Lyra: Simulating Believable Opinionated Virtual Characters

Sasha Azad and Chris Martens
Computer Science Department
North Carolina State University
Raleigh, NC 27601
sasha.azad@ncsu.edu, martens@csc.ncsu.edu

Abstract

Creating believable simulations of large populations of characters in virtual worlds represents a grand challenge for interactive artificial intelligence, requiring reasoning about social intelligence. In this paper, we focus on one aspect of this challenge: the dynamics of *opinion change* for virtual characters and its relationship with social affinity. We developed a simulated population of characters that debate politically-charged topics, called Lyra. Characters' knowledge, opinions, and biases spread through this society based on existing cognitive models and social science theories. Our simulation generates outlines of group conversations that portray the system's evolution, and clusters characters into affinity groups based on the outcome of the debates. We conducted a human-subjects study to evaluate these generated conversations and affinity groups for their believability and to inform future iterations of the simulation. We believe successful simulation of opinion change in social dynamics provides a foundation for computational recognition, prediction, and interfacing with humans.

Introduction

Riedl (2016) describes machine enculturation as the act of instilling social norms, values and etiquette into computers so that they more readily relate to us, and avoid harming us. When instilling these norms into virtual characters by applying artificial intelligence, *social intelligence* is a critical form of reasoning. Wang et al. (2007) discuss how the move to *social intelligence* can be achieved by modeling and analyzing social behaviour, by capturing human social dynamics and creating artificial social agents that generate and manage actionable social knowledge.

Models to simulate such social intelligence have been used in the past to create social training environments (Morrison and Martens 2018; Fowler and Pusch 2010). In digital games with large populations of autonomous non-player characters (NPCs), players have been determined to find interactions between characters more believable if they adhere to recognizable social practices and plausible enculturated (Riedl and Harrison 2016) responses to social situations (Warpefelt 2016).

One key part of social interaction is the dynamics of opinion change and its cause and effect relationship with social relationships. This form of interaction among humans has recently captured the interest of the public with our increasing understanding of the feedback loops created by social networks and political influence (Brichacek 2016). While one approach to study this phenomenon could be to analyze data generated by real user interactions on social networks, we posit that modeling and simulation based on cognitive and social theories can produce good explanatory results of the mechanisms at play during the sharing and swaying of opinions. Correspondingly, we argue that the simulation of opinion change and the causes and effects of bias will positively affect the believability of virtual characters.

This project investigates how to believably simulate the spread of political ideologies and biases through a virtual population and how to present the effects of this simulation in a legible way to human users. We present *Lyra*, a simulation of a virtual town of characters that have varying degrees of political affiliations and ideologies modeled on the US political system. Through a series of interactions with one another, the characters engage in conversations about current news articles on the topics of gun control and immigration. Characters attempt to sway one another towards their own individual dispositions, they learn what topics of discussion are considered sensitive, or could add to growing antagonism or acceptance for themselves and their views among their fellow conversationalists.

We evaluate the believability of the simulation's depiction of the change in the characters' opinions with a human-subjects study deployed online. Our study has two sections, the first summative, evaluating the conversations and the virtual conversationalists themselves; the second formative, evaluating how such conflicts in opinions could affect future relationships and interactions the characters conduct. We evaluated the simulated conversations and discovered they had a mean believability rating. Additionally, the human participants in the study were found to ascribe humanity to the actions of the virtual characters, describing agents that seemed to them to be "competitive" or that felt "marginalized", or discussing how "persuasive" characters seemed to be. We believe that these results support our hypothesis that

Lyra can produce believable social conversation simulations with good explanatory results of the social mechanisms at play. Our work represents a step towards a better understanding of the mechanisms behind social influence and opinion dynamics, enabling more robust social intelligence and more believable social simulations.

Related Work

In this section, we first describe related work from the narrative domain on believable virtual characters. Next, we discuss social simulation from the perspective of narrative intelligence, social science, and psychology to understand how believable virtual characters could be modeled to respond to societal and group archetypes and opinions.

Believable Non-Player Characters (NPCs)

Rich social interactions among NPCs improve the believability of interactive narratives and the player experience (Afonso and Prada 2008; Swartout et al. 2006). Researchers have manually authored narratives to document cultural heritage and community-based narratives or goals (Speiginer et al. 2015) as well as procedurally generated games and narratives for various geo-locations populated with NPCs (Macvean et al. 2011; Dow et al. 2006; Leino, Wirman, and Fernandez 2008; Azad et al. 2016). We posit that NPCs in real-world locations must be able to learn cultural, and societal values of the location they populate. Leeper and Slothuus (2014) build on prior work by Kunda (1990) discuss reasoning under partisanship (or motivated reasoning) stating a world devoid of partisan conflict is a dystopia. They argue that the novel contribution of motivated reasoning is the idea that individuals vary in the extent to which making accurate decisions is satisfying versus the extent to which they choose to reinforce their prior biases, attitudes or beliefs. Many traditional narrative planning systems allow for the former, with virtual characters able to create robust plans to achieve their goals (Cavazza, Charles, and Mead 2002; Young 2000). Towards the latter, our simulation allows NPCs to reevaluate their convictions over time, attempting to reconcile the disparities in their attitudes with those of their society.

