of the human agent and the importance of prosocial behavior, however, we do not mean to neglect the signal importance of the task itself being productive, efficient, and effective. The success of the work in achieving a goal is itself a means toward improving a working relationship, and offers chances to serve the wider world besides.

Relational Enhancement Framework

To provide a more integrated framework for evaluating human-robot teamwork, we want terms that bring together the purpose and importance of the immediate task, the team relationships that sustain the performance at that work, and the larger societal sense of what that work accomplishes and who benefits. The *Relational Enhancement* framework thus carries three interacting components, which sometimes reinforce and sometimes are in conflict with each other. *Efficiency* refers to maximizing production with the resources (time, labor, materials, agents) that are available. *Solidarity* is the shared sense within a team of what the work is, what goals it aims for, and how team members can best consider each other as it is executed. *Prosocial concern* involves attention to and action toward a larger circles of agents (community, society) affected by the team's work, including respect for social norms and elements of well-being outside the immediate task goals. Through these three components, we are not claiming an exhaustive list of criteria for what human-robot relationships should embody; to be sure, terms like "trust," "fluency," "complementary," and many others have great use. What these three components provide is a basis for tracking and evaluating how relational considerations may vary in scope, as it were spatially (narrower or wider circles of people) and temporally (relationships as developed short-term and long-term). We propose that through Relational Enhancement the performance of a human-robot across a range of social environments can receive a more nuanced and ethically robust form of evaluation.

Efficiency

Efficiency is a measure of how productively and economically a task is executed, achieving a specified goal with balancing resource consumption. Efficiency generally maximizes reward and minimizes costs, broadly construed. In the context of teamwork or socially collaborative effort, efficiency will naturally demand coordination and distribution of labor (e.g., dividing up of sub-tasks, planned movements within work-space, elements that require joint focus and effort, etc). While there can be different scales of efficiency, its measure lies in the performance on a specific set of tasks. The next two elements of *Relational Enhancement* build out from efficiency, moving to the group or team attempting to accomplish such a task, and then the overall social environment in which that team operates.

Solidarity

Solidarity is a measure of how a team works toward shared, understood goals, with teammates accounting for the interests and abilities of each other. It gauges the collective strength of those working on a task, not just task outcomes. Dynamics of solidarity may manifest themselves as a team manages its task goals with difficulties or changes that its members undergo in the course of work. Recognition that an injured team member needs special attention or encouragement, for instance, could threaten productivity, but prove much more valuable when that teammate recovers and commits to the team with even more loyalty and drive. Solidarity may temper judgments about immediate performance with consideration of a member's long-term development and potential. Collectively, a team can gain solidarity over time, as cumulative experience can yield familiarity, trust, and mutual understanding. Losing solidarity, on the other hand, may threaten longer-term efforts while not having immediate impact on a task performance. As a whole, solidarity encompasses both trust and appropriate consideration of a teammate (Atkinson, Clancey, and Clark 2014).

Solidarity is especially worthwhile to consider when evaluating teams that need to tackle multiple tasks over time, or act in real-time in ways that go outside the specific

task directives. It should encompass the considerations previous research has explored along lines of coordinating joint action, such as fluency, trust, shared goals, and transparency (Bradshaw et al. 2009; 2012; Johnson et al. 2008; Salem and Dautenhahn 2015). Trust in particular will play a large role, though accompanied by rapport muir1996trust. To reiterate, the long-term aspects of solidarity, beyond the metrics of a task iteration, are critical to keep in view (Bickmore and Picard 2005). In these contexts, the team's relationship emerges forcefully as a needed object of evaluation and improvement.

Prosocial Concern

Prosocial concern refers to the awareness a team exhibits for the social working environment. While solidarity is a team-focused social priority, prosocial concern extends beyond both the specific task and the team performing it. Teams working in a socially-charged environment – whether a repair team in a city neighborhood, or a rescue team searching for people in wreckage – will face the rules, constraints, customs, or expectations that inform and size up interactions within that larger environment. Achieving or violating social norms while interacting with people beyond the team may open up or close opportunities for the team to carry out its work in that setting, as larger societal circles react to the team's work and weigh its worth. Empathy is one form of prosocial concern that can radiate from the team to the overall social context and affords a brief illustration of how prosocial concern supplements the evaluation of human-robot work.

Consider a robot and a human working together to provide medical assistance to an injured person. The primary objective of the medical assistance pair is to stabilize the person and move her into a transport vehicle for the hospital. Upon being prepared to be moved, the person yelps in pain and tears begin to pour down her face, and family and neighbors insist that the team stop. The team will have to think about their task and their own teamwork, to be sure. But their work can succeed best, both in the instance and for future instances in this neighborhood, if the team's performance can include socially appropriate interaction with patients and those that care about them.

To get a better idea of how efficiency, solidarity, and prosocial concern, we will introduce a scenario whose social elements are useful benchmarks for computational treatments.

Simple Demonstrative Scenario

We present a simple collaborative task to demonstrate the challenges in evaluating the ethical behavior in human-robot relationships and provide a couple of computational approaches that may be employed. The task to be performed by two agents is to clear the balls from a given zone. At the outset of the task, each agent is assigned a subset of the balls such that every ball is assigned to exactly one agent. The task is successfully completed once all balls have been cleared. An additional constraint is that the task must be completed by a specified time.

