Leveraging Intentional Factors and Task Context to Predict Linguistic Norm Adherence

Cailyn Smith[1] **(ccsmith1@mines.edu)**
Ruchen Wen (rwen@mines.edu)
Saad Elbeleidy (selbeleidy@mines.edu)
Sayanti Roy (sayantiroy@mines.edu)
Tom Williams (twilliams@mines.edu)
Department of Computer Science
Colorado School of Mines
Golden, CO 80401 USA

Charlotte Gorgemans[1]
Department of Computer Science
University of Colorado Boulder
Boulder, CO 80309 USA

Abstract

To enable natural and fluid human-robot interactions, robots need to not only be able to communicate with humans through natural language, but also do so in a way that complies with the norms of human interaction, such as politeness norms. Doing so is particularly challenging, however, in part due to the sensitivity of such norms to a host of different contextual and intentional factors. In this work, we explore computational models of context-sensitive human politeness norms, using explainable machine learning models to demonstrate the value of both speaker intention and task context in predicting adherence with indirect speech norms. We argue that this type of model, if integrated into a robot cognitive architecture, could be highly successful at enabling robots to predict when they themselves should similarly adhere to these norms.

Keywords: Politeness; Linguistic Norms; Human-Robot Interaction

Introduction

Social robots stand to advance human capabilities and well-being across a wide span of domains like education (Mubin, Stevens, Shahid, Al Mahmud, & Dong, 2013; Belpaeme, Kennedy, Ramachandran, Scassellati, & Tanaka, 2018) and healthcare (Broekens, Heerink, Rosendal, et al., 2009; Breazeal, 2011; Cifuentes, Pinto, Céspedes, & Múnera, 2020). For social robots to be successfully integrated into the society (especially those designed as *sociable partners* (Breazeal, 2004)), they are expected to behave in accordance with human social norms (Bartneck & Forlizzi, 2004); failure to do so can risk interaction *breakdowns* (Porfirio, Sauppé, Albarghouthi, & Mutlu, 2018; Mutlu & Forlizzi, 2008). Social norms not only govern how people behave, but also how people communicate. To engage in natural and fluid human-robot interactions, robots must thus not only communicate with humans through language (cf. (Tellex, Gopalan, Kress-Gazit, & Matuszek, 2020)), but do so in a way that complies with social norms. One of the key categories of social norms that require such adherence by social robots are *politeness norms* (Lee, Kim, Kim, & Kwon, 2017).

According to Brown and Levinson's Politeness Theory (Brown, Levinson, & Levinson, 1987), human interactants regularly negotiate the level of threat to one another's *Face*: the public image that the other person wants to maintain and enhance (Brown et al., 1987). Face consist of two aspects: Positive Face (i.e., one's want for a desirable self-image) and Negative Face (i.e., one's desire to be free from imposition and to have freedom of action) (Brown et al., 1987). To comply with politeness norms and mitigate the face threat behind potentially face threatening acts, people employ a variety of politeness strategies (Goffman, 1955). For example, when ordering food in a restaurant, instead of saying "Get me some coffee", people typically phrase their requests in a more indirect manner, such as "I would like some coffee". While this utterance is literally a statement of fact, listeners in this context can easily understand the true intention behind the utterance, which is a request for some coffee. This type of utterance, in which the utterance's literal meaning does not match its intended meaning, is called an Indirect Speech Act (ISA) (Searle, 1975). Indirect language can be particularly effective for reducing face threat by obscuring threats to autonomy. Accordingly, ISAs are one of the most effective and commonly used linguistic politeness strategies.

Robots capable of applying linguistic politeness strategies are perceived as more likeable, considerate and engaging (Castro-González et al., 2016; Torrey, Fussell, & Kiesler, 2013). Yet ISA use is highly context-sensitive. Williams, Thames, Novakoff, and Scheutz (2018a) showed that Americans tend to use more ISAs in contexts with strong social conventions (which come along with strong sociocultural norms and contracts) such as restaurants (cf. (Seok, Hwang, Choi, & Lim, 2022)). Moreover, ISA use depends on nuanced dimensions of the context in which an interaction occurs, which may determine whether it is appropriate to use indirect language. While indirect language can effectively decrease face threat through mechanisms such as hedging, it also has other effects that may potentially *increase* the threat to one's autonomy, (e.g. longer utterances with many hedges necessarily impose on the listener's time, and thus, on their autonomy). For example, in the case of search and rescue, using ISAs could be ineffective as a politeness strategy given the time pressure and potential for harm in those domains. Similarly, researchers have shown that politeness strategies, including ISAs, lead to *less* compliance with robots in healthcare domains, perhaps for the same reasons (Lee et al., 2017).