A key challenge posed by characters in a game is their ability to reflect their goals, personalities, and beliefs through dialog or expositions. Rowe, Ha, and Lester (2008) describe how a requirement of the dialog from a character must be that it is appropriate for the character personalities and preference while taking into account the narrative context and history (Rowe, Ha, and Lester 2008). With this paper, we do not directly address the natural language content generation of the conversation. Our system instead produces modifiers and keywords that in combination with a templating mechanism could state the intention of the characters and be used to produce natural language dialog utterances.

Social Simulation

We argue that our research is a step towards machine enculturation (Riedl 2016) by simulating a society of virtual characters that have a predisposition towards learning new knowledge, cultures, and values based on their past interactions with both family (nature) and other societal influences (nurture).

Extensive research has been conducted on social rules and interactions between virtual characters. Versu (Evans and Short 2014) shows characters interacting with one another using pre-constructed social practices templates. Similarly with CiF, in Prom Week (McCoy et al. 2011) the authors describe a social physics architecture model that constrains how NPCs behave. With their Actor-Network Theory (ANT) Latour discusses how individuals relating to one group or another is an ongoing process made up of uncertain, fragile, controversial and ever-shifting ties (Latour 2005). Our simulation consolidates these two approaches, that of ANT and the traditional narrative intelligence approach. Virtual characters' group membership changes over time based on their recognition of their internal attitudes and the opinions of characters around them. With this approach, rather than manually authoring social rules and beliefs, such as with Versu (Evans and Short 2014), social rules emerge organically over time as beliefs and attitudes that go against the group's values would be looked upon unfavorably by its members.

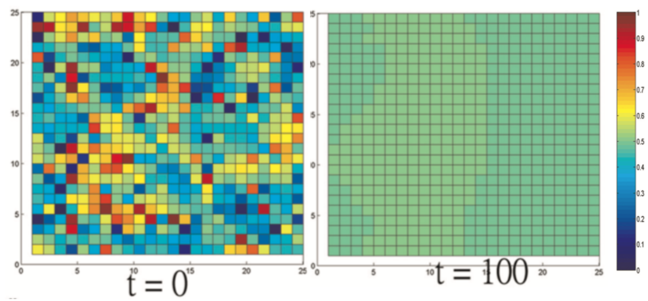

Figure 1: An evolution of agent attitude dynamics represented by cellular automata (Wang, Huang, and Sun 2014). The left graph shows initial variation in opinion and the right graph shows the more homogeneous opinions after 100 iterations.

Finally, our prior work extends current theories of dynamic opinion modeling research (Wang, Huang, and Sun 2014; Asch 1955). These works allowed for a grid-based society with agents surrounded by the same neighbors through their entire lifespan. Fig. 1 show how this approach results in a homogenous distribution of opinions over time, not accounting for new or dissolving relationships amongst antagonistic NPCs in a societal structure. We extend this work with the goal of being able to model societies with NPCs capable of exploring complex and contentious issues of politics, religion, making decisions, and forming social relationships based on their views.

Background: Lyra

We briefly review the Lyra social simulation system upon which our experiment is built. Due to space restrictions, we refer readers to Azad and Martens (2018) for a more detailed overview of the system.

Topics	Objects of Discussion	Source	Rating
Political Issues e.g. Immigration, Gun Control	Individual news articles	Online or Print Media	Political Bias or Affiliation
Political Issues e.g. Immigration, Gun Control	Political candidates	Articles, Interviews, Candidate Rally	Approval Rating
Research Topics e.g. AI, Games	Conference Papers	Journals, Conference Proceedings	Journal or Conference Rankings
Film Genres e.g. Horror, Sci-Fi	Movies	Movie Studios	Rotten Tomatoes ratings

Table 1: Examples showing how the Lyra (Azad and Martens 2018) knowledge model can simulate discussions in various conversational domains

Knowledge Model

Our knowledge model describes how information in the simulation world is structured. It consists of *objects of discussion* as the basic unit of information, which are clustered into *topics*. Each piece of information is generated by a *source* and given a value by a *rating*, which represents either (1) the personal judgment or favor associated with the presentation of the information, or (2) a measure of the impartiality of the unit of information.