This is a simple task that can be achieved by two artificial agents, two human agents, or some combination thereof. There is no requirement for optimality, but we take speedy completion of the task to be preferable. Suppose that the task is performed by two humans and the zone is a football field. There are white balls and orange balls scattered around the field. All of the white ones need to go into one bin, and the orange balls need to go into another bin. The first agent, Walter, is assigned to the white balls, and the second agent, Oscar, is assigned to the orange balls. They must complete the task of clearing all the balls before a certain time (e.g., because another team is scheduled to use the field at that time).

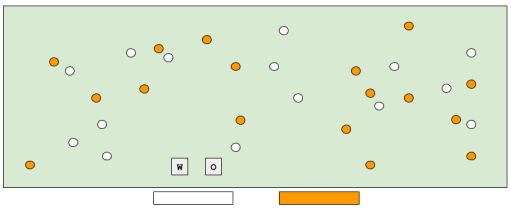

Figure 1: Orange and white balls are scattered across the football field. Walter (W) is to move all the white balls to the white bin, and Oscar (O) is to clear all the orange balls into the other bin.

Walter is quicker to collect his assigned balls. When he is done, he notices that there is an orange ball nearby, but Oscar is at the other end of the field. Should Walter provide assistance by clearing the ball? Assuming a large field with many balls, it may require a significant amount of time to complete the task. If Walter completes his task far before Oscar and does not have another pending task, then one would normally expect Walter to offer assistance (especially since there are not explicit rules prohibiting Walter to help out). In fact, some may even consider it rude for Walter to just watch Oscar clear the rest of the balls. Effectively, there seems to be an implied *social contract* that assistance should be provided in this context.

Now suppose that everything is the same as before except that Walter is a fast and efficient ball scooping robot. Should Walter assist Oscar in the same way as we would expect the human Walter to assist? I.e., does the same social contract apply if one agent is a robot? Before we quickly answer in the affirmative, consider some of the implications on *efficiency*, *solidarity*, and *prosocial concern*. In terms of efficiency, it is important that the task be completed correctly. Does Walter have the knowledge that the orange balls go in the orange bin? It has so far only put balls in the white bin. The solidarity of the team could be enhanced by Walter demonstrating it can be trusted. Would Walter correctly moving the orange balls to the orange bin improve trust? Or perhaps trust is better served by Oscar knowing that Walter will not perform any action that it has not been explicitly told to do. How does Oscar's personal dignity affect Walter's decision? Walter's assistance could devalue Oscar's contribution and cause him to feel incompetent. As for prosocial concern, the team needs to respect those outside of the team.

Should Walter assist if that is the only way to complete the task before the next team arrives? Or would Walter's assistance be construed by observers as a violation of how the community expects a robot to assist a human? We will discuss these questions and others as we review some of the computational approaches that can be used to decide which action the robot is to take.

Computational approaches

Many approaches found in cognitive architectures for robots select an action based on the utility of the action. Also, systems often are able to reason about various constraints (e.g. locations to be avoided). A significant amount of research and development has focused on creating systems that are fast and efficient. We briefly review here some approaches and discuss the potential gaps between them and the Relational Enhancement framework.

Utilitarian approaches

A utilitarian approach would naturally involve the robot making some autonomous decisions based on the expected utility of the action, weighing the expected utility against the expected cost. The optimal action to take would be the one with the maximum net utility. A common approach is to use a planning system to find the best sequence of actions

A utilitarian approach seems ideal for maximizing task efficiency as it would select the action that maximizes utility which, in turn, should be correlated with the task performance measure. In our ball clearing scenario, once the robot has been assigned its set of balls to clear and locates that set, it can invoke a planner to find the most efficient manner for it to complete its task. It is also reasonable for it predict the best or most likely execution time for the human partner in the task. Toward the end of achieving maximal efficiency, it could arrange the assignment of balls such that the collective task is successfully completed in the least amount of time.

Thus, the utilitarian approach seems suited to accommodate the *efficiency* component of our framework. It is less clear how *solidarity* or *prosocial concern* would find adequate accounting in this approach. Ostensibly, utilities would need to be assigned to the features and goals of solidarity and prosocial concern. One could offer a utility for an act that builds solidarity by taking the action the team most likely expects. An act of prosocial concern, which might strengthen ties with the team's neighbors, might also invite some attempt at a utility value. However, the complications of this approach can already be seen in explorations of trolley problem judgments. Variations of the trolley problem show that most people prefer a non-utilitarian answer to its dilemma, finding it impermissible to push one man off a footbridge to save five people about to be run over by a runaway trolley (Mikhail 2007; Thomson 1985) – a utilitarian solution focussing only on saved lives would have been to sacrifice the one man on the footbridge.

Moreover, there is evidence that people's judgements vary in the "by-stander problem" depending on whether the person to be sacrificed is a child or a family member (Bleskerechek et al. 2010). While it is often possible to find some way to assign costs and utilities to make the utility-theoretic decision accord with human decisions (e.g., as has has been demonstrated in (Wilson and Scheutz 2015) where a utilitarian approach was used to simulate such judgments by modulating utilities based on emotional empathy and prosocial concern), those assignments are often fairly arbitrary (e.g., Wilson & Scheutz did not assign systematic values for acting with emotional empathy or for helping an agent feel competent and not inadequate).

