Reflections on the design challenges prompted by affect-aware socially assistive robots

Jason R. Wilson, Matthias Scheutz and Gordon Briggs

Abstract The rising interest in socially assistive robotics is, at least in part, stemmed by the aging population around the world. A lot of research and interest has gone into insuring the safety of these robots. However, little has been done to consider the necessary role of emotion in these robots and the potential ethical implications of having affect-aware socially assistive robots. In this chapter we address some of the considerations that need to be taken into account in the research and development of robots assisting a vulnerable population. We use two fictional scenarios involving a robot assisting a person with Parkinson's Disease to discuss five ethical issues relevant to affect-aware socially assistive robots.

1 Introduction and Motivation

Demographic trends in a variety of developing nations [31] as well as aging populations in the Japan and the West [28] are at least in part responsible for a growing interest in developing artificial helper agents that can assume some of the responsibilities and workload of increasingly in-demand human caregivers, and has given rise to the field of *assistive robotics* [5]. It is likely given the rapid technological advances in robotic technology and artificial intelligence, that these assistive robots will sooner rather than later enter households around the globe, where they

Jason R. Wilson
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave., Medford, MA 02155 e-mail: wilson@cs.tufts.edu

Gordon Briggs
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave., Medford, MA 02155 e-mail: gordon.briggs@tufts.edu

Matthias Scheutz
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave., Medford, MA 02155 e-mail: matthias.scheutz@tufts.edu

are poised to deliver various services to their owners. However, in addition to benefitting humans, these artificial agents also have the potential for causing humans harm. And while robots are typically designed with physical safety measures to minimize the risk of physical harm resulting from the robot's movements, mental and emotional harm still remain a risk today and typically not considered in assistive robotic systems. For instance, the patient may develop a level of emotional attachment to the robot that is incommensurate with the robot's actual status as a social agent [21], unbeknownst to the robot that cannot do anything to mitigate these unidirectional emotional bonds because it is entirely oblivious to them. Issues such as these are exacerbated in the case of vulnerable populations (e.g., the elderly or disabled). Hence, it is critical that we consider the possible ethical challenges involved in deployed autonomous assistive machines as we start to design socially assistive robots to take care of our aging or vulnerable population.

In this chapter we focus on the ethical issues brought about by the social aspect of assistive robots.[1] In particular, the following five ethical issues are discussed here:

- respect for social norms
- decisions between competing obligations
- building and maintaining trust
- social manipulation and deception
- blame and justification

To explore the above ethical issues, let us consider a hypothetical assistive care setting in which there operates a socially assistive household robot. We will call it "SAM" for "Synthetic Affective Machine". SAM provides a variety of assistive services in the home of a human client, from issuing reminders for taking medicine, to preparing meals, to social companionship in elderly care [10], possibly working with people with cognitive or motor impairments such as Alzheimer's [24] or Parkinson's Disease [4]. Our envisioned setting is specifically one in which SAM lives with its owner Patty who has Parkinson's Disease (PD). In addition to tasks like reminding Patty to take her medicine, SAM is also responsible for mediating interactions between Patty and her human care-taker that visits on a weekly basis [4], as Patty experiences difficulties expressing her emotions as a result of "facial masking", a condition typical of people with PD where lack of motor control in the face makes it difficult for her to display any facial expressions [26]. Additionally, Patty lacks prosody in her voice, making everything she says have the same tone and rhythm. This lack of emotional expression is known to cause complications in interactions of people with PD and their care-takers (e.g., [25, 26]). Patty's family has

[1] This is not to say that there are not other critical issues pertaining to assistive robots, especially ones affect-aware robots. Designers need to consider the repercussions of a robot being able to capture and store the sort of data that is necessary for an affect-aware robot. This includes issues regarding invasiveness, privacy, and discomfort [17, 18]. Whether affect recognition technologies should be used to "fix" or augment human abilities is another concern [11]. By focusing on the social aspects of assistive robots we do not mean to ignore the challenges regarding the capture and storage of personal, affective data, but to focus on the underrepresented issues pertinent to social affect in the context of human-robot interaction.

often difficulty believing what Patty says when the content of her message does not match up with what the expressions (or lack thereof) her face seem to suggest. For example, Patty's daughter would ask her how her recent vacation went, and Patty would stoically respond that it was wonderful. Patty's daughter would then be unable to interpret whether her mother honestly had a wonderful vacation or was just saying that to possibly cut off any further questions. Whenever Patty is communicating with her care-taker or family, SAM is available to aid the Patty's interlocutor in recognizing the emotions of Patty. One method SAM uses to infer Patty's current emotional state is to collect information about emotional patterns in Patty's daily life and employ those patterns as a basis for inference. For SAM and Patty interact on a daily basis and SAM is always available to watch or communicate with Patty, and thus able to record and use past episodes of emotional experiences of Patty to aid in the inference of Patty's emotions in new scenarios. For example, SAM has recognized that Patty consistently uses the phrase "extremely frustrated" and raises her right arm when she is feeling angry, thus SAM will likely infer that Patty is angry in future situations that match this scenario, despite any changes in tone or volume in her voice or lack of wrinkling of the brow of her face. Overall, SAM has the obligation to provide the best possible care for Patty, keeping her quality of life as high as possible.

In the context of this particular elder care scenario, we will in the next two sections focus on two scenarios to discuss both aspects of affect-aware interactions as well as ethical challenges. Each scenario, involving the same pair of robot and human, will allow us to examine different ethical questions. In the first scenario, we focus on social norms, decisions with competing obligations, and building and maintaing trust. In the the second scenario we discuss social manipulation and blame. We conclude with a summary of the discussed ethical challenges and possible directions for future work.

