Lyra: Simulating Believable Opinionated Virtual Characters

Sasha Azad
Principles of Expressive Machines (POEM) Lab
NC State University
sasha.azad@ncsu.edu

Abstract

Creating believable simulations of large populations of characters in virtual worlds represents a grand challenge for narrative intelligence, requiring reasoning about social interaction, cultural norms, and human decision-making. 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 present a preliminary computational investigation into modelling opinion change in virtual characters with this goal in mind. 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. We conducted a human-subjects study to evaluate the generated conversations and affinity groups for their believability and to inform future iterations of the simulation. The conversations were found to be moderately believable, rating 3.3 out of 5 on a Likert Scale. Additionally, survey respondents ascribed humanity to the actions of the virtual agents, associating them with competitiveness, persuasiveness, and intention. We believe this further adds to the believability of our system. In the long run, successful simulation of opinion change in social dynamics provides a foundation for computational recognition, prediction, and interfacing with human social behaviour.

Introduction

Atticus: *"You never really understand a person until you consider things from his point of view–"*
Scout: *"Sir?"*
Atticus: *"–until you climb inside of his skin and walk around in it."*

Humans are rational and emotional beings. Their social systems are complex and contextual. The quote above from Lee's fictional yet well-beloved lawyer, Atticus Finch, is a familiar one (Lee 1960). Atticus believes that the prevalent bias and anti-social behaviour against members of their town is a consequence of not understanding different perspectives, leading to townsfolk being discriminated against or shunned. Understanding such behaviour and simulating it with virtual characters requires reasoning not just about observable social network graphs or social interactions, but also about geography, economics, and increasingly, online participation and discourse. 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 modelling 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 with artificial 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 find interactions between characters to be more believable if they adhere to recognizable social practices and plausible enculturated (Riedl and Harrison 2016) responses to social situations (Warpefelt 2016). However, these simulated models typically do not account for some of the most important features of social networks, namely that of the *social dynamics of opinion change* and its cause and effect relationship with social relationships.

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 studying this phenomenon could be to analyze data generated by real user interactions on social networks, we posit that modelling 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

Submitted to the Department of Computer Science at North Carolina State University in partial fulfillment for the Doctoral Written Preliminary Qualifier Exam 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 modelled 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 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 on a Likert scale ranging from 1-Not Believable at all to 5-Very Believable. We discovered the discussions had a mean believability rating of 3.3. 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 (Togelius et al. 2013) 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. In summary, this work (1) overviews the previously established Lyra system (2) describes the design process and generation of conversational metadata (3) evaluates the generated conversations with a human subject study for their believability (4) extracts insights from the study to inform future research on how contentious discussions could affect social relationships amongst NPCs to more believably simulate the spread of opinions.

Related Work

In this section, we first describe related work from the narrative domain on believable virtual characters. Next, we discuss group formation from the perspective of social scientists, and psychologists to understand how believable virtual characters could be modelled to respond to group (or societal) archetypes and opinions.

Believable Non-Player Characters (NPCs)

There is no generally agreed-upon definition of believability. Instead within the narrative field, believability is used linguistically to describe that which is believable by someone. In terms of virtual characters, this could imply some aspect of their viewed interactions (either with the player or with each other) is believable. Togelius et al. (2013) describe how games that incorporate believable elements can elicit particular emotional responses to a player. They discuss how the generation of believable, human-like opponents lead to increased player enjoyment. Additionally, rich social interactions among NPCs have been found to improve the believability of interactive narratives and the player experience (Afonso and Prada 2008; Swartout et al. 2006).

With the wide-scale availability of mobile devices, and more recently the adoption of augmented reality (AR) technologies, researchers have manually authored narratives to document cultural heritage and community-based narratives or goals (Speiginer et al. 2015) as well as procedurally-generated narratives for various geo-locations populated with NPCs (Macvean et al. 2011; Dow et al. 2006; Leino, Wirman, and Fernandez 2008). 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 this goal, our simulation allows for an NPC to evaluate their convictions over time, attempting to reconcile the disparities in their attitudes and beliefs with those of the other NPCs they interact with via conversation.

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

Bias

Bias, in algorithms or decision making, can impact government, businesses and personal lives. Studying the impact of bias has recently become an emerging trend to study in computer science (Budak, Goel, and Rao 2016; Entman 2007; IBM Research 2018). Entman describes how the term can apply to news that distorts or falsifies reality (distortion bias), or news that favours one side rather than providing equivalent treatment to both sides in a political argument (content bias), or even with respect to the motivations and judgment behind decision-making processes (decision-making bias) (Entman 2007). Entman describes how studying media bias can provide insight into how the media influences the distribution of power. However, the bias in the media is not necessarily all bad. AllSides notes on those media outlets ranked with Center biases may leave out valid arguments from the left or right perspectives (AllSides 2018).