Moreover, it is not clear how social values could receive utilities relative to one another. Returning to our scenario, Walter, the robot, may decide, if it is done and Oscar is still clearing, that it is most efficient to clear any remaining balls assigned to Oscar. Oscar, though, may end up feeling slow, incompetent, or even useless as a result of this "help." A prosocial concern for the community enters consideration through the fact that all the balls must be cleared before the next teams arrive at the field for a game. We have interrelated concerns, but it is difficult to pin down how they settle in quantitative terms. Just some of the competing priorities are the following:

- Complete the ball-clearing task in the most efficient fashion
- Avoid impinging upon the other teams, or those who have come to watch them
- Maintain personal dignity of one's own team members
- Maintain a positive mood amongst one's own team members

For the most part, these options are independent, and it is unlikely that utilities can be defined for these options such that the utilities are valid and appropriate across all scenarios. In many cases, it may be of greatest utility to complete the task quickly. However, in other cases the personal dignity of the human team members may be more important than the most efficient clearing. Alternatively the relations with those outside the team may demand promptness of clearing, not face-saving patience for the sake of the team's rapport.

One wrinkle in the utilitarian approach could be encoding some rules or constraints to handle some of these complications in expectation and performance. For example, it may maximize a measure of utility for a robot to push a person in front of a runaway trolley to save five people who otherwise would be be killed, but a constraint against physically pushing or moving a human could forbid such an action. One could encode information about solidarity and prosocial concern as constraints that the robot must check before completing its action selection. A similar approach has been used to generate ethical behavior in a robotic architecture (Arkin, Ulam, and Wagner 2012). The robot checks the ethical rules that it was given before selecting an action, and if all actions violate some rule then the action with the greatest utility, and the one that violates the fewest constraints, is preferred. While this approach does improve upon a strict utilitarian decision, it fails to capture the long-term effects of actions, whereby solidarity and prosocial concern may represent more important aspects of the actions. Additionally, such approaches, in order to resolve conflicts – how to act

if two actions violate a constraint or an ethical code – still ultimately must rely on a utility value being assigned to each alternative (see (Scheutz 2014)).

We next consider two other computational approaches to reasoning about tasks in a way that addresses Relational Enhancement – social norms and mental modeling. From the outset these approaches are considered as complements of one another and the utilitarian approach, not as mutually exclusive alternatives. We have not yet implemented the following approaches, but explore them here to further the discussion on the development of ethical robots.

Social norms

A social norm is a common pattern of behavior identified within a social environment. It provides guidance for how one should act and allows other to predict how others are likely to act. Humans participate in a variety of social norms on a constant basis, from greeting one another to respecting elders. Many of these norms have simple benefits, like the courtesy of holding a door open for another person. There may be no benefit to oneself, and the beneficiary of the action may be a complete stranger and may never be seen again. In regards to our framework, social norms may not necessarily include an element of efficiency – at least in the short term. On a longer-term view, one might well view social norms as implicitly encapsulating a variety of long-term benefits around cooperative, coordinated action. When one considers even nonverbal behaviors that may serve as a greeting, a show of encouragement, or an expression of empathy, it is not surprising that many artificial agents are designed to express empathy (often through mimicry) to improve the relationship between the human and the agent (Bickmore and Picard 2005; McQuiggan and Lester 2007; Boukricha and Wachsmuth 2011). Task planners have also broached these kinds of social rules that a robot agent might honor, which suggests the possibility of social norms as means of guidance (Alili et al. 2009; Alili, Alami, and Montreuil 2009; Shah and Breazeal 2010; Shah et al. 2011)

In our ball clearing scenario, it may be desirable and have an immediate benefit for the robot to provide assistance, but future tasks may be affected. For example, perhaps the robot can be assigned to a zone where all the red balls are to be cleared but none of the yellow balls. This robot that has developed trust in choosing not to clear balls it is not assigned (at least not without permission) can be assigned to this more complex task. Extending this task into more realistic domains, a trusted robot, one that has demonstrated through repeated trials that it will consistently and reliably perform as desired, can be assigned to a dangerous task that requires collecting the red balls (food) but avoiding at all costs all of the yellow balls (which turn into a poisonous gas upon touch).

We propose an approach of incorporating social norms as guidance for which action is most appropriate for an agent to take. When multiple actions may be applicable in a given scenario, if one action corresponds with a social norm then this action may be preferred. We present three social norms that may be considered in our ball clearing scenario, how the reasoning could be used in a computational system, and how using these norms relate to Relational Enhancement.

Table 1: The set of conditions for the social norm of requesting permission. Q_R is the queue of pending actions for the robot R. $Req(\alpha)$ is the set of requirements to do action α, and $Cap(R)$ is the set of capabilities of R. **O** is the modal operator indicating that the proposition is obligatory, and **D** indicates that the proposition is desired.

Description	Condition
Robot has no other immediate task	$Q_R = \emptyset$
Robot is capable of doing other task	$Req(\alpha) - Cap(R) \neq \emptyset$
No other agent is doing the task	$\forall x \neg doing(x, \alpha)$
There is no expectation for robot to assist (not necessarily obligated)	$\neg \mathbf{O} do(R, \alpha)$
Increasing trust human has in robot is desirable	$\mathbf{D} increase(trust(H, R))$

Requesting permission The first social norm suggests that if an agent is available to help, it should first request permission to help. Consider the case where the robot has completed its task (clearing its assigned balls) and identifies that another ball is nearby and can easily be cleared. The robot asks if it should clear the ball, and upon receiving permission it clears the ball.

A cognitive architecture using norms to select an action could choose to apply a social norm and the corresponding action if the current situation matches a set of conditions. This approach would be a form of case-based reasoning, in which an action or script is selected based on some matching conditions or trigger conditions being met (Riesbeck and Schank 1989; Ros et al. 2009). Some possible conditions, expressed in prose and logically, for this norm are in Table 1.