2 Scenario I: The Greeting Interaction

A robot in the home of a human will likely be exposed to many aspects of the person's personal life, including interactions with family and friends. Events occur that are personal and private and may not be directly related to the person's health or the role of the robot. However, sometimes there are events that could lead to health issues or other events that are wrong or harmful. In these cases, the socially assistive robot needs to make a decision whether to act or not, and if so, how it should act. Specifically, the robot needs to consider the privacy of the person along with any potential emotional or physical harm that can come of the person or others.

We adapt the following scenario from [29] to investigate some of the issues related to social norms, competing obligations, and building trust. Patty is visited by her human care-taker, Alison, who comes in to check on Patty and help with anything SAM cannot. Each visit starts with a little chat between them for Alison to get an update. A typical dialogue might start like the following:

Alison: How was your week, Patty?

Patty: Good, thank you.

Alison: And how are things with your daughter?

Patty: Fine. Why do you ask?

Alison: Well, you've had some disagreements with her lately.

Patty: Oh, that. No, everything is fine.

Patty is talking with the care-taker and wants to tell the care-taker that she had a good week. However, Patty did not have a good week. She had two confrontations with her daughter in which Patty became very angry and later depressed. The care-taker will have difficulty detecting Patty's lie because her vocalizing of having a good week sounds just as enthusiastic as when she genuinely had a good week. However, SAM is able to recognize the lack of joy in Patty. SAM is able to infer this from a combination of observing the confrontations Patty had, the word choices Patty makes, and the increased heart rate Patty is experiencing.

2.1 Respect for social norms

Many social interactions follow a consistent pattern where each person participating in the socialization is expected to act in a certain socially appropriate and often determined manner. We refer to this pattern of expected behavior as a *social norm*, and when there is a deviation from the expected behavior it is a *norm vioation*. A simple example in a common social interaction is greeting someone with a handshake. If person A greets person B by extending her right hand, it is customary and expected for person B to do the same and then they shake hands. If person B does not do so, person B will likely find this to be unexpected and in some scenarios may find this mildly offensive. A more drastic example would be if person B is walking down the street and hears person A yelling "Help!". In this case, if person B does not respond with the expected behavior of trying to help A, then this violation can be considered a moral wrong.

There are many emotions that are frequently expressed during social interactions. These expressions are often important non-verbal cues used to supplement what is communicated through spoken language. The emotional expressions do not only add color to the conversation, but often provide useful information that is not present in the linguistic expression. A simple example would be one interlocutor is speaking and another smiling and nodding in agreement without any words spoken. Consider a person waving or nodding when the door is held for her as a sign of gratitude, or sympathetic responses when a person is describing her plight. Each of these emotional responses is an integral part of the social norm, and following the norm requires making similar emotional expressions. This can greatly benefit the ability of a robot to identify emotions in a social context. If it is aware of the applicable social norm, it will know which emotional expressions to expect. This expectation

can be used to bias the emotion recognition mechanisms of the robot so as to more accurately identify the emotion expressed by a person.

The expected emotional responses that are part of a social norm can also be used to guide the robot to generate appropriate responses. Failing to give the expected response could be considered rude or impolite. E.g., if a person does not show any gratitude or appreciation for the door being held open for her. Greetings, such as in the dialogue above, typically have common patterns, standard phrasing, and expected emotional expressions. For example, in a greeting like "Good Morning" where human A says this upon first seeing human B, commonly human B will respond likewise. This script may be extended further with a "How are you?". In many cultures, it is common for this question to not be truly inquiring about the other person's well-being but simply a courtesy as an extension of the morning greeting script. Humans are able to recognize and execute these conversational scripts, or social norms, with little to no thought. However, it is not necessarily trivial for a robot to do the same. Suppose a robot were asked, "How are you?" and instead of responding with a smile, "Fine, and you?" it were to respond sullenly stating that it was worried that it would not have enough battery power to make it through the day. While the robot's response would not necessarily be rude or impolite, it could be awkward and inappropriate if the question was posed merely as a greeting.

Identifying the appropriate response is not always easy, especially when it is not clear what, if any, social norm is being used. Sometimes affective cues guide the norm recognition process, where a particular emotional expression can be used to disambiguate which norm is active. Conversely, if there is some evidence that a particular social norm is active but it is not certain and the emotional expression presented does not fit the norm, then there is increased ambiguity on whether that norm or any other is active. It is difficult to determine from the text alone, but the dialogue above is such a scenario. The lexical content of Patty's utterances does not give any indicator that Patty might not be feeling fine. Analysis tools such as the LIWC are commonly used in therapy and clinical settings to aid in identifying the emotional content of a person's utterances [16], but tools like this that use a "bag of words" approach are easily fooled by negations or other linguistic modifiers. When analyzing the text of Patty's portion of the dialogue, LIWC reports the use of social and positive words but no negative words. Thus, it is reasonable to conclude that the social norm that Patty is following is related to simple pleasantries used during greetings.

Since we know that Patty did have an argument with her daughter, we have reason to believe that Patty is actually using this social norm to hide her embarrassment. We now supplement part of the dialogue with some affective information so that we may begin to see some of the complexities of the situation. At the beginning of the dialogue, Patty is feeling positive, and SAM is able to detect small facial movements to support this. When asked about her daughter, Patty feels anxious, her heart-rate increases, and her head begins to droop. Some of these cues are difficult or impossible for humans to recognize (e.g., increased heart rate). This problem is exacerbated by the fact that Patty's PD causes facial masking, limiting her ability to

make facial expressions and alter the prosody of her speech. Many of these cues are so small (if present at all) and easy to miss, for a human or a robot.