Our work aids these efforts by endeavouring to use existing computational social psychology models to simulate how humans respond to and make decisions when faced with authority bias, or even how they respond to the cheerleader effect in conjunction with our social simulation.

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. These templates are constructed manually, can be time-intensive and require domain knowledge. 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 (as in systems like 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 unfavourably by its members.

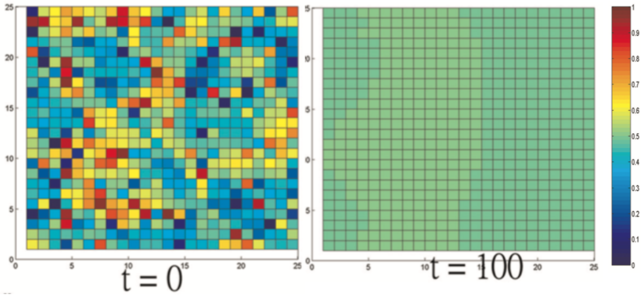


Figure 1: An evolution of agent attitude dynamics represented by cellular automata (Wang, Huang, and Sun 2014). The left graph shows initial variation in opinion and the right graph shows the more homogeneous opinions after 100 iterations.

Finally, our prior work extends current theories of dynamic opinion modelling research (Wang, Huang, and Sun 2014; Asch 1955) with the goal of being able to model societies with NPCs capable of exploring complex 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.

Knowledge Model

Our knowledge model describes how information in the simulation world is structured. This can be overviewed as follows. For a single discussion, the participants in the discussion choose an *Object of Discussion* to converse on, obtained from a *Source*. The Source and the Object of Discussion are associated with a *Rating*. Multiple objects of discussion can be clustered to form a *Topic*.

Our model of the knowledge base can be used for a large variety of datasets while affording the same discussion and opinion modelling. For instance, simulating debates among NPCs about *current news articles* clustered by *political issues* and ranked by their *bias*. Similarly, we could use our model to discuss the merits of various *journal articles* clustered together by *research topics* and ranked by *journal rankings* or have audience members discuss their *movie preferences* clustered by *movie genres* and ranked by their *Rotten Tomatoes rankings*. Some datasets considered during the design phase of this model have been highlighted in Table 1.

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.

Ratings are defined as the value of the information learned by the NPC in the system. This rating could represent either (1) the personal judgment or favour associated with the presentation of the information, or (2) a measure of the impartiality of the unit of information.

Topics are a clustering of information regarding a specific subject, or field of information. A specific information unit can be a part of multiple topics at the same time. For instance, a discussion of procedural content generation could belong to the topics of both artificial intelligence or game design.

Objects of Discussion This single unit of information forms the basis of our discussion model. While interacting with one another, virtual characters search through their knowledge base and conversational repertoire, choosing a single object of discussion to debate. An NPC that adds a new object of discussion to his knowledge base will note the original authorial rating intended to be affiliated with the information, and associate with it their own opinions on the topic. These views could be based on prior discussions of the information with conversationalists that introduce the character to the information, as well as on the character's current view of the topic to which the information belongs.

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

Sources may create information covering a wide variety of objects of discussions and topics. Sources may also have associated with them a rating, representing the expected rating of the information they produce. NPCs may use this rating to choose to subscribe or unsubscribe to these over time based on their current inclinations.

Virtual Character's Views

Every participant in the discussion has their own *Bias* and *View* on the information and can express their opinions on the object of discussion at hand. These elements and our dataset have been described in further detail below. The attributes of an agent's view are modelled based on those by Wang, Huang, and Sun (2014).

We represent these NPC 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 a *Uncertainty* about their views. 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.

Bias is the agent's predisposition to adopt a particular view on a topic in a discussion. This bias is informed by either (1) the agent's views inherited from their parents or (2) a mean of their views on all objects of discussion under the said topic or (3) the initial bias they learn from the conversationalists when the topic was added to their knowledge base during a discussion.

Attitude (*att*) is the agent's private views on a specific issue. Attitude is a real number in the range $[-1, 1]$ and represents an evaluation of the object of discussion.

Opinion (*op*) is an agent's outwardly expressed or shared views on a specific issue. Like attitude, opinion is a real number in the range $[-1, 1]$ and reveals the agent's opinion on the object of discussion to the other dialogists. There may be a discrepancy in the attitudes and opinions of the character since a character may not represent their attitudes accurately to participants. A human example of the situation where this is apparent can be seen in examples of an employee in conversation with his managers who choose not to express his disagreement to avoid being punished.

Uncertainty (*unc*) is a measure of an agent's confidence in their view. The higher the uncertainty, the more likely the agent is to change his mind or be accepting of other perspectives. As an example, an NPC may express opinions about the legality of abortion in their town. However, 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 a majority of whom disagrees with him. unc is a real number in the range $[0, 1]$.