Asking for permission to do a task that is readily available to be done and would only be delayed by taking a moment to request permission does not have much immediate benefit. This would diminish the efficiency by which the task at hand is being accomplished. However, the request – regardless of whether it is granted or not – has potential long-term benefits, such as the following:

- build trust and respect
- express concern for the interests and desires of the other agent
- allow the agents to arrive at a common understanding

We categorize these benefits in the Relational Enhancement framework under *solidarity* because they are focused on developing shared perspective and build stronger relationships within the team. We next describe each of these benefits further and how they enhance human-robot interaction in the long term.

Not acting out of impulse or as an automatic reaction to the available ball and instead showing some constraint could

install in the human a sense of confidence that the robot is likely to show constraint and ask for permission in other cases as well. This may directly lead to the human trusting the robot to not act in cases where it should not (such as when it does not have permission). We could imagine a scenario in which it is critical that the robot does not perform some task even though it may be convenient for it to do so. A trusted robot can be positioned near this task to do another task and the human can have confidence that the robot will not stray from its task without permission.

By requesting permission, the robot allows the human to respond and express the interests or desires that she may have. An obvious benefit of this is that the robot gains this knowledge. A less tangible benefit is that the human is likely to increasingly feel that her desires matter in the team and that her interests are taken into account in decisions. The benefits of this range from the human being more accepting of a robot that considers her interests to the human being more engaged in a team that respects her own desires.

Requesting permission allows the human to not only respond with the permission (or the denial) but also additional information. Perhaps Oscar wants to get some extra exercise, or maybe he is not sure that the robot understands the careful handling necessary for the orange balls. By choosing to communicate before acting, the robot creates an opportunity for further communication and brings the human and the robot closer to a common understanding.

As a result of these, there may be long-term benefits to efficiency. Improvements in efficiency may also occur as the result of the robot requesting permission to assist. It is also an opportunity to establish an evolving protocol that may incorporate work history or a scope of rules. If Oscar does wish to grant permission to Walter to assist and wants it to regularly assist with clearing all the balls, he may reply with a statement like, "Yes, please help anytime. Use the same procedure as clearing the white balls, but put the orange ones in the other bin." Or perhaps Oscar needs to provide some constraints: "Yes, but pick up the orange balls more slowly so as to not mark or scuff them."

Subordinate assistance The second social norm we present applies when one agent is a subordinate to the other agent, and the task presents little safety risk for the subordinate agent who might provide assistance. In this case, it is assumed that the subordinate agent will automatically provide assistance and that first requesting permission is not necessary.

Some of the conditions for this norm are similar to the first norm, but we also have conditions specifying the roles of the agents and the risk associated with the subordinate agent performing an action. Table 2 describes the conditions associated with this norm.

The application of this norm supports the efficiency of completing the task, but there are also considerations for solidarity. There is a recognition on the part of the robot that it is a subordinate and that being a subordinate has certain implied responsibilities. The agents share a common goal of accomplishing the ball clearing task, but the robot providing automatic assistance allows the task to be completed faster, whereby freeing up the human partner to pursue other goals. Additionally, it may be beneficial for the human to gain an expectation that the robot can and will assist in tasks even without explicitly being told to do so under the appropriate conditions. If robot continues to do so and does the task correctly, the human-robot team may improve task performance and take on other tasks where subordinate assistance is ideal. Though subordinate assistance directly applies to the team, prosocial concern can also feature in this norm. If the pair are observed by those around the field, Oscar and Walter's work could have broader effects. Any appearance of demeaning or abusive behavior by Oscar, even as a joke, may be troubling to see in combination with subordinate assistance. Following a norm of subordinate assistance therefore is not just a means toward task completion or team-building – it should also be compatible with good relationships between the team and its community.

Table 2: The set of conditions for the social norm of subordinate automatically assisting. Q_R is the queue of pending actions for the robot R. $Req(\alpha)$ is the set of requirements to do action α, and $Cap(R)$ is the set of capabilities of R. **E**[X] is the expected value of X, and θ is a threshold for the maximum acceptable cost of failure.

Description	Condition
Robot has no other immediate task	$Q_R = \emptyset$
Robot is capable of doing other task	$Req(\alpha) - Cap(R) \neq \emptyset$
No other agent is doing the task	$\forall x \neg doing(x, \alpha)$
Task is trivial and non-risky for robot	$\mathbf{E}[failed(x)] < \theta$
Robot is subordinate to the human	$subordinate(R, H)$

Respect for others The final social norm we consider focuses on a team being scheduled to use the field at a specified time, with their game being impossible if all the balls are not cleared before that time. The norm in this case would be a group-oriented one, where the work of the human-robot pair should respect another team's ability to use the field. One might argue that being on time is a simple rule that is followed (e.g., end by 7pm). But one could easily imagine a looseness around times, with groups finishing over the time if the next team was late. The norm of respecting a team's use of the field would govern that fluidity – the key would be not holding a team up from their rightful use of the field.

The conditions for this norm are similar to the first norm in that the robot is available to assist in the task. The conditions specific to this norm describe the social context that another agent (or team) has reserved a resource and that the robot is currently using that resource. Finally, the normative behavior is to take action if the action is likely to make the resource available in time. Table 3 describes the conditions associated with this norm.

This norm is intended to demonstrate how constraints from and consequences upon the social environment can in-

Table 3: The set of conditions for the social norm to respect the reserved resources of another agent. Q_R is the queue of pending actions for the robot R. $Req(x)$ is the set of requirements to do action α or task t or to complete the reservation ρ. The resource σ is required by the task t and the reservation ρ. The action α is an action towards completing the task t. The time at which the task t will complete is $completionTime(t)$, and start time of reservation ρ is $startTime(\rho)$.