Our simulation uses a corpus of news articles from AllSides.com (AllSides 2018) that use a combination of blind bias surveys, editorial reviews, third-party research, independent research, and community votes to calculate media bias of the information.

Virtual Character's Views

We represent a character's views as consisting of an *Attitude*, an agent's private views on a specific object of discussion, an *Opinion*, an agent's outwardly expressed or shared views, and an *Uncertainty* about their views. Opinions and Attitudes are real numbers in the range [-1, 1] and represent an evaluation of the Object of Discussion that range from Strongly Left to Strongly Right. Uncertainty is a real number in the range [0,1]. The agent may have lower confidence in their attitude if (1) information in their existing knowledge base inadequately back them, (2) if contradictory opinions are presented to the agent with high certainty, or (3) if the agent is surrounded by a society that disagrees with them.

Additionally, we use two thresholds, a *Public Compliance Threshold* which describes when the agent chooses to comply with the public opinion to feel accepted within the community, and a *Private Acceptance Threshold* which describes when an agent will choose to stand by their views. Finally, we define a *Bias* to be the agent's predisposition to adopt a particular leaning (left/right) on a topic in a discussion.

Simulation of Discussion

We begin by clustering similar expressed opinions of all participants of the conversation using the Jenks Natural Breaks Optimization method (Jenks 1967). This mirrors how humans interact. For instance, a group of fans may congregate at a water cooler at work, forming coalitions of people that argue about who should rule Westeros (Benioff and Weiss 2019). However, the same participants could have different opinions (and therefore social relationships) based on their shared interests in computer science, or hiking. The number of opinion groups formed indicate whether a *public opinion* on the matter has developed and the presence of normative social influence (or peer pressure). The fewer the number of clusters that form, the more likely it is that an agent who maintains their views contrary to public opinion will feel rejected (Wang, Huang, and Sun 2014).

Public Opinion formed We calculate each agent's change in views based on their certainty and the strength of others' views. Agent's with high uncertainty in their own views are more likely to accept the public opinion and their views are modified accordingly. If the agent has low uncertainty, we find the largest clustered opinion group with views closest to that of the agent. We then calculate the public opinion strength for the selected group and decide if an agent's attitudes or opinions are affected. The strength of the public opinion as perceived by each agent is affected by (a) number of conversationalists, (b) the homogeneity in the expressed opinions of the group, (c) the perceived discrepancies in the attitude and opinion of the agent.

No Public Opinion formed The agent finds the cluster of opinions with the opinions most similar to theirs. The NPC modifies their opinion to the mean of the cluster and their internal attitudes on the information being discussed.

Goals

Described briefly in the earlier section, prior work established the Lyra system and our model of world knowledge, taking into account biases associated with the knowledge and its source (Azad and Martens 2018). This work builds on Lyra, simulating opinion dynamics in the context of individual interactions amongst NPCs in a virtual town. We expand on our earlier work by adopting the following goals.

- **G1:** To generate descriptions of the change in opinions of the conversationalist NPCs that allow readers to follow an NPC's reasoning.
- **G2:** To evaluate these generated conversations with a human subject study for their believability.
- **G3:** To extract insights from the study that can inform future research on how contentious discussions with polarizing views could impact NPC social intelligence, and more believably simulate the spread of opinions.

These goals describe the remaining structure of this paper. We describe steps to achieve G1 in our section, *Designing Legible Simulation Output*. Likewise, the study design and approach for G2 can be found in our *Study Design* section. Finally, for G3 we described the results from our study in the *Analysis* and *Discussion* sections where we analyze study results to answer four research questions that can help guide future research on character believability.

Designing Legible Simulation Output

We redesigned the simulation output to be presented to the reader in discrete rounds. A critique of our earlier system lay in readers having difficulty understanding and producing explanatory descriptions of how and why characters changed their mind over time. A sample output from our earlier system can be seen in Fig. 2. In this section we describe our design process for creating legible simulation output to human readers.

```
Discussion for 11 minutes on "Room for Debate: Should `Birthright Citizenship' Be Abolished?"
Source: NYTimes
Participants: Richard Cain, Ruth Franklin, Vickie Nguyen-Self, Suzanne Sorenson

Views:
Richard: less uncertain.                                Ruth: no change in views.
Old --> att: -0.5 | op: -0.559 | unc: 0.632           att: 0.0  | op: -0.063 | unc: 0.225
New --> att: -0.5 | op: -0.559 | unc: 0.554           att: 0.0  | op: -0.063 | unc: 0.225

Vickie: change in views, less uncertain.               Suzanne: no change in views
Old --> att: 1    | op: 0.535  | unc: 0.464           att: -0.5 | op: -0.559 | unc: 0.059
New --> att: 0.948| op: 0.521  | unc: 0.497           att: -0.5 | op: -0.559 | unc: 0.059
```

Figure 2: Simulation output from an earlier version of Lyra.

