Blame, What is it Good For?

Gordon Briggs[1]

Abstract— **Blame is an vital social and cognitive mechanism that humans utilize in their interactions with other agents. In this paper, we discuss how blame-reasoning mechanisms are needed to enable future social robots to: (1) appropriately adapt behavior in the context of repeated and/or long-term interactions and relationships with other social agents; (2) avoid behaviors that are perceived to be rude due to inappropriate and unintentional connotations of blame; and (3) avoid behaviors that could damage long-term, working relationships with other social agents. We also discuss how current computational models of blame and other relevant competencies (e.g. natural language generation) are currently insufficient to address these concerns. Future work is necessary to increase the social reasoning capabilities of artificially intelligent agents to achieve these goals.**

I. INTRODUCTION

Continued development and improvement in the capabilities of autonomous robots will enable the deployment of these artificial agents in a variety of domains, including contexts that will involve human-robot interactions (HRI) with significant ethical import [1]. Due to this likely future, there has been a burgeoning community of researchers interested in the various ethical concerns that surround the deployment of autonomous agents in these context [2]. These concerns revolve around both the human-centered issues of responsibility (legal and moral) attribution in the case of unethical action by robotic agents and engineering-centered issues of how to design agents that act in these contexts. In particular, the latter questions have interested researchers in the field of machine ethics, who have sought to computationalize ethical reasoning and judgment in ways that can be used by autonomous agents to refrain from performing unethical actions. Various approaches to implementing moral reasoning that have been proposed range from use of deontic logics [3], [4] and machine learning algorithms [5]. Though much future work is warranted, these initial forays into computational ethics have demonstrated the plausibility of robots that have the ability to modulate their behavior with regard to moral considerations.

However, there are still a variety of challenges that face robot-ethicists seeking to ensure ethical outcomes to human-robot interactions. For instance, it is not clear that correct behavior modulation is as easily formalized in certain social situations as it is in others. Extreme behaviors such as harm towards humans (e.g. lethal force) may be constrained by near universally agreed upon principles that have been codified into formal regulations [3], but behaviors such as deception depend on a large number of circumstantial factors [6] and reside in a much more ambiguous moral territory. Additionally, it is not enough to be able to correctly reason about moral scenarios to ensure ethical or otherwise desirable outcomes from human-robot interactions, the robot must have other social competencies that support this reasoning mechanism. Deceptive interaction partners or incorrect perceptions/predictions about morally-charged scenarios may lead even a perfect ethical-reasoner to make choices with unethical outcomes [7]. What these challenges stress is the need to look beyond the ability to simply generate the proper answer to a moral dilemma in theory, and to consider the social mechanisms needed in practice.

One such social mechanism that others have posited as being necessary to the construction of a artificial moral agent is *blame* [8]. I agree with this assessment. In this paper, I will present the case for the need to incorporate mechanisms to reason about blame in a social robot, which stems from a few key concerns:

1) The need to adapt behavior in the context of repeated and/or long-term interactions and relationships with other social agents.
2) The need to avoid behaviors that are perceived to be inappropriately connotative of blame.
3) The need to avoid behaviors that could lead to blame, and consequently the deterioration of working relationships with other social agents (in the absence of well-formalized and overriding rules to modulate certain behaviors in specific domains).

Blame is conceived of as having both a cognitive and social component [9]. We begin by addressing concern 1, which pertains to the cognitive assessment of the characteristics of interaction partners and how these characteristics should influence interactions. Next, we consider blame as a social act, raising concern 2 and showing how the ability to model blame can allow robotic agents to avoid such impolite behaviors. Finally, we discuss how the ability to reason about blame by other agents toward itself (concern 3) can allow a social robot to modulate its behaviors in ways that are more helpful and prosocial.

II. COGNITIVE FACET OF BLAME

What blame is, precisely, has been a topic of philosophical debate for quite some time. In this paper, we are interested in the conception of blame as pertaining to assessments of the *character* of an agent. This notion of blame is not novel and can be traced back to prominent philosophers such as David Hume, who stated:

[1]Human-Robot Interaction Laboratory, Tufts University, Medford, MA USA gbriggs@cs.tufts.edu

If any action be either virtuous or vicious, 'tis only as a sign of some quality or character. It must depend upon durable principles of the mind, which extend over the whole conduct, and enter in the personal character. Actions themselves, not proceeding from any constant principle, have no influence on love or hatred, pride or humility; and consequently are never consider'd in morality. [10]

While some of the stronger claims of Hume in this regard (i.e. the strict coupling of action evaluation and character judgment) have been critiqued by contemporary philosophers [11], the idea of blame as fundamentally concerning inferences about character is an important one that has significant implications for the design of socially intelligent agents. To illustrate the need for character-oriented blame reasoning, we will present a simple HRI scenario in the following section.