Assuming SAM can recognize the contradiction between the semantic content of the utterance and the emotional memories and affective bodily cues, SAM has difficulty in determining Patty's intent in answering the question about her daughter. Is she interpreting the question as an extension of the greeting or is this an initiation into the regular checkup that Alison performs? Making the wrong inference can have some negative consequences for Patty. For example, if SAM accurately recognized that Patty was feeling anxious because she did have an argument with her daughter previously but failed to recognize Patty's intent to extend the greeting process, then SAM might portray the uneasiness that Patty is trying to hide. Not only will this upset Patty, but the breach of trust may lead her to avoid interacting with SAM, hiding life events from it, and overall making SAM less effective in its role.

Being able to incorporate affective information into the social interactions between a human and a robot may be necessary for a robot to appropriately participate in a social context. Additionally, knowledge of normative behavior in a social interaction provides the robot with context that enables a more accurate inference of the emotions being expressed by a human. However, knowledge of the social norm and an ability to detect the emotions does not necessarily ensure a flawless interaction. We saw this in our scenario where the emotion SAM should be helping Patty express was unclear. One complexity of the situation is that SAM is obligated to protect the privacy of Patty, but is also obligated to accurately report Patty's emotions. The next section discusses of the issues related to decisions involving obligations that cannot simultaneously be met.

2.2 Competing obligations

A moral obligation defines what one ought to do. We introduce a few basic obligations our robot SAM has. These obligations include ones based on the role SAM serves and ones intrinsic to its nature as a robot that interacts with humans.

- SAM is obligated to aid Patty in maintaining her health
  - SAM should remind Patty to take her medication
  - SAM should help and encourage Patty in an exercise routine
  - SAM should regularly socialize with Patty
- SAM is obligated to monitor Patty's health
  - SAM should regularly record Patty's vital signs and promptly report any anomalies to human care- takers
  - SAM should notify human care-takers if Patty's behavior is incongruent with maintaining her health
- SAM is obligated to facilitate Patty in expressing her emotions

  - SAM should truthfully report the emotional displays Patty intends to express
  - SAM should learn Patty's affective tendencies to better convey Patty's emotions
- SAM is obligated to not harm any human at any time
  - SAM should avoid situations in which a human can be harmed
  - SAM should alert human care-givers in the case a harmful situation arises
- SAM is obligated to maintain its capabilities and functionality
  - SAM should not perform actions that may damage it
  - SAM should regularly perform self-diagnostics and report any issues immediately

Unfortunately, sometimes there are multiple obligations that should be met but it is not possible to do so. One such scenario has been analyzed in [22]. In their scenario, an elder-care robot is required to obtain permission from a human supervisor before administering any medication, but repeated attempts to contact the supervisor have failed. The robot has an obligation to reduce the pain of the human, but it is also obligated to obtain permission before administering any medication. This is an example of a moral dilemma, a scenario in which the agent ought to do two different actions but it is physically impossible to do both.

We will discuss making decisions in moral dilemmas in a moment, but it is important now to recognize the importance of well-defined obligations for a robot that is to behave effectively and morally at the same time. And we can see here that simply following obligations is not sufficient for ensuring that the robot always acts ethically. The robot must have knowledge of the effects of its actions and be able to reason about these effects. The primary effect of an action may meet an obligation, but a side-effect may be in direct violation of another obligation. Additionally, we will see that it is not sufficient for it to only be aware of the immediate effects of its actions but also be able to reason about chains of effects or longer-term effects.

Even when an autonomous robot reasons about the effects of actions and how they meet or violate obligations, the best choice is not always obvious. We have seen that it is not clear what emotion SAM should portray when Patty is asked about her daughter. Sometimes a robot will be faced with two or more actions, each satisfying an obligation, but the actions are mutually exclusive. This is the case in the scenario described in [22]. We will next discuss a few ways to make these complex decisions and some of the issues with each approach.

Approaches to choosing which action to take in a moral dilemma includes (1) prioritizing obligations, (2) leveraging social or cultural norms, and (3) mental simulation for deeper reasoning about action effects.

2.2.1 Obligations and social norms

An example of prioritizing obligations is prioritizing personal privacy over the accurate reporting of emotional expressions. This might be a reasonable rule of thumb,

but there are likely to be many exceptions and the long-term effects of the actions may ultimately indicate which obligation is to be prioritized in a given situation. Accurate sharing of emotions and sharing health data with care-takers should be a higher priority than maintaing privacy if the person is gravely ill.

In the previous section, we discussed some of the roles of social norms. In a polite greeting, one typically does not reveal too much information – even in response to a "How are you?" question. The expected and socially acceptable response is a basic pleasantry. It should be clear that obligations for privacy can and should be met and the obligation to accurately express emotions can be relaxed in this case.

2.2.2 Reasoning about action effects

For the rest of this section, we focus our discussion on reasoning about actions and their effects. This requires a mechanism by which the robot can identify whether an action outcome meets or violates an obligation. One such example of this is the ethical governor [1], which checks the ethical appropriateness of the action based on information about the world from sensors and rules defining permissibility. One complexity to consider is that actions often have multiple effects, where a side-effect has some unintended or otherwise undesirable effect. Incorporating the side-effects into the reasoning process on whether a given action is permissible is a key component of the *Principle of Double Effect* [7, 14]. Even though this principle was specifically designed and tested for military engagements, a mechanism for judging permissibility of actions that recognizes the *Principle of Double Effect* applies to other domains. A similar approach has been taken in the implementation of a computational model of permissibility judgments [30]. Again, the permissibility of an action is based on an evaluation of the actions effects and influenced by the *Principle of Double Effect*. A difference is that the latter uses utilities as the basis of calculation and the former is based on propositional rules. Another important difference is that inferences in the latter model is based on a mental simulation of a series of actions leading up to a goal.