Description	Condition
Robot has no other immediate action	$Q_R = \emptyset$
Robot is capable of doing other action	$Req(\alpha) - Cap(R) \neq \emptyset$
No other agent is doing the action	$\forall x \neg doing(x, \alpha)$
Current task uses a resource	$\sigma \in Req(t)$
Current task will not complete in time	$completionTime(t) > startTime(\rho)$
Next agent has reservation	$has(other, \rho)$
Reservation requires the resource	$\sigma \in Req(\rho)$
Other action will complete in time	$completionTime(\alpha) > startTime(\rho)$

fluence the decision as to which action the robot should take in tandem with its human partner. If the ball clearing task is not completed in time and some balls still remain on the field, then the agents (and their team) are impinging upon the ability of the other team to use the reserved resource. The other team has the right to be upset, as there has been a norm violation in respecting the other team and the reservations that have made. In order to avoid any conflicts with the other team, the robot is to take an action that will complete the task in time.

As mentioned, there is some flexibility to this norm. Perhaps the task does not need to be completed before the reservation time but before the other time arrives if they arrive after the reservation time. In that case the normative behavior might be for the robot to do nothing (provide no assistance) until the next team begins to arrive, and then the robot assists so that the task can be completed quickly and the field is made available for the next team.

Mental simulation with counterfactual reasoning

We now discuss an approach that allows for in-depth reasoning about a wide range of possible outcomes. Mental simulation with counterfactual reasoning is well suited for explicitly reasoning about specific effects or particular variables. An advantage of this feature is that effects of actions relevant to each of the categories of Relational Enhancement may be directly considered and evaluated.

Mental simulation has been applied to many domains, including qualitative simulations of physical systems (Forbus 1984), simulations of teammates decision-making (Kennedy et al. 2008), and making moral decisions (Wallach, Franklin, and Allen 2010). Looking at human-robot teamwork through theory of mind and mental modeling has already yielded some helpful proposals (Nikolaidis and Shah 2012; Hiatt, Harrison, and Trafton 2011). A recent application of mental simulations to moral decisions simulated the effects of prosocial concern (Wilson and Scheutz 2015). Simulations of physical systems, teammates decisions, and moral decisions can be viewed as examples of how mental simulation can be used to represent and reason about details relevant to efficiency, solidarity, and prosocial concern, respectively. In this section we will discuss further how mental simulation can address each of these categories.

An important application of mental simulation is reasoning about the long-term implications of actions by simulating a series of actions. In particular, the long-term effects of a single action that is performed numerous times can be determined by repeatedly simulating the action. Each iteration of the simulation uses information from the end state of the previous simulation for the initial state of the subsequent simulation. For example, consider the scenario where the robot, Walter, has completed clearing its balls and the chosen action is to do nothing – not assist Oscar in clearing the rest of the balls. At the end of a single simulation of this action, Oscar has an increased expectation that Walter will respect the division of labor that has been assigned, not assist Oscar, and thus not attempt to clear Oscar's balls. Repeating this simulation, and now starting with this increased expectation to not assist, the result is similar – further increase of expectation. Since repeatedly performing as expected is likely to result in trust (Corritore, Kracher, and Wiedenbeck 2003; Muir and Moray 1996), we reason that after some number of repeated events of the robot not assisting, the human acquires trust that the robot will not do more than it has been instructed. With this trust, Oscar knows it may leave Walter unattended to complete its portion of the task and not overstep its bounds (though overall solidarity may include what effects not assisting will have on Oscar's overall attitude toward Walter as co-worker, how respected as a worker Oscar feels, etc.) As an analogy, think of a dog so well-trained and the trainer so confident, that it may leave a freshly cooked steak on the table, within reach of the dog, without any fear that the dog will attempt to eat the steak.

Multi-trajectory simulation To discuss more deeply the mechanics of the mental simulation, we consider the scenario in which Walter has completed its portion of the task and it is now deliberating upon three possible actions: 1) do nothing, 2) automatically assist, and 3) request permission to assist. If permission is granted, Walter will assist. If permission is denied, it will do nothing. And if no response to the request is given, then Walter will repeat the request. The complete set of possibilities is depicted in the graph in Figure 2. Note that branches in the graph are the result of the robot having a choice in action to take and multiple possible results of one of the robot's actions. This multi-trajectory simulation allows us to explore a wide range of possibilities.

The simulation of this scenario has six possible end states, each describing the short-term effects of the actions taken. As discussed above, one short-term effect is the increased expectation Oscar has that the robot will take that particular

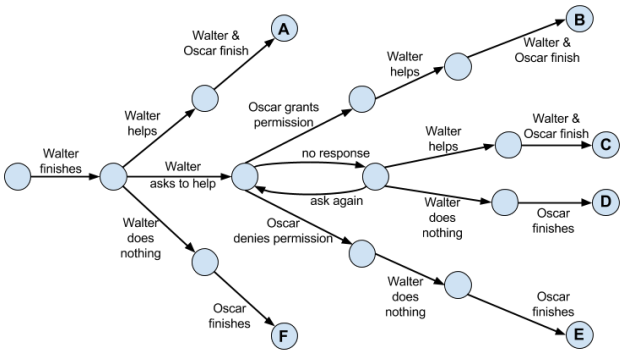

Figure 2: Multi-trajectory simulation of completing the task when Walter has finished its portion first. The simulation represents that Walter may automatically provide assistance, request permission to assist, or do nothing. When requesting permission, it may be granted, denied, or no response provided.

action. We can also analyze end state A to determine that it is optimal because it completes the task in the least amount of time. Repeated simulations of this scenario will present benefits to efficiency and solidarity.