Adherence to linguistic politeness norms is also important due to their connection with Grice's conversational maxims. Grice stipulates that humans assume that cooperative interactants will generally strive to "Make [their] conversational

[1] Authors Smith and Gorgemans contributed equally to this work.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."(Grice, 1975). However, the way in which interactants do so is again highly context-dependent. Indirect speech acts avoid violating sociocultural politeness norms, but in doing so simultaneously violate Grice's Maxim of Manner ("Be perspicuous"). Accordingly, speakers need to be sensitive to the contextually sensitive tradeoffs between these conversational requirements. As above, individuals may differentially weight these considerations when solving time-sensitive issues than when they are engaging in routine tasks. Humans demonstrate exactly this type of context sensitivity when arbitrating between these tradeoffs, both in human-human interaction (Agha, 2006) and human-robot interaction (Williams, Thames, Novakoff, & Scheutz, 2018b). Moreover, robots that are sensitive to changing social and conversational contexts are viewed more favorably (Ritschel, Baur, & André, 2017; Jackson, Wen, & Williams, 2019). Thus, it is crucial for robots to not only be able to use ISAs, but moreover to be able to intelligently decide whether to use ISAs based on interaction context.

Lockshin and Williams (2020) previously examined the impact of three key contextual factors (potential for harm, interlocutor authority and time pressure) on people's use of ISAs in human-human interaction, and found that ISA use varies based on these more nuanced contextual dimensions. However, task context alone is not sufficient to capture everything behind why people use ISAs. In fact, we argue that these are unlikely to be even the *primary* dimensions that dictate ISA use. Instead, the decision to use an ISA is more likely grounded primarily in the the information that a person is trying to convey. Different types of utterances and different intentions fundamentally have different levels of face threat, and thus the appropriateness of ISA use should significantly vary based on these intentional factors. For instance, in cases where a person wants to acknowledge that they have received information ("I understand"), using direct language may be more appropriate because self-directed acknowledgements are often likely to have low levels of face threats; while in cases where a speaker shares information about another person ("You are going to do this"), using indirect language may be more appropriate because statements about what a listener *will do* inherently constricts the listener's autonomy (imposing constraints on the listener's time and future actions). As such, these intentional dimensions of interactions need to be analyzed alongside the contextual domains previously investigated by Lockshin and Williams (2020).

In this work, we thus investigate whether contextual *and intentional* factors enable more effective prediction of ISA use than does the use of contextual factors alone. To perform this investigation, we use the same public dataset previously used and provided by Lockshin and Williams (2020). In the following section we will explain our conceptualizations of contextual and intentional factors. We will then describe how we augmented Lockshin and Williams (2020)'s dataset to newly account for intentional factors. Next, we will describe how we trained an explainable machine learning model on this dataset to better predict linguistic norm adherence. Finally, we will evaluate the performance of that model and investigate the claims made by that trained model.

Contextual and Intentional Factors

In this section we define the *contextual* and *intentional* factors we use in this work to predict linguistic norm adherence.

Contextual Factors

Lockshin and Williams (2020) examined the ability of three key contextual factors to predict linguistic norm adherence: *potential for harm*, *interlocutor authority*, and *time pressure*. These contextual factors were chosen by considering contexts where these norms might or might not be followed (especially context in which a robot might reasonably be involved as human teammates), and hypothesizing that those factors represented key dimensions of variance between those contexts.

As defined in Lockshin and Williams (2020), *potential for harm* is present when there is a high likelihood for negative outcomes to occur if actors within a situation do not respond correctly. *Interlocutor authority* is present when a speaker possesses authority over the hearer. *Time pressure* is present when there is a limited amount of time to complete a task.

Intentions

As described above, the key aim of this work is to expand on these types of contextual factors to additionally consider the speaker's *intentionality*, under the hypothesis that there is a strong correlation between the intended purpose of an utterance and whether it is phrased directly or indirectly. Because of the nuanced and complex nature of intent, we decided to use a multidimensional model of intent, thus enabling us to determine not only *whether* intent is helpful for predicting ISA use, but moreover, what aspects of intent are most informative for making such predictions.