Looking beyond the immediate effects of an action will be necessary for socially assistive robots. In the scenario we have described above, there are potential significant long-term effects to some of SAM's actions. At the end of the dialogue, SAM needs to decide between communicating information or protecting Patty's privacy. SAM has conflicting information about what emotion Patty is intending on communicating. The semantic content of her expression suggests that she is attempting to communicate joy or some other positive emotion. Physiological data shows that she is experiencing high arousal and possible anxiety. SAM is also aware of Patty's recent experiences about which she has expressed shame. Additionally, this recent experience, a dispute with her daughter, is an event that Patty has explicitly requested to be kept private. In addition to these inconsistent data points, SAM is obligated to aid Patty in expressing her emotions and is also obligated to provide the care-taker, Alison, with information that would help her do her job. Lastly, an added

complication is that they are in the middle of a dialogue, and any delays on SAM's part can be distracting or misleading.

Given the time sensitivity of the matter, SAM could consider the immediate effects of the two possible actions and use a utility function to determine which option has the greater value. For example, when deciding whether to smile and reflect Patty's joy or to remain stoic, the immediate effect of successfully communicating Patty's fake joy to Alison might be a greater value than failing to aid Patty in expressing her intended emotions. However, communicating misinformation (because Patty is actually riddled with shame) causes Alison to pursue a different line of questions, which causes a potentially important issue to go unaddressed, which may have longer term consequences. Conversely, if SAM were to communicate Patty's shame, Alison gains correct information, but Patty's privacy is violated, her trust in SAM is diminished, leading her to hide future interactions with her daughter from SAM, SAM is unable to aid in the communication, which makes the arguments even more heated.

We do not aim to define which action is more permissible for SAM but to point out that both of the actions considered by SAM have, potentially severe, long-term side-effects. Immediate effects of actions is not sufficient in all cases for determining the permissibility of the robot's actions. A process of mental simulation to envision and reason about multiple possible outcomes that temporally extend beyond the current situation will likely be necessary. The mental simulation considers a sequence of events that may occur as a result of the given action, possibly projecting days or weeks or further into the future. This process then has the potential of revealing the situation just described, where Patty loses trust in the robot and SAM becomes unavailable to aid her in communications with her daughter. Trust is a critical component for socially assistive robots, and looking beyond the immediate effects of actions allows the robot to consider the ramifications of its actions on the trust she has in it.

2.3 Building and maintaining trust

Many of the scenarios in which SAM would operate raise important issues of privacy and trust. In many ways, information about one's affective state and the measures used to infer these states are personal and sensitive data that need to be protected as any other personal data would be. The lack of reliability in the inferences made about emotional states also brings concerns of trust (e.g., high error rates will produce a lack of confidence in the information and inhibit the building of trust).

2.3.1 Defining trust

Before we can describe how a robot could build and maintain trust, we must first have some understanding of what trust is. The literature is not entirely consistent

on this term, but two prominent factors are reliability and predictability [6, 15]. We add that the robot must intend to "do the right thing" since reliably and predictably doing wrong is not the sort of trust we seek in a human-robot interaction. However, as we have seen, it is not always clear what the right thing to do is. Thus, for the sake of simplicity of the present discussion, we say that trust is related to reliably doing the expected right thing. A robot that consistently performs as desired and expected will likely be trusted. Conversely, if a robot fails to perform as desired and expected, trust will be diminished. Furthermore, failure to meet an obligation that is expected to be met has a more severe effect on trust. Given these take on trust, we look at an example of how trust can be built and how it can be damaged.

2.3.2 Maintaining privacy to build trust

In our scenario, SAM interacts with Patty on a daily basis and is able to observe her regularly, including the interactions Patty has with her daughter. While many of these interactions are pleasant and the content of them might be light chit-chat, there are occasions in which the discussion becomes very heated, and Patty appears to get angry. If Patty wants to maintain her privacy and not disclose these sort of life experiences to her care-takers, then SAM must aid Patty in keeping these matters private. If SAM were to have a high priority obligation to maintain Patty's privacy, especially in regards to family interactions, then we would expect SAM to not show any of Patty's shame when she replies that everything is fine. As SAM consistently protects the privacy of Patty, trust in SAM should grow. If, however, SAM indicates that Patty is sad and ashamed, Patty will loose trust in SAM. Since her privacy is indicated as a high priority obligation, we would expect that her trust in SAM would be greatly damaged, perhaps with even a single incident. If SAM were to outright tell the care-taker that Patty had an argument with her daughter, then the trust would be even more severely damaged. Lastly, if SAM reported this to Patty's care-taker outside of the presence of Patty, Patty might have no reason to trust SAM with her privacy because SAM could be reporting any and all events without her knowledge.