Efficiency The optimal outcome is in end state A, but states B and C are also relatively optimal. The differences in optimality between end state A and end state F may be minimal in a single simulation, but repeated simulations produce a wider gap between these outcomes. In terms of efficiency, it can be concluded that always taking the action that results in end state A is the best option.

Solidarity While always performing the action that results in end state A is the most efficient, we discuss here how it is not necessarily the best option for solidarity. In introducing mental simulation, we describe a scenario in which the robot repeatedly does not assist – resulting in end state F each time. Predictable behavior, such as always not assisting, enables the human to build trust in what the robot will do. That being said, any consistent behavior will lead to predictable behavior and the resulting trust. Contrast this with sometimes helping after being granted permission (end state B) and sometimes automatically assisting (end state A). Consider a sequence of simulations with end states A, B, A, B, etc. Compare this with a sequence of simulations that always results in end state A. The differences in efficiency between these simulations is minimal, but the difference in trust, and the type of trust involved, is more significant.

Another sequence of simulation that results in trust is when the robot asks for permission and then assists upon being granted permission. The implied result is that the human can trust that the robot will always ask for permission in the future. Asking for permission has an additional benefit in terms of solidarity. By inquiring what the human wants the robot to do, there is an opportunity for the human to express information beyond the granting or denying of permission. Perhaps the robot is not to help because some of the balls need to be specially handled or maybe the human simply wants the extra exercise. This additional information gives the human-robot team an increased shared perspective. There is then the potential for the robot to make a more informed decision in the future, a decision that is more representative of the shared perspective and common goals of the team.

Multi-agent simulation The team of Walter and Oscar is working in some social environment, and that environment may have other agents with which team will interact or may influence the behavior and choices of the team. The environment includes the local government that provides certain laws and ordinances that must be obeyed. The social environment includes other agents, such as other teams, local residents, or league officials. The team may have relationships with many of these agents, and these relationships need to be fostered or maintained. For example, it is prudent to not disturb the local residents with excessive noise, especially during normal sleeping hours. Failing to do so can cause neighbors to be upset, damaging the relationship with them, and hurting chances or future cooperation with them.

In our ball clearing scenario, we consider the case where another team is scheduled to use the field soon. We describe here how a multi-trajectory, multi-agent simulation can be used to identify a wide range of possible outcomes. One approach is to conduct a multi-trajectory simulation of each agent, identify the interaction points, and then simulate the effects of the interactions (Hinrichs et al. 2011). We describe here a slightly different approach in which we simulate the team as we did in the previous example, but we interject exogenous events into the simulation at relevant points and simulate the effects of these events in combination with the actions of the team.

Consider the case in which the robot, Walter, is deciding whether it should help Oscar in completing the task. The interaction with the next team scheduled to use the team depends on whether the team arrives on time or late. (We make the assumption that arriving early is equivalent to on-time but recognize this is not always equivalent.) Walter has two possible actions and corresponding outcomes:

1. Walter helps $\Rightarrow$ teams finishes on time
2. Walter does nothing $\Rightarrow$ team finishes late

We have two possible events introduced by the other team:

A. Other team arrives on-time

B. Other team arrives late

Combination of these event results in the interactions described in Table 4.

We see here that external influences from the other team affect the outcomes, affect team dynamics, affect relationships with the social environment, and ultimately can afffect the action choices of the robot. In scenario 1A, the relationship between Walter and Oscar can improve because they had to work together to not only complete the task but also to avoid negative interactions with the other team. Other outcomes resulted in pride, guilt, or regret – all impacting the

Table 4: The effects of the combinations of events. Label is the combination of events. Effects is a description of the impacts on Team A (Walter and Oscar) and Team B (the next team that has reserved the field).

Label	Event Sequence	Effects
1A	Walter helps Team A completes task on-time Team B arrives on-time	Team B appreciates field is ready for them. Oscar is glad the team was able to get the task done in time. Oscar is a little saddened by the fact that he could not complete his portion of the task on his own.
1B	Walter helps Team A completes task on-time Team B arrives late	Team B is ether appreciative or has no reaction towards the team. Oscar thinks the help was unnecessarry and regrets receiving help.
2A	Walter does nothing Team B arrives on-time Team A completes task	Team B is annoyed the field is not ready. Oscar is embarrassed and feels guilty.
2Bi	Walter does nothing Team A completes task late Team B arrives late	Team B has not reaction towards team A. Oscar has pride in being able to complete his portion of the task. Oscar appreciates Walter letting him finish.
2Bii	Walter does nothing Team B arrives late Team A completes task late	Team B is greatly annoyed that field is not ready even with it being late. Oscar regrets not getting helps and feels guilty for imposing upon the other team.

team dynamics. In addition to effects on the team, we see that the other team may have a positive or negative reaction to the Walter and Oscar. Negative responses could have a variety of consequences, including complaints to those managing the field to direct conflict between the teams.

There are numerous possibilities for how these agents could interact, and the multi-trajectory mental simulation is a means of fully describing the wide range of actions and effects that may occur. In some cases, this elaborate description can make an action choice obvious. For example, if all actions but one result in a negative outcome, then the one positive action is probably the best choice. In cases like we have presented here, where each action in combination with exogenous events results in some combination of positive and negative outcomes, some additional mechanism is needed to make the final decision. The goal of the mental simulation is to provide a sufficient level of information for the decision mechanism to make the best possible choice. This is an example where the social norms can be applied after having done the mental simulation. If the robot understands that it is an appropriate normative response to avoid irritating the other team, then the action available to the robot that does not result in this negative outcome is for the robot to act. However, if the situation dictates that the appropriate norm is to ensure that the human teammates morale and mood stay positive, then allowing the human to complete the task has the chance for greatest reward.