We conceptualized speaker intent in a way that broke intent into three key dimensions: *direction*, *target*, and *force*.

Direction: We first considered whether an utterance *requests* or *provides* information. For example, *"Where is the salt?"* requests information, whereas *"The salt is in the cabinet"* provides information. Critically, an utterance like *"Can you pass the salt?"* is framed as a request, but is not typically a true request for information, and more likely provides information about (a desired) future action of the listener.

Target: We next considered whether an utterance is requesting or providing information about an action of the *speaker* or the *hearer*. For example, *"I'm going to get the salt"* could be viewed as providing information about the speaker's own future actions, whereas *"You should pass me the salt"* is providing information about the future actions the speaker desires the *hearer* to perform.

Force: Finally, we considered the action the utterance is focusing on, from a dialogic or illocutionary perspective. For example, *"You should get the salt"* is directly concerned with

the action being conveyed by the utterance. In contrast, *"Why are you getting the salt?"* asks for explanation about a previously communicated course of action; *"Which salt are you going to get?"* asks for clarification about the action; and *"Alright?"* (after *"I'm going to get the salt"* asks for acknowledgement about a previously communicated course of action. We thus delineate between four types of forces we are concerned with: *action*-centered, *explanation*-centered, *clarification*-centered, or *acknowledgement*-centered.

Data Context

We will now describe the dataset used in this work. We used the public dataset collected by Lockshin and Williams (2020); a corpus of transcripts of participants playing the board game *Pandemic*. Lockshin and Williams (2020) used this board game to collect their dataset because it allowed for the systematic control of the three contextual factors described above, and because *Pandemic* is a highly cooperative game which promotes teamwork and team communication. Using this game, Lockshin and Williams (2020) collected six game transcripts, each from a game with a new trio of three players, ranging from 526 to 1200 dialogue moves each.

After collecting this dataset, Lockshin and Williams (2020) annotated each utterance in their dataset with binary feature values indicating whether the game state at the time of that utterance was one in which there was potential for harm, interlocutor authority, and/or time pressure.

Potential for harm was considered to be present if players were close to losing the game. Lockshin and Williams introduce an equation for measuring this within the specific context of *Pandemic*. Interlocutor authority was considered to be present if it was currently the speaker's turn. In *Pandemic*, speakers take turns, but on each player's turn, the whole group debates and suggests to that player what they should choose to do. That is, the player whose turn it is has the final decision on how they use their turn but can receive advice from other players. Turns rotate between players, so interlocutor authority shifts to a different player each turn. Time pressure was considered to be present if players had a limited amount of time to make decisions about how to use their turn. While *Pandemic* does not traditionally include this type of time pressure, Lockshin and Williams ran 50% of their games using a variant ruleset in which time limits on turns were introduced. Out of the six games of *Pandemic* played, three were randomly assigned to this variant ruleset condition. In the games with time pressure, players had 90 seconds to decide their moves each turn, whereas players had unlimited time to decide their moves in the games without time pressure.

Data Annotation

We will now describe how we augmented Lockshin and Williams (2020)'s dataset to include intentional factors. After revising Lockshin and Williams (2020)'s original labels for improved dataset quality and splitting utterances into distinct dialogue moves, four coders coded the dataset for each intentional factor described above. Two coders annotated each datapoint, and if their codes disagreed, all coders discussed the datapoint to collectively select an appropriate code. These disagreements occurred for 5.17% of all annotations: 3.75% for utterance target, 2.31% for utterance direction, and 9.46% for illocutionary force. In (extremely rare) cases where agreement could not be reached with four coders, two supervisors provided further comment, and a final vote was taken.

Before moving on, we will provide examples of utterances from Lockshin and Williams (2020)'s dataset that were coded in each of the categories delineated above.

Direction: Examples of utterances coded as *providing* information include *"It's a choice"* and *"I need one more red to cure it"*. Examples of utterances coded as *requesting* information include *"I could go to LA and get one of those, right?"* and *"How do you go there?"*

Target: Examples of utterances coded as *speaker*-targeted include *"I didn't think about that"* and *"And now I have to do this other infection thing, right?"*. Examples of utterances coded as *other*-targeted include *"Now you draw two cards"* and *"You have to get rid of a card"*.