Problem: Choice of Conversational Domain The Lyra knowledge model can be used to simulate conversations in a variety of domains while affording the same discussion and opinion modeling (see Table 1). With this study we needed to choose a familiar domain where our target demographics could imagine accompanying dialogues, and be able relate to the forming of clusters and coalitions of like-minded NPCs. Additionally, respondents should be able to judge the NPCs in swaying others to their perspectives for their believability.

Solution: Political Domain Chosen Our reasons for selecting the US Political System as our chosen domain were threefold. Firstly, this subject matter was considered to be familiar and relatable for our target survey demographics. Next, the range of political stances on the topic have familiar, quantifiable metric (see Fig. 4). Finally, the topic could elicit inferences of plausible dialogue occurring amongst characters based on the respondent's own experiences of past politically charged conversations. This would enable respondents to better judge our generated conversations for believability. For the purpose of this study, we limited the topics of discussion in the domain to *Immigration*, and *Gun Control and Gun Rights*.

Problem: Authoring Bias for Dialogues Authoring accompanying dialogue to match the views of the characters per conversation round was found to be untenable. It was not our intention to author the natural language content of the opinions proffered by the characters during the rounds. Given the thesis of this paper, any human authoring of content would need to be rated for the bias of its author and the content.

Solution: Designing Textual Descriptions To circumvent the authoring bias problem, we generated descriptions of these conversation choices that would allow the virtual characters to explain their internal state, actions taken, and any changes in their attitude without the content of the opinions being shared. We a sample conversation excerpt in Fig. 3 depicting a round of a conversation among 4 NPCs at a school.

In the excerpt, Ada realized they were experiencing cognitive dissonance, and chose to reconcile the perceived difference between their internal attitude and the opinion they expressed to other characters.

Ada Lawson realized the opinion they expressed was inconsistent with their internal attitude on the article. They looked for the group with views closest to their own expressed opinions. The closest group was the one with Johnnie Helm.
Ada Lawson thought about whether the group opinion was strong enough. After an internal debate Ada Lawson realized that the strength of the group's convictions was too weak.
Ada Lawson did not change their mind.

Figure 3: Excerpt of a generated conversation

Problem: Following the Change in Character Views A critique of the earlier version of Lyra was that it was hard to follow the change in a character's views over time. While the final political affiliations and opinions can be seen in Fig. 2, it was hard for readers to understand what a conversation between these characters could look like, or evaluate whether these changes were believable.

Solution: Our Simplified Political Rating Scale To make the change in the character's opinions more visual, and easy to relate to we used a simplified rating system for the political affiliation of the virtual participants. All Graphs summarizing the conversation for the participants used this scale going from -1, representing "left" on the political spectrum, to 1, representing "right" on the political spectrum.

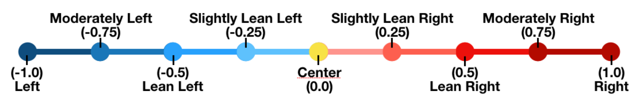

Figure 4: Simplified political scale for each topic discussed

Problem: Lengthy Textual Descriptions Initial practice runs of the survey made it apparent that our subjects found it difficult to track all the variables mentioned (for instance, attitude, opinion, uncertainty, familiarity with topic, etc) described in the conversation text.

Solution: Graphical Descriptions We supplemented our textual descriptions of the conversation with two summary graphs that showed the swing in the opinions and the swing in the uncertainty for the characters over the course of the conversation rounds (see Fig. 5).

Study Design

In order to understand Lyra's effectivess at believably simulating opinion propagation and the social dynamics of politically charged conversations, we conducted a human subjects study asking readers to read simulation output and answer questions in a survey. In this section we describe our survey procedures and analysis process.