A. Scenario

Consider a situation in which a delivery robot, whom we shall affectionately dub "Postie the PR2," has a long list of items to deliver throughout a large office complex. It determines that it cannot deliver all the items by the end of the day without help, so it enlists the assistance of a human employee, whom we shall call Bob, the helpful human. Bob graciously agrees to assist his robotic colleague in this parcel logistics task and takes several packages that are to be delivered to the far side of the complex. At the end of the day, Postie has successfully delivered its reduced set of packages, but before it is about to plug in a recharge for the night, it receives a complaint: an urgent and valuable package has not been delivered as expected! The robot checks its logs and determines that it indeed handled the parcel in question. It was one of the packages that it gave to Bob!

Postie abduces many possible explanations for the failed delivery, including, but not limited to:

- *Case 1* : Bob had taken the package for his own use.
- *Case 2* : Bob had forgotten to deliver the package as promised.
- *Case 3* : Bob had attempted to deliver the package, but found out the delivery destination was in a secure office where he did not have access.
- *Case 4* : Bob had mistakenly dropped off the package at the office of an employee with a similar name.
- *Case 5* : Bob had successfully delivered the package, but someone else took it.

Of course, determining which of these cases reflect the reality of the situation is a huge challenge in and of itself (requiring quite sophisticated social reasoning and interaction capabilities). Nevertheless, assuming this knowledge could be obtained (either by itself or from human investigation-partners), it is clear that these different cases paint a different picture regarding how Postie should interact with Bob in the future.

In Case 1, Bob has acted with antisocial intent (i.e. theft of someone else's property). Such a violation should significantly damage the trust Postie has in Bob, causing the robot to try to avoid similar interactions in the future. It should also trigger an obligation to report such a serious violation to the relevant authorities. Cases 2-4 present a less dire assessment of Bob's character, where it is not the case that Bob acted with malicious intent. However, these outcomes are perhaps indicative of other aspects regarding Bob that negatively impact how optimistic a delegating agent should be in assigning a similar task in the future. The precise impact, however, is dependent on the case and what it *specifically* conveys regarding Bob's characteristics. In Case 2, for instance, this outcome may be indicative of a general quality of forgetfulness that Bob may possess. Not only should this make Postie more hesitant to ask Bob to deliver packages in the future, but it perhaps means that Postie should be more hesitant to make any urgent/important requests of Bob (as he might forget them too). In contrast, Cases 3 and 4 are possible deficiencies that are more limited in scope. Specifically, Case 3 implies Bob should not be given tasks in the future that require deliveries to restricted areas, whereas Case 4 merely implies that Bob needs to be more careful with reading recipients' names in the future (or memorize a special case). Finally, in Case 5, Bob was not causally responsible for the disappearance of the package. Therefore, no inferences can be made regarding Bob's moral character or competencies.

What this simple scenario is intended to show is that appropriate adaptation for future interactions with other social agents depends on the precise *details* of why a particular outcome occurred–specifically, what characteristic of that agent makes sense of the outcome. Knowing the degrees to which an agent possesses particular traits is necessary to know what goals and actions (and consequently) what plans could be expected to result in task completion with particular interaction partners. For instance, delegating a task to a forgetful agent (case 2) has a lower probability of task success over the space of all possible tasks, whereas an agent that simply does not have access to one area of the workplace (case 3) will have zero probability of task success for all tasks that require such access and high probability of task success for tasks that do not (assuming no other modulating characteristics). However, do current approaches to modeling blame allow for this form of character inference? We discuss this concern below.