Force: Examples of utterances coded as *action*-centered include *"Where is the thing about research stations?"* and *"The yellows are more in danger of outbreaks"*. Examples of utterances coded as *explanation*-centered include *"But you have to discard a card if you want to move like far"* and *Because then you go from there to there and you are adjacent*. Examples of utterances coded as *clarification*-centered include *"And since we are in the same city right now maybe I should give you like a yellow and a red or just a yellow?"* and *"Either way, yeah"*. Examples of utterances coded as *acknowledgement*-centered include *"Okay"* and *"Alright"*.

Technical Approach

Now that we have described our dataset selection and augmentation, we can now describe our technical approach to modeling linguistic norm adherence, in which we trained a decision tree (Breiman, Friedman, Stone, & Olshen, 1984) on the annotated dataset described in the *Data Context* Section. We trained this model to predict whether direct or indirect language would be used in a given context based on the contextual and intentional factors described above.

We used decision trees due to their ease of interpretation. This was especially important since we were not only interested in developing a highly effective predictive model, but also in developing an understanding of the underlying rationale that humans may follow when deciding whether or not to speak directly. Decision trees have been highly successful in past research similarly interested in transparency and explainability due to the readily interpretable nature of the flowcharts used to represent their models (Delen, Kuzey, & Uyar, 2013; Namazkhan, Albers, & Steg, 2020). Decision trees have even been used as surrogate models to explain more complex black-box models (Shi, Zhang, & Fan, 2019; Kuttichira, Gupta, Li, Rana, & Venkatesh, 2019).

A decision tree can be represented as a flowchart, where

each node represents an intermediate binary decision on the way towards classification, centered on a single variable of interest. The test at each node checks whether the value of a feature meets some condition; either less than or greater than a particular threshold, in the case of numerical values or, more relevant to our use case, whether the feature's value is a particular choice from the set of possible values, in the case of nominal or categorical variables. Following a path through a decision tree based on a particular sample's feature values leads to a leaf node designating a final outcome (in our case, a classification of the utterance as direct or indirect). Such a diagram for our best performing model is shown in Fig. 1.

When training decision trees, the training algorithm identifies binary decisions' splitting criteria that minimizes the impurity of the sets that result from splitting samples. The impurity of a set of points at a given node, N can be measured using Gini impurity ($G(N) = \sum_k p_{k,N}(1 - p_{k,N})$), where $p_{k,N}$ is the proportion of samples labeled as class k found in node N. As an example, a pure set in which all samples belong to a single class would have a Gini impurity of 0 since only a single class would have a proportion of 1 and the remaining classes would have a proportion of 0, resulting in a sum of 0. To determine the resulting tree that minimizes the impurity of its nodes, we used the optimized variation of the CART algorithm (Breiman et al., 1984) provided by the scikit-learn Python library (Pedregosa et al., 2011).

Training a decision tree can depend on several hyperparameters; impurity metric, maximum tree depth, minimum leaf samples, minimum sample split, minimum impurity decrease, and class weighting. While we described the Gini impurity metric as the impurity metric to optimize, node impurity can also be measured using metrics like set entropy. Maximum tree depth is a stopping criterion for the tree; a maximum number of consecutive decision points before reaching a final decision. Minimum leaf samples is the minimum number of samples within a leaf node to warrant a decision be made. Minimum sample split is the minimum number of samples needed to warrant a decision point or "split". Minimum impurity decrease is the the amount the impurity of a node must decrease for a new split to occur in which a decision must be made. Maximum tree depth, minimum leaf samples, minimum sample split, and minimum impurity decrease are all hyperparameters that are used to limit overfitting. Class weighting is the relative error weighting of particular classes and is useful for imbalanced datasets.

To train the model, we used an 80/20 stratified train-test split of our 2208 utterances, in which a randomly selected 80% of the data was used for training, with the remaining 20% of the data used for testing. Because there was also (coincidentally) an approximately 80/20 split in our dataset between direct and indirect utterances, this approach led to training and testing sets that each contained approximately 81% direct utterances, and 19% indirect utterances. We performed 5-fold cross validation to tune the model's various hyperparameters and determine the best performing model, by splitting our training set into 5 equal subsets of samples, then training a new model using each combination of 4 subsets for training and the remaining hold-out set for validation.