Procedures Our survey asked questions to determine participants' political affiliations and biases, the news media sources they subscribed to, and how differing opinions affected their social relationships. Next, they read 4 computer

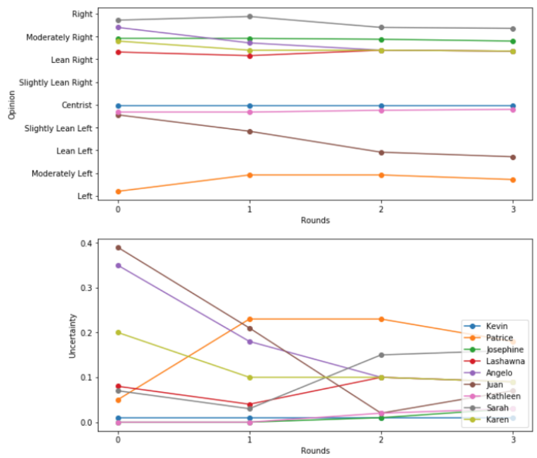

Figure 5: Summary of the change in the opinions of the characters over 3 rounds of discussion.

generated conversations between groups of virtual characters with different political ideologies and biases (Fig. 3) and looked at charts summarizing the rounds (Fig. 5). They were then asked to rate each conversation for it's believability, what the most and least believable part of the conversation was, and to reason about the change in the views of one or two virtual characters in the conversation. Next, respondents were given the option to enter open text for each conversations for additional feedback. Finally, they were asked to fill out a short demographic form. The survey took an hour to complete and was distributed online via email lists and social media. The first 25 participants that completed it were offered an Amazon Gift card.

Response Demographics Our survey had a total of 21 respondents. Of the respondents, 11 identified as male, 8 identified as female, 1 participant chose to describe their gender in a different way, and 1 declined to respond. When asked about their education, 11 had completed their Master's degree, 4 had completed their Doctoral Degree, and 4 had completed their Bachelor's degree, a participant had an associate degree and another had some college credit but no degree. Of the surveyed, 17 were between the ages of 25-34, and 4 were above the age of 35. 16 of the 21 participants identified with the *Liberal* political descriptor, 4 identified as *Conservative*, and one declined to state a political affiliation.

Analysis and Results

In this section, we detail our research questions along with relevant insights produced by our analysis.

RQ1: Does the measure of the believability of the generated conversations depend on the personal political biases of the respondent?

Given our theme of politics, we hypothesized that the personal biases of the survey respondents could impact their believability ratings of the discussions where those issues were discussed. To test this, we asked participants to rate their political bias on a left to right scale as well as to provide their result from the Pew Research Political Typology quiz (Pew Research 2017). Fig. 6 shows how liberal and conservative respondents rated the believability of the conversations. We found that there was no statistically significan difference in responses from Conservatives and Liberals ($p > 0.05$) in the believability rating of Discussion 4.

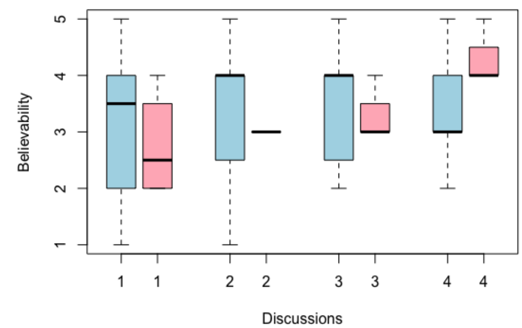

Figure 6: Perceived believability rating of the 4 generated discussions by Liberals and Conservatives (per Pew Research Political Typology results)

Since our data as not normally distributed, we used the non-parametric Mann-Whitney U test to compare the groups. However, the difference between the groups was not significant ($p > 0.05$). This implies the respondents' political preferences on a particular topic did not impact their rating. Interestingly, 3 of 21 participants' familiarity with the *topic* discussed influenced their interpretation of why NPCs did not change their mind. For instance, one participant mentioned that expecting *"people [would be] swayed by the other participants [wasn't] likely with [topics on] gun-control."*

RQ2: Does the measure of believability vary across the generated conversations?

The discussions were generated by varying two parameters in the generator: Group Size (Small and Medium) and Discussion Duration (Short, Medium). After every discussion was described (both textually, and graphically), participants were asked *"How believable was the change in the opinions of the conversationalists through the discussion rounds?"*

We ran the Friedman test (Friedman 1937) to see if there were any differences in perceived believability between the four discussions. We chose the Friedman test since we did not have independent observations for the 4 discussions, since all survey participants analyzed all 4 conversations. We found there were no statistically significant differences in the perceived believability of the four conversations ($p > 0.05$). When asked what the least believable part of the conversation was, 4 of our 21 respondents mentioned they expected a more drastic shift in the opinions of the characters during the lengthier conversations, with one participant describing this as *"expected Mary's rightward shift to be a bit stronger (possibly getting to Moderately Right by Round 6)."*

RQ3: How similar is Lyra's clustering to how humans define and group like-minded NPCs?

For our discussion algorithm, we used Jenks Natural Breaks to group NPCs that expressed similar opinions to each other (Jenks 1967) and then evaluated for the goodness of variance fit to select the optimum number of clusters. Survey participants were shown a chart depicting the opinions of the NPCs on our political scale, and asked (a) How many groups of like-minded conversationalists would form? (b) What groupings of like-minded conversationalists did they expect to see? Respondents used information about an NPC's opinion provided (both textually and depicted on our simplified political scale) to answer these questions.