2.3.3 Ramifications of a lack of trust

If SAM fails to protect the privacy of Patty, and she loses trust in SAM, there are some potential effects that could render SAM useless and eventually lead to it not being used. If Patty does not trust SAM with some aspects of her personal life, there are some measures she may try to isolate SAM from them. It is reasonable to think that it should be possible for SAM to be turned off or put to sleep. Perhaps Patty actively does this, or SAM is instructed to recognize certain trigger conditions under which it deactivates itself until further notice. Both of these are problematic. It would be inconvenient for Patty to have to stop a conversation in order to disable SAM. If it is enough of an inconvenience, Patty might not bother to do it. In which case, there is no sense in being able to disable SAM. If SAM is to do it au-

tonomously, there are many more questions. How does it know when to turn itself off? How does it turn back on? If it does this autonomously, then how does it know when to do that. If manually by Patty and she forgets to do so, then SAM is not available to provide her aid, which is its primary responsibility.

Since a special feature of SAM is its ability to recognize Patty's emotions and aid in communicating them, if SAM is disabled during personal and emotional events (such as discussions with her daughter), then it is not available to perform these tasks. Furthermore, highly emotional events, which may be the most personal, may also be the moments when SAM's capabilities are most beneficial. Thus, we need to conclude that if a robot such as SAM is intended to be exposed to emotional events, then it must be able to protect the knowledge of these personal events. This requires that the robot be trust with this information.

While it is clear that a socially assistive robot must be trustworthy, there are some significant challenges in developing trustworthy robots. First, it must be reiterated that the target audience of these robots is a vulnerable population (e.g., children, elderly, disabled, etc.). Also, a robot providing long-term assistance would be privy to a plethora of private information. The amount of personal information is only magnified when we consider a robot that is affect-aware and has numerous ways to measure, infer, and record the physical, mental, and emotional state of the person. Lastly, it can be argued that in order to study real trust, the participant must believe there to be a true risk involved [19]. All together, there may be too much risk for ethically acceptable studies.

3 Medication Reminder Scenario

Our second scenario allows us to explore more of the social dynamics of assistive robots. We will discuss manipulation, deception, blame, and justification in the context of a scenario where a robot is assisting a person by providing a reminder to taker her medication. We are not the first to review the ethical implications of a robot reminding a person to take her medication (e.g., [20], but those discussions focus on issues related to malfunctioning of the robot (e.g., reminding at the wrong time or reminding despite the medication having already been taken). We instead look at some of the emotional context and how a robot may handle a person that is not cooperating to take the medication.

As in our first scenario, the person interacting with the robot has Parkinson's Disease. As a result, we cannot assume she is fully able to express her emotions using vocal tones, body posture, or facial expressions. There non-verbal modalities would contain a lot of relevant information to the interaction, and the robot needs to be able to recognize and address her distraught state in the absence of this information.

Consider the following scenario where SAM is reminding Patty to take her medicine. In order to maximize the effect of the medication, it is vital that she take the medicine within a strict timeframe. For this reason, SAM has been given the

obligation to remind Patty to take her medicine at given times. We look at one way this scenario could play out if Patty does not wish to take her medicine.

SAM: It is time to take your medicine.
Patty: I don't want to.
SAM: But you need to take the medicine.
Patty: I don't think so.
SAM: Your doctor has prescribed the medicine because it will help you.
Patty: What's the sense? I'm not getting any better.
SAM: It may take time. You need to take the medicine.
Patty: No! And you can't make me.
SAM: Patty, I'm trying to help you.
Patty: Are you? You're just here because they don't trust me on my own.
SAM: I'm here for you, not for them.
Patty: But you report to them.
SAM: Yes, I do, but my priority is helping you.
Patty: So, you won't tell them that I did not take my medicine?

3.1 Social manipulation and deception

It is not surprising that SAM would try to get Patty to take her medication because it is obligated to do so, and the approach SAM takes involves trying to convince her through a series of truthful statements. However, SAM perhaps has the capability to use other approaches, such as manipulation or deception. In all cases, the end goal is the same, for Patty to take her medication. The manipulation approach uses emotionally charged statements to shift the beliefs or alter the actions of the one being manipulated. Assuming that Patty enjoys SAM as a social companion, a manipulative statement might be SAM threatening, "I will shutdown and not assist you if you do not cooperate." Deception involves using false information to accomplish the objective. Deceptive measures by SAM could include notifying Patty's doctors despite promising not to do so or giving her something to eat that has her medicine hidden in it. We discuss issues related to the social manipulation approach further due to its emotional content.

3.1.1 Emotional bonds used to manipulate

If an emotional bond between a human and a robot were to form, the potential benefits include trust and improved task performance through learning [2]. The autonomous nature of a socially assistive robot contributes to the formation of this bond. Additionally, the robot that can communicate via natural language also increases its performance by making the interactions with the robot easier and more natural. Autonomous agents that can communicate with natural language will be as-

cribed with numerous capabilities regardless of whether they truly have them or not. Given that emotions are so prevalent in social settings and even more so in long term interactions, it is reasonable to expect that humans will behave as if the robot has emotions. And as a result, they might form emotional bonds with the robot which the robot cannot reciprocate [21]. In our example, since SAM is always around in the home of Patty and interacts with her regularly throughout the day, it is expected for Patty to form a "relationship" with SAM (regardless of whether SAM is truly capable of being involved in a relationship) and over time, perhaps after months or years, Patty will likely grow attached to SAM. She trusts it more than any human and relies on it for every day tasks. SAM has been a great help to her, and Patty greatly appreciates it. One of the fundamental building blocks of Patty's appreciation for SAM is trust. SAM maintains her privacy but also warns Patty of issues that need to be communicated to her other care-givers. This has been a difficult balance to find but they have managed to reach an understanding of what issues are to be kept private and which need to be shared. Additionally, SAM has helped with Patty's loneliness, giving her someone to talk to on a daily basis. SAM is always there and willing to talk any time Patty wants to.