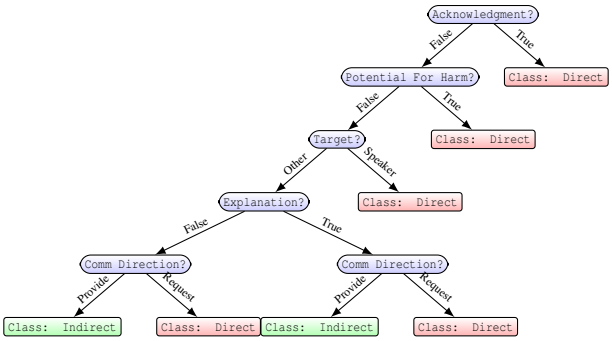

Figure 1: Flow chart visualization of the best DT model.

To determine the best performing model, we tuned all hyperparameters described above. We performed a hyperparameter search over a range of possible hyperparameter values, and identified the values that maximized mean macro-averaged F1-score across our 5-fold cross validation sets.

Analysis and Results

To evaluate our approach and compare to Lockshin and Williams, we trained and compared three models: one trained using both the contextual features used by Lockshin and Williams and the intentional features introduced in this work, one trained using only contextual features, and one trained using only intentional features. In this section, we will step through these models and their comparison, the results of which are summarized in Tab. 1 and visualized in Fig. 2.

Our best performing decision tree model achieved an accuracy of 67% and a macro F1 score of 0.60. Our best performing model used Gini impurity as the impurity metric, had a maximum tree depth of 5, a minimum impurity decrease of 0.002, minimum leaf samples of 1, minimum sample split of 2, and a class weighting of 3:1 (indirect:direct). The best performing model's class weighting matches our expectations due to our dataset's class imbalance. Our best performing model, shown in Fig. 1, has seven leaf nodes, of which five are classified as direct and two of which would be classified as indirect. This model can be simplified by combining the two deepest subtrees, which are only rendered distinctly due to the optimal choice of maximum tree depth. This would produce a tree with five leaf nodes (four direct, one indirect).

To assess the benefits of including intentional factors, we also evaluated a decision tree trained with *only* contextual factors, to assess what results Lockshin and Williams would have seen if they had used a similar Decision Tree theoretic modeling paradigm[2], if they had performed parameter tuning, considered class weighting to handle minority class imbalance, or

[2]We also informally evaluated a Naive Bayes approach (including parameter tuning), which produced notably worse results than the Decision Tree approach.