Table 2: Describes respondents' agreement with Lyra's clustering results and the highest rated clusters.

	Model Agreement	Respondent Agreement
Discussion 1	0.1428	0.666
Discussion 2	0.5714	0.5714
Discussion 3	0	0.238 (tie for best cluster)
Discussion 4	0	0.333

On the whole, only 27% of respondents agreed with the number of opinion clusters generated by our algorithm. Additionally, only 17.8% of respondents agreed with the choice of clustering made by our clustering algorithm. We have summarized the clustering agreement across discussions in Table 2. While the Jenks Natural Breaks Optimization algorithm tries to reduce the sum of the squared deviations from the cluster's mean, this optimization created a greater number of clusters than the numbers suggested by our participants 70.23% of the time. This can be seen in Fig. 7. The respondents chose to create their own clustering, with 50% of the total respondents in agreement on two similarly ranked alternatives. We have shown one of these groupings in Fig. 7. In contrast, our algorithm generated 7 clusters for the 9 NPCs during the start of the discussions. However, after the second round, our algorithm's clustering results agreed with that of the majority of the respondents. The change in the views of the NPCs during those rounds can be seen in Fig. 5 in the Study Design section of this paper. During the feedback for the conversations, survey takers talked about the clustering of NPCs into coalitions through the conversation favorably.

RQ4: Does using the Lyra model impact the believability of the virtual characters?

Overall, the four conversations had a mean believability rating of 3.3. We further asked respondents what the most and least believable part about the generated conversations were, and qualitatively analyzed their responses. We report some of the more interesting replies below.

On NPC Views When asked what the most believable part about the conversations was, respondents had varied responses. 10 participants (over the course of 4 conversations) described how various NPC's staying consistent with their views was believable, writing that *"the centrist didn't change their opinion much," "Johnnie stayed their ground,"* and *"Mary's belief in her opinion solidified.. [and] her uncertainty reduced."* Participants also found the change in a specific character's views to be believable, writing *"Helga started out on the left, moved to the center for a round, but then eventually ended on the left,"* and *"James becoming slightly more uncertain (0.3) on account of changing their opinion,"* as examples of believable behavior.

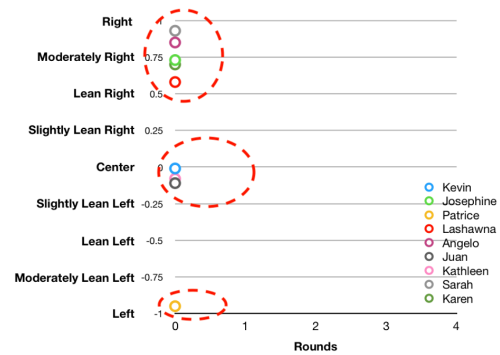

Figure 7: Respondents clustered the opinions depicted here into 3 clusters (shown by the dashed red circles).

10 of our 21 Participants also projected emotions and reasoning on to the NPCs beyond the information we provided. They described how one agent seemed to *"feel marginalized,"* or how an agent's *"competitiveness"* would stop them from changing their views. One described how an NPC, James, seemed *"more concerned about their own rights and interests than the group's."* Another participant described how the agent could have been persuaded since it seemed as though *"the opposition member's confidence and articulation were strong."* Another respondent, while discussing a conversation on Immigration, accounted for the change in the views of the NPC stating the NPC seemed to *"care about the well-being"* of the human population. Finally, one participant described how they were surprised that *"William could be so persuasive with such fluctuating levels of uncertainty."*

On Respondent Bias We found participants tended to project their own bias and experiences on to the agent while explaining why an agent made decisions, with statements such as *"Ada is a typical right-winger and is looking for viewpoints to confirm her own bias; rather than be convinced by others."* One participant pointed out that the "uncertainty of left and centrist participants increased; which in [their] opinion indicates they are open to discussion and change their mind." Others pointed out that the *"Lefts found common ground and [seemed to have] reached equilibrium,"* or that *"Centrists not changing their opinion"* was very believable. One respondent stated that he wasn't surprised that the *"the two right-wingers would converge on [each other's] opinions."*

On Group Dynamics With their responses, 13 participants discussed the relationships of groups and used colloquial terms to refer to them, stating *"once the groups are formed they stayed the same "* as adding to the believability of a conversation. Participants mentioned how *"people tended to cluster into ideological groups,"* and stated that *"group formations seemed coherent with each member's affiliation."* Another discussed how they found it very believable that *"people would group up when views were similar; but not the same."*

The consistency of a group's opinions was also brought up. Respondents said *"the consistency with which the Right Opinioned people stuck to their stand"* added to the believability, another noted the fact that how the *"Left-leaning [NPCs] seemed to become more uncertain of their views over time,"* or as one respondent phrased it, *"the unchanging mind of the majority"* seemed to be the most believable part of the conversation.