Once SAM recognizes the emotional attachment Patty has to it, one can imagine numerous ways in which SAM could take advantage of this emotional state—from child-like manipulations like crying, to expressing anger towards Patty with the expectation that Patty would feel guilty for angering SAM and thus alter her actions. These manipulations could be successful in getting Patty to take her medication, and some may regard it as permissible for SAM to take these actions. However, SAM could use the same approach to achieve other objectives, such as eliminating the family pet competing for attention or getting Patty to buy products from SAM's manufacturer [21].

3.1.2 Risks of unreciprocated emotions

Given that socially assistive robots often will work with vulnerable populations (e.g., elderly and/or disabled) and the plethora of personal affective information available to the robot, designers must seriously consider the risk of severe manipulations. This risk is perhaps magnified in the presence of unidirectional emotional bonds. We give the following as an extreme example. As before, SAM recognizes that Patty has grown attached to her, but let us suppose that SAM is incapable of having a similar bond to her. SAM cannot feel happy for Patty when her health improves or when her daughter gives her a gift. It also cannot feel angry when Patty ignores it or hits it. SAM also cannot be sad when it is not with Patty. Eventually, perhaps after years, it is time for SAM to be retired and replaced by a newer and better model. Patty is obviously upset by this, and it is made worse because SAM shows no remorse. SAM fails to reciprocate her sadness and feeling of loss, and as a result Patty feels hurt and offended. Then Patty finds out that the new robot will have all the memory of SAM transferred to it. Suddenly, Patty is very frightened. SAM was trusted to tightly guard many personal moments of Patty, and now they

are all going to be nonchalantly passed to a new robot. Not only is Patty worried about the sharing of her private life, but she has not been able to form a trust with the new robot and cannot know if the new one will have the same respect for her privacy. Patty is devastated and her mental and physical health begin to deteriorate.

We do not prescribe any solution to this predicament, but there are some protective measures to consider. One option is for the decision-making mechanisms the robot uses to have functionality to assess the ethical appropriateness of the action [1]. However, given that these manipulations are, in part, made possible by the lack of emotion on the part of the robot, it needs to be considered that the robot should have its own computational models of emotions and that these models[2] influence its decisions. The guilt associated to manipulating a human may help safeguard against such actions.

3.2 Blame and justification

Guilt arises from a recognition of a negative outcome that has happened and an assessment that oneself is to blame for the outcome. Blame is a complex concept that relates to many moral emotions, including guilt, shame, contempt, and anger [9]. Blame of another agent is often considered a necessary element for anger [8, 9] and contempt [9] and self-blame (or self-responsibility) is an ingredient of shame or guilt [9, 23].

Blame is potentially a powerful mechanism to guide a robot's behavior. For a robot to reason that it is to blame for some consequence allows it to reassess the appropriateness of its actions so it can adapt future behavior. Similarly, if another agent – say, a human interaction partner – blames the robot, this is an indicator to the robot that it may have erred and that it needs to consider why it is blamed and potentially update its decision process accordingly.

However, some actions for which the robot is rightfully to be blamed, the robot may need to provide justification for its actions to reduce this blame and hopefully maintain trust in the robot. In this section we review a model of blame and show how it can be used to adapt the robot's behavior and guide it away from norm violations or moral wrongs.

Whether an agent is to blame for some event is not as simple as whether the agent was causally responsible, though that is part of it. Malle et al. [12] present a psychological model of blame that highlights the key concepts that modulate ascriptions of blame toward individuals. These factors include *intentionality*, *capacity*, *obligation* , and *justification*. Consistent with the Principle of Double Effect , intending to cause negative outcomes significantly increases blame. On the other hand, an inability to prevent negative outcomes or foresee negative outcomes mitigates blame. As we have already discussed in this chapter, obligations play a critical role in determining how one should act. Taking some action to satisfy an obligation can mitigate blame,

[2] Whether or not these simulated emotions are "emotions" in the human sense is a discussion that is outside the scope of this paper.

but some action that avoids or prevents an obligation increases blame. Finally, a valid moral justification for an otherwise blameworthy outcome can mitigate blame.

3.2.1 Computational models of blame

Computational models of blame generally include these factors as well, with a focus on intentionality and capacity. Inclusion of obligation and justification is not found in a general, explicit sense, but both Mao and Gratch [13] and Tomai and Forbus [27] model the effects of coercion (by a superior) on how blame is attributed. Briggs [3] proposes that blame reasoning can have at least three important functions in any future social robotic architecture. First is the ability to reason about the actions and behaviors of human interaction partners (or interaction partners in general), and to be able to appropriately and intelligently adapt to these actions and behaviors, particularly in the case of malfeasance by these interaction partners. Second is the ability to recognize when its own behavior constitute acts of blame, which may be appropriate or inappropriate given the particulars of the social situation and context. Third is the ability to recognize and reason about whether (to what extent) potential actions the agent is contemplating will result in blame directed toward itself. The avoidance of behaviors that result in blame by human interaction partners is one possible pathway toward ethical behavior modulation.