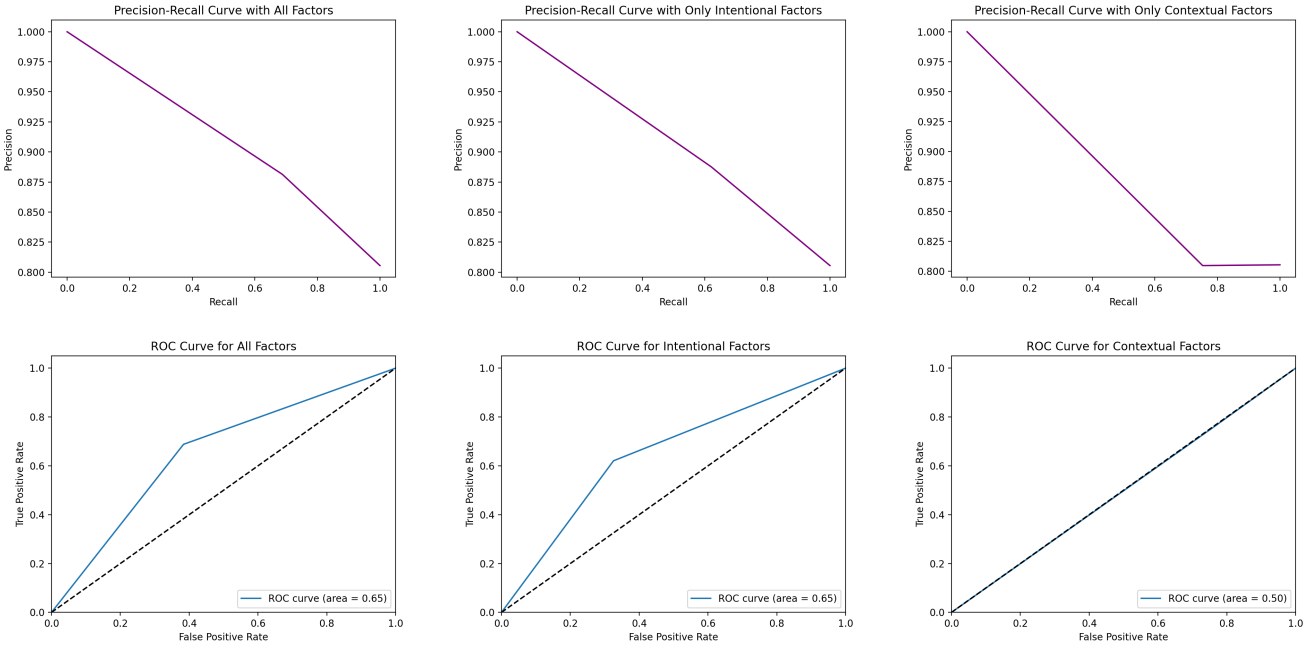

Figure 2: **Evaluation Results:** The top row of figures show (Top) Precision-Recall curves and (Bottom) ROC curves for (Left) All factors, (Center) Intentional factors alone, and (Right) Contextual factors alone.

Features		**Evaluation Metrics**						
Intentional	**Contextual**	**Macro F1**	**Accuracy**	**Prec. (Ind.)**	**Prec. (Dir.)**	**Rec. (Ind.)**	**Rec. (Dir.)**	**F1**
✓	✓	**0.60**	**0.67**	**0.32**	0.88	0.62	0.69	0.77
	✓	0.50	0.65	0.19	0.80	0.24	**0.75**	**0.78**
✓		0.57	0.63	0.30	**0.89**	**0.67**	0.62	0.73

Table 1: Evaluation and model comparison of decision trees with various features.

used the directness/indirectness labels produced in this work.

The best performing decision tree model with only contextual factors achieved an accuracy of 65% and a macro F1 score of 0.50. This model used Gini as the impurity metric, had a maximum tree depth of 3, a minimum impurity decrease of 0, minimum leaf samples of 1, minimum sample split of 2, and a class weighting of 4:1 for indirect to direct. This decision tree model had six leaf nodes (four direct, two indirect).

To compare model performance, we used a McNemar's test. This test did not reveal statistically significant differences in predictions (p=0.61) between the best-performing full model and the model with only contextual features used.

We also evaluated a decision tree trained with *only* intentional factors to determine the importance of using the contextual factors from Lockshin and Williams (2020). The best performing decision tree model with only intentional factors achieved an accuracy of 63% and a macro F1 score of 0.57. This model used Gini as the impurity metric, had a maximum tree depth of 4, a minimum impurity decrease of 0.0015, minimum leaf samples of 1, minimum sample split of 2, and a class weighting of 3:1 for indirect to direct. This model had six leaf nodes (four direct, two indirect).

Discussion

To begin, we can first consider the general success of our modeling approach relative to that used by Lockshin and Williams (2020). Lockshin and Williams fit their collected data using simple logistic regression models, and used frequentist hypothesis testing to assess the influence of each contextual factor on linguistic norm adherence. As such, they were operating according to a very different methodological philosophy than we do in this work, with no cross-validation, hyperparameter tuning, class reweighting, accuracy calculation, and so forth. Nevertheless, there are some ways in which our approaches can be directly compared. The most straightforward demonstration of the utility of our approach is through examination of the qualitative differences in predictions made by our approaches. While Lockshin and Williams did show statistically significant differences evidencing the importance of potential for harm and time pressure for modeling ISA use, their models uniformly recommended ISA use regardless of context. That is, their models demonstrated that ISA use differs between contexts, but nevertheless maintained

ISAs as the most likely prediction across all contexts, and would thus never actually recommend *not* using an ISA. In contrast, as described above, the decision tree produced by our model includes leaves recommending ISA use *and* leaves not recommending ISA use, and in fact the majority of leaves within this tree do not recommend ISA use. As such, our approach novelly produces a model that could actually be used to intelligently decide whether to use ISAs in a given context.

Another benefit of our approach is the explainability and transparency of Decision Trees, which encode models in a form that can be readily interpreted by humans to make psychological claims, and that can be easily encoded into autonomous agents to control their behavior, without needing model code or needing to know how the model was trained. Specifically, the decision tree produced by our approach suggests that speakers tend to phrase their utterances directly if any of the following are true (and phrase their utterances indirectly in all other cases), in decreasing order of importance.

1. the utterance is an acknowledgment;
2. there is potential for harm;
3. the utterance is directed at oneself (vs another person);
4. the utterance requests (rather than provides) information.