B. Limitations of Current Approaches

Malle et al. (2012) present a psychological model of blame that highlights the key concepts that modulate ascriptions of blame toward individuals. These factors include: *intentionality*, intending to bring about negative outcomes significantly increases blame; *capacity*, inability to prevent negative outcomes or foresee negative outcomes mitigates blame; *obligation*, the presence or lack of obligation toward achieving/preventing an outcome can mitigate or increase blame; *justification*, having a valid moral justification for a typically blameworthy outcome can mitigate 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 (2012) and Tomai and Forbus (2007) model the effects of coercion (by a superior) on how blame is attributed.

The sole output of these models is a *degree of blame*, which is often an ordinal value [12], [9], [13], rather than a degree to which specific traits or characteristics are ascribed to the agent. This is in part due to the fact that these models were designed to account for the results from human subject data regarding philosophical thought-experiment type scenarios[1]. That is to say, these models were designed to account for how humans responded to Likert scale questions (e.g. How much do you blame Bob? 1 = minimal blame to 9 = maximum blame) after reading about a moral transgression.

When applying this form of model to the scenario described in Section II-A, we can see that a high degree of blame is attributed to Bob in Case 1, as the failure to deliver the package stemmed from a intentional act of theft. Case 2 and 4 would likely result in lower due to the lack of malicious intent. Case 3 might result in an even lower blame attribution, as Bob's lack of capacity to actually achieve the goal could be exculpatory. Finally, Case 5 would have a level of no blame assessed as Bob is not casually responsible for the outcome at all. In some sense, this is consistent with the character inferences made in the previous section. We can imagine that a multi-dimensional character assessment derived when reading about a moral situation is being collapsed into a single-dimension when asked to do so in a one-dimensional blame assessment question. The question, "how much do you blame Bob?", is pragmatically interpreted as, "how pejorative a character inference do you make about Bob?", or "to what extent does this reflect negatively on Bob's character?" Unfortunately, this means that the social planning and interaction adaptation proposed in the prior section will receive insufficiently detailed input. What does "high" or "low" blame tell us about another agent's future actions? These processes require that more multi-dimensional computational models of blame, that model character inferences, need to be developed.

III. SOCIAL FACET OF BLAME

As previously mentioned, blame has a social component, whereby the cognitive assessments of agents' blameworthiness are communicated. Whether or not humans appreciate being blamed by social robots is a matter for future HRI studies. Initial work has indicated that people assess a robot more negatively when it directs more blame at them for a task failure [14]. It is also possible for natural language generation (NLG) architectures to make blame-related statements unintentionally. To illustrate this, let us revisit our Postie the delivery robot example.

A. Scenario

In Section II-A, we described a scenario in which Postie the PR2 is responsible for delivering several packages throughout a large facility. Having determined that it needs human help, Postie needs to formulate a request to Bob (the helpful human). In the previous section, we glossed over the details of the NLG process, where Postie must choose the precise form of the request. These possible request forms include, but are not limited to:

1) "Deliver this package to Building 23."
2) "Can you deliver this package to Building 23?"
3) "It would be great if you could deliver this package to Building 23."
4) "This package has not been delivered to Building 23 yet."

Human speakers choose between these various request forms based on social concerns, such as politeness [15]. For instance, in certain social contexts (e.g. military) directness is favored for practical purposes (and to reinforce the explicit strict organizational hierarchy). As such, request form 1 would be favored in these situations. In more general contexts, more polite speech is favored, including conventionally indirect request forms such as 2 and 3. These request forms are commonly understood to imply requests, but are not literal commands. Conventionally indirect forms are a form of *negative politeness strategy*, which seek to mitigate threats to *negative face*, or the autonomy of an agent's future actions [15]. Finally, unconventionally indirect forms (such as 4) require more inference to understand the implied request, but are considered to be even more polite, as they are less associated with autonomy-limiting requests than conventionalized forms. However, as one may expect from the term *negative face*, there also exists a notion of *positive face*, which deals with the self-image and perceived character of the agent. An utterance that is polite in terms of mitigate threats to negative face may still be perceived as rude or impolite due to its threats toward positive face. This leads to some complications in NLG modulation, which we discuss below.