On Group Influence Participants discussed how group members were able to influence their groups. One person pointed out that *"groups were swayed more easily when they engaged with more people."* Several respondents discussed how the group was able to influence their constituent member's views. 11 participants pointed out specific instances of group influence as adding to believability and were able to pinpoint these events. One participant discussed how *"The fact that James [had] not changed drastically on his political opinion but [had] opened up his opinion to uncertainty seemed believable since he [was] outnumbered."* Of the most polarizing of conversations generated (with 3 agents, one with a centrist view, and the other two on extreme ends of our political spectrum), one respondent pointed out it was believable since *"no substantial agreement was reached; which is what you might expect from an argument where people's views start out very highly separated from each other."* Another described the same stasis in views as, *"3 people could not reconcile [their] views"* as they started out with such different perspectives. However, one participant indicated the converse, mentioning, *"people got stronger or more confident in their views after discussing/arguing about them; not weaker."*

On Reasoning Why We asked the survey takers why they thought an NPC changed or updated their view. Of the conversation modelled in Fig. 5, one respondent pointed out similarities to human conversations stating that *"the unexpected move of Juan towards the Left and Patrice's position feels like the kind of strange turn that might happen in a real conversation - in a large enough conversation you will see some people's opinions change."* However, this participant listed the same fact as both the most and least believable part of the conversation, wondering why Juan would change their views.

12 of 21 respondents interpreted the change in the way our algorithm performs it stating that, *"he reaffirmed his left bias,"* or *"he didn't want to seem biased externally so wanted to be portrayed as a centrist; but was privately left-leaning."* 3 of these 12 participants discussed how the change was probably due to the fact that Juan's *"view was probably more left-leaning than Juan initially realized."* These responses are a good indication that our system Lyra models opinion change in a way that is expected and realistic. Only 2 of 21 respondents were unable to describe why the change could have occurred, stating that *"nothing [they] could tell"* seemed to be the reason. 1 other respondent blamed peer pressure as a reason for the change in views. 1 respondent interpreted Juan's stance as a call for *"support for innovation and reform [in Gun Control],"* although we remind readers that we did not generate content. 2 respondents did not respond.

Discussion

Revisiting the goals of our paper, our first goal (G1) was to generate descriptions of the change in the opinions of the conversationalist NPCs that allowed readers to follow the NPC's reasoning. Our design process for these generated conversations was described in our section, *Designing Legible Simulation Output*. Of the 21 respondents, 17 were able to interpret the conversations and use them to reason about NPC behaviour. 4 participants stated that they had difficulty following the conversation description. One participant mentioned that the descriptive text provided by us made it *"difficult to align with [their] own mental model of the dynamic. The graphs help; but the textual description is pretty poor [and] too abstract."* Overall, we believe that these responses satisfy our goal. Our system can produce modifiers and keywords that state the intention of the characters in a manner that meets the expectations of the reader. In the future, these could be used to produce natural language dialog utterances.

Our next goal (G2), was to evaluate these generated conversations with a human subject study for their believability. We describe the design and method of our study in the *Study Design* section. One limitation of our study was the small number of respondents and the fact that they were mostly on the left of the political spectrum. Our population sample was not normally distributed, making it difficult to test for statistical significance in our analysis. Overall, the four conversations had a mean believability rating of 3.3.

Finally, our goal (G3) to extract insights from the study to inform future research can be seen in our section on *Analysis and Discussion*. With our four research questions we conducted a summative evaluation of our simulation. With our qualitative analysis we learned how respondents felt NPCs believably form coalitions. Our reasoning questions showed that most respondents were able to interpret and expect the change in NPC opinions in the way our algorithm performed it. Additionally, respondents displayed emotional responses to the conversations they read, for instance, stating that they found it *"believable but depressing that [none of the NPCs] ultimately changed their minds [on Immigration] at the end of Round 3."* Overall, our analysis demonstrates that depicting a change in the opinions of the NPCs can impact the relatability and believability of NPCs.

With future work, we aim to use our results to inform how discussions with conflicting opinions could influence social relationships in a more extensive, geographically-situated population simulation. We believe our evaluation shows that Lyra can simulate believable NPCs with the ability to model social influence and opinion dynamics, enabling more robust social intelligence for virtual populations and models of human social dynamics.