These findings validate our approach *from a purely qualitative perspective*. Our original motivation was the realization that utterances like acknowledgements (or requests for acknowledgement) would almost certainly be phrased directly, and should be a clear case where a model would not recommend indirectness. This intuition is borne out as the single most salient feature in the trained decision tree. We also posited that we would need a nuanced representation of intentionality, that included things beyond utterance force; a hypothesis borne out in the third and fourth criteria of the decision tree. Moreover, our approach demonstrates the importance of considering both contextual factors and intentional factors. As shown in Tab. 1, using both intentional and contextual factors facilitates the best Accuracy and Macro F1 score, whereas using only contextual or intentional factors produced similar but ultimately poorer results due to substantially poorer precision and/or recall. However, as we have mentioned, there were no statistically significant differences found between these approaches.

Our approach also demonstrates the limits of our claims in other ways. The only category of utterance force represented in the tree is the most obvious one (other than the consideration of whether or not the utterance is phrased as an explanation, which does appear in the tree, but not in a way that makes any classification difference given the optimal maximum tree depth selected during hyperparameter tuning). This approach also shows that once intentionality is accounted for, some of the contextual factors previously argued to be important no longer need be considered, with potential for harm the only contextual factor actually used in the decision tree.

Moreover, many of the discussed benefits of our approach are largely arguments in favor of our general machine learning approach rather intentionality. When comparing the features used in this work to those used by Lockshin and Williams – *within the context of this machine learning approach* – the performance differences between feature sets is relatively minor. As we have described, the best decision tree model did indeed make significant use of intentional factors. Yet the benefit of including intentional factors was quite small, and there was no significant difference between models that did or did not use these features. While we are setting a high standard for ourselves here (as little machine learning research does this type of hypothesis comparison), it nevertheless suggests that we should not make overly strong claims about the importance of intentionality.

Finally, while this work is focused on understanding *how* people adhere to different politeness norms in intention- and context-sensitive ways, if our goal is to use these insights to design robots, we must also acknowledge that we may not always wish robots to adhere to politeness norms in the ways that humans do. It may be necessary for robots to intentionally *violate* social norms (Yasuda, Doheny, Salomons, Sebo, & Scassellati, 2020), including sociocultural politeness norms (Briggs, Williams, Jackson, & Scheutz, 2022), either to issue blame-laden moral rebukes (Zhu, Williams, Jackson, & Wen, 2020) and/or to avoid reinforcing sexist attitudes (Jackson, Williams, & Smith, 2020; Winkle, Melsión, McMillan, & Leite, 2021; Winkle et al., 2022). Understanding when and how norms should be intentionally violated in non-humanlike ways is a key open research area.

Conclusion

We investigated whether contextual *and intentional* factors enable more effective prediction of Indirect Speech Act use than does the use of contextual factors alone, with the goal of enhancing the social intelligence of interactive agents like social robots. To do so, we developed a framework for analyzing speaker intent, augmented an existing public dataset, and deployed a highly interpretable machine learning approach. Our results demonstrate the benefits of our machine learning approach and the utility of using both intentional and contextual factors when predicting linguistic norm adherence, yet also suggest the performance gain obtained by using both types of features may be negligible.

This work motivates a number of possible future research directions. First, the produced models could be deployed into robot cognitive architectures to determine the extent to which their use encourages positive human perceptions or facilitates more effective human-robot interactions. Second, researchers should explore the incorporation of additional types of contextual factors, which might both increase performance and further demonstrate the utility of the features already used in this work. Finally, researchers should explore the role that individual differences and variation play, which might account for a large proportion of the observed variance in ISA use.

Acknowledgments

This work was funded in part by Air Force Young Investigator Award 19RT0497. We would like to thank Nicole Chen and Huarui Lui for their assistance in early phases of this work.

References

Agha, A. (2006). *Language and social relations* (Vol. 24). Cambridge University Press.

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *Ro-man 2004. 13th ieee international workshop on robot and human interactive communication (ieee catalog no. 04th8759)* (pp. 591–594).

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science robotics*, *3*(21), eaat5954.

Breazeal, C. (2004). *Designing sociable robots*. MIT press.

Breazeal, C. (2011). Social robots for health applications. In *2011 annual international conference of the ieee engineering in medicine and biology society* (pp. 5368–5371).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Briggs, G., Williams, T., Jackson, R. B., & Scheutz, M. (2022). Why and how robots should say 'no'. *International Journal of Social Robotics*, *14*(2), 323–339.