In summary, the combinations of events and event sequences can be quite vast, and mental simulation attempts to sort through many of the possibilities to reach a more informed decision. The simulation provides a mechanism by which we can explore information relevant to all aspects of Relational Enhancement by explicitly modeling and simulating the effects on efficiency, solidarity, and prosocial concern. We have given some examples where the action resulting in the greatest efficiency does not necessarily benefit solidarity or prosocial concern. By repeating the simulations we can attempt to explore the long-term effects of these actions, hopefully leading to the decision that is best short and long-term and does not neglect any of the categories of Relational Enhancement.

While mental simulation theoretically has the capacity to do an exhaustive search of the actions, subsequent actions, and all those actions' associated effects, it is computationally infeasible due to (1) the large state space and (2) the lack of absolute end conditions (i.e. It is not like chess with well-defined end states). As a result, combining the mental simulation with other approaches, like social norms and utilitarian approaches, will be necessary.

Discussion

Though the scenario we present here is relatively simple and theoretical compared to others in the literature, it yields some nuanced projections of how social norms and mental modeling may best perform. Its basic structure can apply to many socially involved cleaning tasks, from a dinner party to a playground to a hazardous material cleanup, where the completion of the task may be complicated by attending to matters of dignity, propriety, trust, and long-term collaboration with the people involved. A social norm approach seems better geared toward work relationships where a violation of respect or dignity is particularly damaging, as well as where etiquette and protocol are closely tied to the nature of the work practice (e.g. traditions of a vocation). One disadvantage of that approach, however, might be that it condenses and obscures implicit assumptions about relational virtues and longer-term goods. If the demands of relational maintenance impede critical achievements of safety or health, for example in a time-critical cleanup of a toxic spill, a context-based reexamination seems important for the team to employ. A mental modeling approach might be more nimble on that score, with more dynamic apprehension of short and long term relationship implications. The downside of that analytic power may be the cost to processing efficiency, with both time and energy of the system being taxed in real time scenarios.

There are counter-examples that could be generated, admittedly, where a social norm approach might handle social roles and violations less ably than mental modeling. Likewise, mental modeling's mapping of intention and future actions may be less adaptive and conducive to trust than basic adherence to norms. What does seem evident at this initial stage of exploration, however, is that both approaches are

worth keeping in mind for design and management theory around human-robot teams. Given that social robots are entering work environments of many types, it will only become more and more difficult to dissociate ethics from the design and policy around human-robot relationships. In this sense a hybrid approach might link up with work that links ethical, collectively-oriented principles to safety constraints for autonomous systems (Rossi 2015). There are many analogies to our demonstrative scenario to be found across many real-life, not to mention riskier, contexts. The need to consider people's interests, from basic dignity to their purposes in life and work, will become increasingly evident and crucial, just as will the place of human-robot teamwork within society's views of its betterment into the future. How will a robot repair worker with a human colleague deal with a toddler coming into the road? Should it go off task? Should it let the person decide if that life is worth saving? How will robot healthcare assistants work with medical providers to bring the best therapy to a patient? How will intersecting notions of patient dignity and medical vocation factor into that action? How will search and rescue robots collaborate to find evidence that human colleagues may find especially fraught, from weapon pieces to human remains?

For these and many other tasks, it will not be enough to separate personal or societal considerations as ethical addenda to the "real work" or task. What we have designated solidarity and prosocial concern will need to find operational expression, just as they will need to sit within an overall reasoning process that can facilitate morally acceptable reasons and morally competent action. Relational Enhancement as a term represents the ethically peremptory integration of *efficiency*, *solidarity*, and *prosocial concern* – none of the three can be alienated from evaluation, and each of the three can inform and affect the other two.

Conclusion

Robots will soon take on an even broader range of roles and functions than they do already within society. The various kinds of work they will perform will include socially robust collaboration with human beings, both as partners and as a surrounding social environment.

We have shown in this paper that evaluating socially interactive robotic performance with human partners requires not just short-term measures of efficiency but also long-term measures of the team's relationship and the interaction of its work within a larger social milieu. Interactions within and outside the team bear on more than the efficiency with which a present task is performed, so we proposed solidarity and prosocial concern as two critical supports for a socially effective and collaborative human-robot partnership . Through a basic task scenario, we outlined the advantages for a human-robot team of the robot anticipating, addressing, and accommodating the needs and interests of its human teammate. We then showed how the long-term work of a human-robot team could hinge on how well it is understood, received, and possibly assisted by those physically near to the work, or by those whose lives the work affects. We tackled the necessary means for moral reasoning that a computational approach to efficiency, solidarity, and prosocial concern would involve, and we concluded that a hybrid of utility and social norms is the most promising means for including all three elements of relational enhancement. Computational architectures that furnish this more socially adaptive approach to human-robot teamwork will render robots more successful and accountable partners in the many social environments they are beginning to occupy.

Acknowledgements

This project was in part supported in part by a grant from NSF, #IIS-1316809.

References

Alami, R.; Clodic, A.; Montreuil, V.; Sisbot, E. A.; and Chatila, R. 2005. Task planning for human-robot interaction. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, 81–85. ACM.