B. Limitations of Current Approaches

Current approaches to politeness modulation in NLG architectures often simply vary the directness of the request [16], [17]. However, utilizing forms similar to those presented above, Gupta et al. (2007) gathered human subject data on politeness rankings, which showed that contrary to the expectations based on negative politeness, the unconventionally indirect request form (similar to form 4 above) was perceived by human listeners to be even *more* impolite than the direct request. Gupta et al. (2007) commented that their unconventionally indirect forms, such as, "X is not done yet," can be construed as a complaint. Indeed, as pointed out in [18], modeling blame is necessary to detect whether or not certain forms of positive face threats are occurring, and propose some mechanisms to integrate such reasoning into an NLG architecture. Additionally, Briggs and Scheutz (2014) also make the point that not only does reasoning have to be conducted as to whether or not a candidate utterance can be construed as connotative of blame, but reasoning must be conducted as to whether or not such an assessment of blame is valid or appropriate to communicate

[1]Also in part likely due to the appeal of considering the affective component of blame judgments (and its amplitude).

given the current social context. No current NLG architecture modulates generate language based on these considerations. As such, future work is needed to tackle this particular challenge in enabling social robots and other artificially intelligent agents to avoid such social faux pas.

IV. AVOIDING BLAME TOWARD SELF

In the previous sections we have discussed how socially-appropriate behavior adaptation by a social robot can occur based on: blame-reasoning about other agents and reasoning about other agents perceptions of blame directed toward them. We conclude by discussing how prosocial behaviors can be derived by reasoning about how other agents may direct blame towards the robot itself. This form of social reasoning would be akin to processes in human social cognition such as *impression management*. Indeed, recent research in the neurological basis of moral reasoning has found that regions associated with thinking about how others perceive one's self are often activated in cases where people are confronted with difficult and counterintuitive moral dilemmas [19]. Other studies have shown how contextual cues that imply social surveillance drive behaviors toward more prosocial outcomes [20].

Yet, how could this process be achieved in a robotic system? I imagine it could be done using the mechanisms described in Section II-A in conjunction with an ability to reason about *long-term utilities*. That is to say, some mechanism by which to quantify and compare the damage or enhancement to particular relationships and future interactions. To illustrate this, consider the following actions used to represent dialogue moves (similar to those in [21]):

$inform(\alpha, \beta, \phi)$
  precondition: $want(\alpha, bel(\beta, \phi))$
  effect: $bel(\beta, want(\alpha, bel(\beta, \phi)))$

where α denotes the speaker, β the listener, and ϕ a proposition to be conveyed. This action serves to model a declarative speech act, in which the speaker wants to get the listener to believe in that ϕ holds. This results in the listener believing that speaker wants him or her to believe ϕ. However, another action is needed to translate this belief in an interlocutor's intention into actual belief adoption:

$decide\_to\_believe(\alpha, \beta, \phi)$
  precondition: $bel(\alpha, want(\beta, bel(\alpha, \phi)))$
  precondition: $bel(\alpha, is\_honest(\beta))$
  effect: $bel(\alpha, \phi)$

this action serves to model the interaction partner and their internal process of deciding to adopt a suggested belief, where again α denotes the speaker, β the listener, and ϕ a proposition to be conveyed.

These basic actions would comprise building blocks to a considerable number of social plans. For instance, if Postie the PR2 was re-assigned from mundane mail delivery duties to something more bucolic, say, guarding a flock of sheep, we can imagine the following plan may be used at some point:

$[inform(postie, townsfolk, wolf\_present),$
  $decide\_to\_believe(townsfolk, postie, wolf\_present),$
  $goto(townsfolk, field)]$

This plan will succeed, assuming competency on the part of the townsfolk, in proportion to the degree to which the precondition $bel(townsfolk, is\_honest(postie))$ is met. If the belief of Postie's honest character is diminished by reasoning of the sort described in Section II-A, let's say, by repeated false alarms, then the viability of this plan (or any plan dependent on impressions of honesty) is also diminished. The long-term utility consequences, then, of dishonesty can be quantified by the loss of expected rewards from increased social plan failure.

Of course, there are some limitations to this sort of behavior modulation method. Such a pathway to ethical behavior is not a means by which to achieve super-human moral capabilities, but rather human-level capabilities, subject to many of the same faults. For instance, if a potentially unethical behavior is not deemed unethical by the group of interactants the robot is surrounded by, and the robot is only concerned with the perceptions of that group, then it would not be discouraged from that behavior (unless other overriding behavior modulating mechanisms existed). Regardless, we would argue that in the absence of any formalized and well-agreed to standards of behaviors in specific domains, or conflict between competing principles, such impression management is a good fallback option for a social robot.