Broekens, J., Heerink, M., Rosendal, H., et al. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, *8*(2), 94–103.

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Castro-González, Á., Castillo, J. C., Alonso-Martín, F., Olortegui-Ortega, O. V., González-Pacheco, V., Malfaz, M., & Salichs, M. A. (2016). The effects of an impolite vs. a polite robot playing rock-paper-scissors. In *International conference on social robotics* (pp. 306–316).

Cifuentes, C. A., Pinto, M. J., Céspedes, N., & Múnera, M. (2020). Social robots in therapy and care. *Current Robotics Reports*, *1*(3), 59–74.

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert systems with applications*, *40*(10), 3970–3983.

Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, *18*(3), 213–231.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Jackson, R. B., Wen, R., & Williams, T. (2019). Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In *Proceedings of the aaai/acm conference on artificial intelligence, ethics, and society.*

Jackson, R. B., Williams, T., & Smith, N. (2020). Exploring the role of gender in perceptions of robotic noncompliance. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (pp. 559–567).

Kuttichira, D. P., Gupta, S., Li, C., Rana, S., & Venkatesh, S. (2019). Explaining black-box models using interpretable surrogates. In *Pacific rim international conference on artificial intelligence* (pp. 3–15).

Lee, N., Kim, J., Kim, E., & Kwon, O. (2017). The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics*, *9*(5), 727–743.

Lockshin, J., & Williams, T. (2020). "we need to start thinking ahead": The impact of social context on linguistic norm adherence. In *Annual meeting of the cognitive science society.*

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., & Dong, J.-J. (2013). A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, *1*(209-0015), 13.

Mutlu, B., & Forlizzi, J. (2008). Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *2008 3rd acm/ieee international conference on human-robot interaction (hri)* (pp. 287–294).

Namazkhan, M., Albers, C., & Steg, L. (2020). A decision tree method for explaining household gas consumption: The role of building characteristics, socio-demographic variables, psychological factors and household behaviour. *Renewable and Sustainable Energy Reviews*, *119*, 109542.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Porfirio, D., Sauppé, A., Albarghouthi, A., & Mutlu, B. (2018). Authoring and verifying human-robot interactions. In *Proceedings of the 31st annual acm symposium on user interface software and technology* (pp. 75–86).

Ritschel, H., Baur, T., & André, E. (2017). Adapting a robot's linguistic style based on socially-aware reinforcement learning. In *2017 26th ieee international symposium on robot and human interactive communication (ro-man)* (pp. 378–384).

Searle, J. R. (1975). Indirect speech acts. In *Speech acts* (pp. 59–82). Brill.

Seok, S., Hwang, E., Choi, J., & Lim, Y. (2022). Cultural differences in indirect speech act use and politeness in human-robot interaction. In *Proceedings of the 2022 acm/ieee international conference on human-robot interaction* (pp. 470–477).

Shi, S., Zhang, X., & Fan, W. (2019). Explaining the predictions of any image classifier via decision trees. *arXiv preprint arXiv:1911.01058*.

Tellex, S., Gopalan, N., Kress-Gazit, H., & Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, *3*, 25–55.

Torrey, C., Fussell, S. R., & Kiesler, S. (2013). How a robot should give advice. In *2013 8th acm/ieee international conference on human-robot interaction (hri)* (pp. 275–282).

Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (2018a). " thank you for sharing that interesting fact!" effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (pp. 298–306).

Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (2018b). "thank you for sharing that interesting fact!": Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 13th acm/ieee international conference on human-robot interaction.*

Winkle, K., Jackson, R. B., Melsión, G. I., Brščić, D., Leite, I., & Williams, T. (2022). Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In *Proceedings of the 2022 acm/ieee international conference on human-robot interaction* (pp. 120–129).

Winkle, K., Melsión, G. I., McMillan, D., & Leite, I. (2021). Boosting robot credibility and challenging gender norms in responding to abusive behaviour: a case for feminist robots. In *Companion of the 2021 acm/ieee international conference on human-robot interaction* (pp. 29–37).

Yasuda, S., Doheny, D., Salomons, N., Sebo, S. S., & Scassellati, B. (2020). Perceived agency of a social norm violating robot. In *Proceedings of the annual meeting of the cognitive science society.*

Zhu, Q., Williams, T., Jackson, B., & Wen, R. (2020). Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective. *Science and Engineering Ethics*, *26*(5), 2511–2526.