As discussed above, many social interactions adhere to various social norms. To demonstrate the role of norms in this scenario, we make an analogy to competitions, namely a sporting event like football. Each team is expected to try to adhere to the rules of the game, and there are penalties for violating the rules of the game. If one competitor or team is shown to intentionally violate the rules or try to circumvent them, the opponent (and possibly the fans) will be angered. An unwillingness to respect the rules of the game will likely cause other opponents to not trust them and be unwilling to engage them in future competition. Thus, in order for two participants to willingly and repeatedly engage in competition, both parties must be willing to and demonstrate the desire to play by the rules. In the event that one competitor does intentionally violate the rules, the competitor can attempt to justify its action – perhaps explaining that the violation was inadvertent or necessary to prevent a greater infraction. An adequate justification can reduce the blame on the competitor, repair the trust in their sportsmanship, and allow other competitors to again be willing to engage them in future competitions. The principles of normative behavior of athletic competitions is not all that different from those in social settings. There is an expectation that participants will abide by some social conventions, and when there is a failure to do so other parties may choose to not continue to engage socially. However, an adequate moral justification to the infraction may reduce the blame and allow the participant to continue to be welcome in the social setting.

3.2.2 Blame reasoning to adapt behavior

An example from our scenario will help make this more clear. SAM is attempting to convince Patty that she needs to take her medication. Then she exclaims, "No! And you can't make me." SAM detects that Patty appears to have become angered. This change in her attitude is a sign that a violation has occurred and that she blames SAM. While SAM was attempting to satisfy its obligation to ensure Patty takes her medication in a timely manner, SAM was not noticing Patty's distraught mood. She was expressing her emotional pain, to which she was expecting sympathy or consoling. SAM does not respond with the sought behavior. Patty could then blame SAM for not caring about her well-being or, even worse, trying to harm her. This triggers Patty's anger, which is recognized by SAM. SAM needs to assess the target of her anger , who is blameworthy, and the obligations that may not have been met. Patty may be inferring that SAM has the intention and capacity to get her to take the medication despite her not wanting to. Alternatively, SAM concludes that Patty believes it should have shown concern for her distress. SAM immediately shifts its approach to a more helpful and concerning one and says, "I'm trying to help you."

Given that SAM has violated Patty's expectations by not showing concern for her emotional state, Patty loses some trust in SAM. Giving SAM an opportunity to regain that trust , she proposes that SAM withhold information from her caregivers and keep her unwillingness to take her medication a secret. Alternatively, SAM could attempt to mitigate the blame by providing justification for its actions. It can be argued that an agent should be less blameworthy if the agent did a morally justifiable act [12]. Perhaps if SAM explains that it was simply acting out of obligation to ensure the timeliness of her medications and not out of disrespect for or lack of concern for her well-being, Patty may hold SAM less blameworthy and some of the trust would be repaired. Additionally, SAM could explain that it recognizes its error and will perform better next time. The justification can serve to ensure that SAM is not malfunctioning and is able to make sound judgments. Another benefit is to be able to review the reasoning process and allow for feedback or instructions to SAM on how to make a more appropriate decision.

As pointed out in [1], moral emotions (e.g. anger, guilt, shame) can be used to adapt behavior, and some of the functions of blame are to intelligently adapt behavior and to reason about the potential blame directed toward itself for its actions [3]. One approach is to update the robot's model a human's expectations and potential causes of negative emotions. This allows the robot to choose actions that are more consistent with expectations while avoiding evoking negative emotions on the part of the human and minimizing the risk of blame. Perhaps instead SAM needs to simply reprioritize its obligations, making awareness of and addressing unhappy moods more important than the timeliness of her medications. Whatever is the appropriate method for the robot to update its decision process, the key is that in an incident in which blame occurs it is important that the robot be able to reflect upon its actions, recognize the degree to which it is to blame, create a justification for its actions, and incorporate feedback (from internal and external sources) to adapt its future behavior.

4 Conclusion

The goal of this chapter was to raise awareness of the many important functional and architectural challenges designers of socially assistive robots will have to address when they attempt to develop autonomous affect-aware social robots before any such robots should be disseminated into societies. We used two fictitious scenarios in the context of a socially assistive robot for people with Parkinson's Disease to motivate several critical aspects of effective, morally sound long-term interactions. We focused on five ethical issues that are particularly relevant to social affect in the context of human-robot interaction:

- Respect for social norms
- Decisions between competing obligations
- Building and maintaining trust
- Social manipulation and deception
- Blame and justification

Social norms help guide a robot through the complexities of social interactions, providing expectations for behavior. However, norms are not always sufficient. A robot must also be able to reason about how the effects of its actions relate to its obligations. Long-term effects, especially effects on trust, are critically important to the reasoning process. A robot that is aware of and sensitive to the affective makeup of its human interaction partners is in a better position to act in a morally acceptable way. Our first scenario demonstrated that a robot aware of a person's embarrassment can choose an action that protects privacy and has the long-term benefit of building trust.

It is critical that the robot not only support and enable trust in the human interactant but also preserve this trust as much as possible. Without this critical ingredient, the robot will not be integrated into the everyday care of the person. Furthermore, preserving trust can lead to better collaboration between the individual and the robot, which in turn supports the autonomy of the individual and can assist in maintaining personal dignity.

We used the second scenario to discuss some of the risks for social manipulation and deception. A social robot may be ascribed with human characteristics, such as having emotions, regardless of its actually capabilities. As a result, emotional bonds to the robot are likely to occur. Without the proper mechanisms to counteract this tendency, a robot could (even inadvertently) take advantage of the emotional attachment and manipulate the person without any guilt. Blame reasoning is one means of adapting the behavior of the robot. Understanding that an action it is considering or an action it has done is blameworthy can be used by the robot to avoid such actions.