V. CONCLUSIONS

We have made the case that blame-reasoning mechanisms are necessary to enable future social robots to: (1) appropriately adapt behavior in the context of repeated and/or long-term interactions and relationships with other social agents; (2) avoid behaviors that are perceived to be rude due to inappropriate and unintentional connotations of blame; and (3) avoid behaviors that could damage long-term, working relationships with other social agents. However, these social competencies would require blame-reasoning mechanisms to be integrated throughout the architecture of a robotic system, including components such as the natural language system. Perhaps more importantly, current computational models of blame need to be expanded to account for degree to which certain characteristics can be attributed to an agent, in addition to the degree to which "blame" can be attributed toward an agent. We hope that by illustrating these challenges, that other researchers interested in enabling social robots to act in socially appropriate ways are inspired to work toward their solutions.

REFERENCES

[1] W. Wallach, "Robot minds and human ethics: the need for a comprehensive model of moral decision making," *Ethics of Information Technology*, vol. 12, pp. 243–250, July 2010.

[2] P. Lin, K. Abney, and G. A. Bekey, *Robot ethics: the ethical and social implications of robotics*. MIT Press, 2011.

[3] R. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," Georgia Institute of Technology, Tech. Rep. GIT-GVU-07-11, 2009.

[4] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a general logicist methodology for engineering ethically correct robots," *IEEE Intelligent Systems*, vol. 21, no. 5, pp. 38–44, July/Aug. 2006.

[5] M. Guarini, "Particularism and the classification and reclassification of moral cases," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 22–28, July/August 2006.

[6] J. Shim and R. C. Arkin, "A taxonomy of robot deception and its benefits in hri," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2328–2335.

[7] G. Briggs, "Machine ethics, the frame problem, and theory of mind," in *Proceedings of the Symposium on Moral Cognition and Theory of Mind at AISB/IACAP*, 2012.

[8] P. Bello and S. Bringsjord, "On how to build a moral machine," *Topoi*, vol. 32, no. 2, pp. 251–266, 2013.

[9] B. F. Malle, S. Guglielmo, and A. E. Monroe, "Moral, cognitive, and social: The nature of blame," *Social thinking and interpersonal behavior*, vol. 14, p. 313, 2012.

[10] D. Hume, *A treatise of human nature*. Courier Dover Publications, 2012.

[11] G. Sher, "In praise of blame," 2007.

[12] W. Mao and J. Gratch, "Modeling social causality and responsibility judgment in multi-agent interactions," *Journal of Artificial Intelligence Research*, vol. 44, no. 1, pp. 223–273, 2012.

[13] E. Tomai and K. Forbus, "Plenty of blame to go around: a qualitative approach to attribution of moral responsibility," DTIC Document, Tech. Rep., 2007.

[14] V. Groom, J. Chen, T. Johnson, F. A. Kara, and C. Nass, "Critic, compatriot, or chump?: Responses to robot blame attribution," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 2010, pp. 211–217.

[15] P. Brown, *Politeness: Some universals in language usage*. Cambridge University Press, 1987, vol. 4.

[16] S. Gupta, M. A. Walker, and D. M. Romano, "How rude are you?: Evaluating politeness and affect in interaction," in *Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 203–217.

[17] G. Briggs and M. Scheutz, "A hybrid architectural approach to understanding and appropriately generating indirect speech acts," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

[18] ——, "Modeling blame to avoid positive face threats in natural language generation," in *Proceedings of the 8th International Natural Language Generation Conference*, 2014, forthcoming.

[19] G. Kahane, K. Wiech, N. Shackel, M. Farias, J. Savulescu, and I. Tracey, "The neural basis of intuitive and counterintuitive moral judgment," *Social cognitive and affective neuroscience*, vol. 7, no. 4, pp. 393–402, 2012.

[20] M. Bateson, D. Nettle, and G. Roberts, "Cues of being watched enhance cooperation in a real-world setting," *Biology letters*, vol. 2, no. 3, pp. 412–414, 2006.

[21] C. R. Perrault and J. F. Allen, "A plan-based analysis of indirect speech acts," *Computational Linguistics*, vol. 6, no. 3-4, pp. 167–182, 1980.