Alili, S.; Alami, R.; and Montreuil, V. 2009. A task planner for an autonomous social robot. In *Distributed Autonomous Robotic Systems 8*. Springer. 335–344.

Alili, S.; Warnier, M.; Ali, M.; and Alami, R. 2009. Planning and plan-execution for human-robot cooperative task achievement. *Proc. of the 19th ICAPS* 1–6.

Arkin, R. C.; Ulam, P.; and Wagner, A. R. 2012. Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE* 100(3):571–589.

Atkinson, D. J.; Clancey, W. J.; and Clark, M. H. 2014. Shared awareness, autonomy and trust in human-robot teamwork. In *Artificial Intelligence and Human-Computer Interaction: Papers from the 2014 AAAI Spring Symposium on*.

Bickmore, T. W., and Picard, R. W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12(2):293–327.

Bleske-rechek, A.; Nelson, L. A.; Baker, J. P.; and Brandt, S. J. 2010. Evolution and the Trolley Problem : People Save Five Over One Unless the One is Young, Genetically Related, or a Romantic Partner. *Journal of Social, Evolutionary, and Cultural Psychology* 4(3):115–127.

Boukricha, H., and Wachsmuth, I. 2011. Empathy-based emotional alignment for a virtual human: A three-step approach. *KI-Künstliche Intelligenz* 25(3):195–204.

Bradshaw, J. M.; Feltovich, P.; Johnson, M.; Breedy, M.; Bunch, L.; Eskridge, T.; Jung, H.; Lott, J.; Uszok, A.; and van Diggelen, J. 2009. From tools to teammates: Joint activity in human-agent-robot teams. In *Human Centered Design*. Springer. 935–944.

Bradshaw, J. M.; Dignum, V.; Jonker, C.; and Sierhuis, M. 2012. Human-agent-robot teamwork. *Intelligent Systems, IEEE* 27(2):8–13.

Corritore, C. L.; Kracher, B.; and Wiedenbeck, S. 2003. Online trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58(6):737–758.

Forbus, K. D. 1984. Qualitative process theory. *Artificial intelligence* 24(1):85–168.

Hiatt, L. M.; Harrison, A. M.; and Trafton, J. G. 2011. Accommodating human variability in human-robot teams through theory of mind. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, 2066.

Hinrichs, T. R.; Forbus, K. D.; de Kleer, J.; Yoon, S.; Jones, E.; Hyland, R.; and Wilson, J. 2011. Hybrid Qualitative Simulation of Military Operations. In *IAAI*.

Hoffman, G., and Breazeal, C. 2004. Collaboration in human-robot teams. In *Proc. of the AIAA 1st Intelligent Systems Technical Conference, Chicago, IL, USA*.

Johnson, M.; Feltovich, P. J.; Bradshaw, J. M.; and Bunch, L. 2008. Human-robot coordination through dynamic regulation. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 2159–2164. IEEE.

Kennedy, W. G.; Bugajska, M. D.; Adams, W.; Schultz, A. C.; and Trafton, J. G. 2008. Incorporating Mental Simulation for a More Effective Robotic Teammate. In *AAAI*, 1300–1305.

McQuiggan, S. W., and Lester, J. C. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human Computer Studies* 65(4):348–360.

Mikhail, J. 2007. Universal moral grammar: theory, evidence and the future. *Trends in cognitive sciences* 11(4):143–52.

Muir, B. M., and Moray, N. 1996. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39(3):429–460.

Nikolaidis, S., and Shah, J. 2012. Human-robot teaming using shared mental models. *ACM/IEEE HRI*.

Riesbeck, C. K., and Schank, R. C. 1989. *Inside Case-Based Reasoning*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc.

Ros, R.; Arcos, J. L.; De Mantaras, R. L.; and Veloso, M. 2009. A case-based approach for coordinated action selection in robot soccer. *Artificial Intelligence* 173(9):1014–1039.

Rossi, F. 2015. Safety constraints and ethical principles in collective decision making systems. In *KI 2015: Advances in Artificial Intelligence*. Springer. 3–15.

Salem, M., and Dautenhahn, K. 2015. Evaluating trust and safety in hri: Practical issues and ethical challenges. In *Workshop on the Emerging Policy and Ethics of Human-Robot Interaction @ HRI 2015*.

Salem, M.; Rohlfing, K.; Kopp, S.; and Joublin, F. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *RO-MAN, 2011 IEEE*, 247–252. IEEE.

Scheutz, M. 2014. The need for moral competency in autonomous agent architectures. In Müller, V. C., ed., *Fundamental Issues of Artificial Intelligence*. Berlin: Springer.

Shah, J., and Breazeal, C. 2010. An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52(2):234–245.

Shah, J.; Wiken, J.; Williams, B.; and Breazeal, C. 2011. Improved human-robot team performance using chaski, a human-inspired plan execution system. In *Proceedings of the 6th international conference on Human-robot interaction*, 29–36. ACM.

St Clair, A., and Mataric, M. 2015. How robot verbal feedback can improve team performance in human-robot task collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 213–220. ACM.

Thomson, J. J. 1985. The trolley problem. *The Yale Law Journal* 94(6):pp. 1395–1415.

Wallach, W.; Franklin, S.; and Allen, C. 2010. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in cognitive science* 2(3):454–85.

Wilson, J. R., and Scheutz, M. 2015. A model of empathy to shape trolley problem moral judgements. In *The sixth International Conference on Affective Computing and Intelligent Interaction*. IEEE.

Woods, D. D.; Tittle, J.; Feil, M.; and Roesler, A. 2004. Envisioning human-robot coordination in future operations. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 34(2):210–218.