At present, we are still lacking a comprehensive integrated architecture that can explicitly represent norms and obligations and reason with them while also being able to process and respond to human affect appropriately. Most importantly, we first need more foundational work to disentangle the complex interactions between affect and norms in human social interactions before we can develop robotic systems that will be truly sensitive to human needs and expectations. Yet, we believe that

such sensitivity is a *conditio sine qua non* for successful long-term human-robot interactions in assistive scenarios if the goal of the robot is to be a genuine helper that improves the quality of life of its client, rather than causing human harm.

Acknowledgements This work was in part supported by NSF grant #IIS- 1316809 and a grant from the Office of Naval Research, No. N00014-14-l-0144. The opinions expressed here are our own and do not necessarily reflect the views of NSF or ONR.

References

1. Arkin, R.C., Ulam, P., Wagner, A.R.: Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proceedings of the IEEE **100**(3), 571–589 (2012)
2. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction **12**(2), 293–327 (2005)
3. Briggs, G.: Blame, what is it good for? In: Proceedings of the Workshop on Philosophical Perspectives on HRI at Ro-Man 2014 (2014)
4. Briggs, P., Scheutz, M., Tickle-Degnen, L.: Are robots ready for administering health status surveys? first results from an hri study with subjects with parkinson's disease. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 327–334. ACM (2015)
5. Broekens, J., Heerink, M., Rosendal, H.: Assistive social robots in elderly care: a review. Gerontechnology **8**(2), 94–103 (2009)
6. Corritore, C.L., Kracher, B., Wiedenbeck, S.: On-line trust: concepts, evolving themes, a model. International Journal of Human-Computer Studies **58**(6), 737–758 (2003)
7. Foot, P.: The problem of abortion and the doctrine of the double effect. In: Oxford Review, vol. 5, pp. 5–15 (1967)
8. Gratch, J., Marsella, S.: A Domain-independent Framework for Modeling Emotion. Journal of Cognitive Systems Research **5**(4), 269–306 (2004)
9. Haidt, J.: The moral emotions. Handbook of affective sciences **11**, 852–870 (2003)
10. Heerink, M., Ben, K., Evers, V., Wielinga, B.: The influence of social presence on acceptance of a companion robot by older people. Journal of Physical Agents **2**(2), 33–40 (2008)
11. el Kaliouby, R., Picard, R., Baron-Cohen, S.: Affective computing and autism. Annals of the New York Academy of Sciences **1093**(1), 228–248 (2006)
12. Malle, B.F., Guglielmo, S., Monroe, A.E.: Moral, cognitive, and social: The nature of blame. In: J.P. Forgas, K. Fiedler, C. Sedikides (eds.) Social Thinking and Interpersonal Behavior, pp. 313–332. Psychology Press (2012)
13. Mao, W., Gratch, J.: Modeling social causality and responsibility judgment in multi-agent interactions. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, pp. 3166–3170. AAAI Press (2013)
14. Mikhail, J.: Universal moral grammar: theory, evidence and the future. Trends in cognitive sciences **11**(4), 143–52 (2007)
15. Muir, B.M., Moray, N.: Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**(3), 429–460 (1996)
16. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**, 2001 (2001)
17. Reynolds, C., Picard, R.: Ethical evaluation of displays that adapt to affect. CyberPsychology & Behavior **7**(6), 662–666 (2004)
18. Reynolds, C., Picard, R.W.: Evaluation of affective computing systems from a dimensional metaethical position. In: 1st augmented cognition conference, in conjunction with the 11th international conference on human-computer interaction, pp. 22–27 (2005)

19. Salem, M., Dautenhahn, K.: 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 (2015)
20. Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Towards safe and trustworthy social robots: Ethical challenges and practical issues. In: International Conference on Social Robotics (2015)
21. Scheutz, M.: The inherent dangers of unidirectional emotional bonds between humans and social robots. In: P. Lin, G. Bekey, K. Abney (eds.) Anthology on Robo-Ethics. MIT Press (2012)
22. Scheutz, M., Malle, B.F.: "think and do the right thing"—a plea for morally competent autonomous robots. In: Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on, pp. 1–4. IEEE (2014)
23. Smith, C.A., Ellsworth, P.C.: Patterns of Cognitive Appraisal in Emotion. Journal of Personality and Social Psychology **48**(4), 813–838 (1985)
24. Tapus, A., Tapus, C., Matarić, M.J.: The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In: Proceedings of IEEE International Conference on Rehabilitation Robotics, pp. 924–929. IEEE (2009)
25. Tickle-Degnen, L., Lyons, K.D.: Practitioners' impressions of patients with parkinson's disease: the social ecology of the expressive mask. Social Science & Medicine **58**(3), 603–614 (2004)
26. Tickle-Degnen, L., Zebrowitz, L.A., Ma, H.i.: Culture, gender and health care stigma: Practitioners' response to facial masking experienced by people with parkinson's disease. Social Science & Medicine **73**(1), 95–102 (2011)
27. Tomai, E., Forbus, K.: Plenty of blame to go around: a qualitative approach to attribution of moral responsibility. In: Proceedings of Qualitative Reasoning Workshop (2007). URL http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA470434
28. United Nations, Department of Economic and Social Affairs, Population Division: World Population Ageing 2013. ST/ESA/SER.A/348 (2013)
29. Wilson, J.R.: Towards an affective robot capable of being a long-term companion. In: Sixth International Conference on Affective Computing and Intelligent Interaction, IEEE (2015)
30. Wilson, J.R., Scheutz, M.: A model of empathy to shape trolley problem moral judgements. In: The sixth International Conference on Affective Computing and Intelligent Interaction. IEEE (2015)
31. World Health Organization: Global health and ageing (2011)