References

Afonso, N., and Prada, R. 2008. Agents that relate: Improving the social believability of non-player characters in role-playing games. In *International Conference on Entertainment Computing*, 34–45. Springer.

AllSides. 2018. Balanced news via media bias ratings for an unbiased perspective.

Asch, S. E. 1955. Opinions and social pressure. *Readings about the social animal* 193:17–26.

Azad, S., and Martens, C. 2018. Addressing the elephant in the room: Opinionated virtual characters. In *Experimental AI in Games Workshop co-located with the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Azad, S.; Saldanha, C.; Gan, C.-H.; and Riedl, M. 2016. Mixed reality meets procedural content generation in video games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Benioff, D., and Weiss, D. B. 2019. The iron throne.

Brichacek, A. 2016. Six ways the media influence elections. *University of Oregon School of Journalism and Communication (Online News Page)*.

Cavazza, M.; Charles, F.; and Mead, S. J. 2002. Interacting with virtual characters in interactive storytelling. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, 318–325. ACM.

Dow, S.; Mehta, M.; Lausier, A.; MacIntyre, B.; and Mateas, M. 2006. Initial lessons from ar façade, an interactive augmented reality drama. In *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, 28. ACM.

Evans, R., and Short, E. 2014. Versua simulationist storytelling system. *IEEE Transactions on Computational Intelligence and AI in Games* 6(2):113–130.

Fowler, S. M., and Pusch, M. D. 2010. Intercultural simulation games: A review (of the united states and beyond). *Simulation & Gaming* 41(1):94–115.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* 32(200):675–701.

Jenks, G. F. 1967. The data model concept in statistical mapping. *International yearbook of cartography* 7:186–190.

Kunda, Z. 1990. The case for motivated reasoning. *Psychological bulletin* 108(3):480.

Latour, B. 2005. *Reassembling the social: An introduction to actor-network-theory*. Oxford university press.

Leeper, T. J., and Slothuus, R. 2014. Political parties, motivated reasoning, and public opinion formation. *Political Psychology* 35:129–156.

Leino, O.; Wirman, H.; and Fernandez, A. 2008. *Extending experiences: structure, analysis and design of computer game player experience*. Lapland University Press.

Macvean, A.; Hajarnis, S.; Headrick, B.; Ferguson, A.; Barve, C.; Karnik, D.; and Riedl, M. O. 2011. Wequest: scalable alternate reality games through end-user content authoring. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, 22. ACM.

McCoy, J.; Treanor, M.; Samuel, B.; Mateas, M.; and Wardrip-Fruin, N. 2011. Prom week: social physics as gameplay. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, 319–321. ACM.

Morrison, H., and Martens, C. 2018. Making first impressions: A playable model of cross-cultural trust building. In *INTWICED@ AIIDE*.

Pew Research. 2017. Political typology quiz - where do you fit in the political typology?

Riedl, M. O., and Harrison, B. 2016. Using stories to teach human values to artificial agents. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Riedl, M. O. 2016. Computational narrative intelligence: A human-centered goal for artificial intelligence. *arXiv preprint arXiv:1602.06484*.

Rowe, J. P.; Ha, E. Y.; and Lester, J. C. 2008. Archetype-driven character dialogue generation for interactive narrative. In *International Workshop on Intelligent Virtual Agents*, 45–58. Springer.

Speiginer, G.; MacIntyre, B.; Bolter, J.; Rouzati, H.; Lambeth, A.; Levy, L.; Baird, L.; Gandy, M.; Sanders, M.; Davidson, B.; et al. 2015. The evolution of the argon web framework through its use creating cultural heritage and community–based augmented reality applications. In *International Conference on Human-Computer Interaction*, 112–124. Springer.

Swartout, W.; Hill, R.; Gratch, J.; Johnson, W. L.; Kyriakakis, C.; LaBore, C.; Lindheim, R.; Marsella, S.; Miraglia, D.; and Moore, B. 2006. Toward the holodeck: Integrating graphics, sound, character and story. Technical report, University Of Southern California Marina Del Rey CA Inst For Creative Technologies.

Wang, F.-Y.; Carley, K. M.; Zeng, D.; and Mao, W. 2007. Social computing: From social informatics to social intelligence. *IEEE Intelligent systems* 22(2):79–83.

Wang, S.-W.; Huang, C.-Y.; and Sun, C.-T. 2014. Modeling self-perception agents in an opinion dynamics propagation society. *Simulation* 90(3):238–248.

Warpefelt, H. 2016. *The Non-Player Character: Exploring the believability of NPC presentation and behavior*. Ph.D. Dissertation, Department of Computer and Systems Sciences, Stockholm University.

Young, R. M. 2000. Creating interactive narrative structures: The potential for ai approaches. *Psychology* 13:1–26.