Public Compliance Threshold (*pub_thr*) : When the strength of the public opinion exceeds this value, the agent will choose to comply with the public opinion to feel accepted within the community.

Private Acceptance Threshold (*pri_thr*) : When the strength of the public opinion is below this value, the agent will choose to stand by their views. The pri_thr is a real number in the range $[0, 1]$. Professors or experts on a particular topic in our simulation have higher values to indicate their expertise.

Simulation of Discussion

Our model accounts for the fact that the same participants could have different opinions (and therefore social relationships) based on their shared interests in other discussion topics, such as computer science, or hiking. This allows for relationships where characters that agree over a few views but disagree over others to change their affinity for one another throughout multiple discussions.

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). The number of opinion groups formed indicates 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 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:

- The size (f_a) of the group. The larger the group, the stronger the public opinion.

$$f_a = \begin{cases} 0\text{,if } x_a \leq 1 \\ x_a/10\text{,if } 1 < x_a \leq 10 \\ 1\text{,if } x_a > 10 \end{cases}$$

- The homogeneity (f_b) in the opinion of the group defining if the group come to a consensus

$$f_b = 1/(1 + e^{24x_b - 6})$$

- The discrepancies (f_c) in the agent's opinion and attitude.

$$f_c = 1/(1 + e^{-12x_c + 6})$$

Next, the agent measures their own uncertainty with the strength of the public opinion by calculating two threshold values, $th_1 = 1 - agent.unc$ and $th_2 = \max(0.6, th_1)$.

- Low Opinion Strength ($op\_str < th_1$): If the opinion strength is too weak, the conversationalist does not change their mind, recognizing the discrepancy between their internal attitudes and ideas and those of the group.
- Moderate Opinion Strength ($th_1 \leq op\_str < th_2$):
  - Members with a low uncertainty find the opinion strength of their group strong enough to modify their opinions to the mean of the group. Agents then find their internal attitudes, and their expressed behaviours are inconsistent, and so change their attitudes to match. In this case, agents believe that the change in their views is a natural and expected evolution, and do not realize they are bending to public opinion.
  - Agents with large uncertainty realize that they are conceding the discussion, and bending to public opinion. They change their external opinions and internal attitudes to match.
- High Opinion Strength ($op\_str \geq th_2$): The agent realizes the strength of the opinion. In this case, the agent may choose to conform to the public opinion with their outwardly expressed views and change their opinion to the mean of the group. However, they *do not* change their inner attitudes, and in the absence of external pressure will revert to their attitudes.

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.

Due to space restrictions, we refer readers to our prior work(Azad and Martens 2018) for further details of the algorithm and simulation described above.

Goals

This work builds on a simulation of opinion dynamics presented in previous work (Azad and Martens 2018). This work established the Lyra system (described briefly above) and our model of world knowledge that can take into account biases associated with the knowledge and its source. Additionally, we discussed a model of the characters internal attitude and expressed an opinion on the topic based on prior work that models self-perception agents (Wang, Huang, and Sun 2014). Finally, our simulated characters were able to form ad-hoc groups to discuss their views and closer relationships with characters that had similar perspectives.

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.

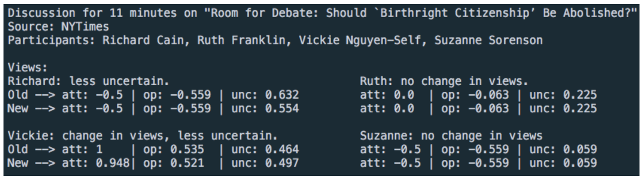
We expand on our earlier work by adopting the following goals.

- **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.



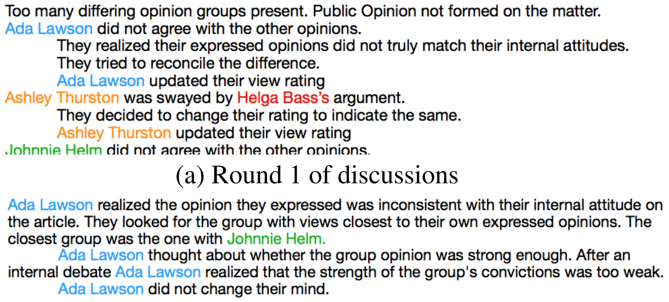
```
Discussion for 11 minutes on "Room for Debate: Should `Birthright Citizenship' Be Abolished?"
Source: NYTimes
Participants: Richard Cain, Ruth Franklin, Vickie Nguyen-Self, Suzanne Sorenson

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 modelling (see Table 1). With this study, we needed to choose a familiar domain where our target demographics could imagine accompanying dialogues and be able to 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 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.


Too many differing opinion groups present. Public Opinion not formed on the matter.
Ada Lawson did not agree with the other opinions.
They realized their expressed opinions did not truly match their internal attitudes.
They tried to reconcile the difference.
Ada Lawson updated their view rating
Ashley Thurston was swayed by Helga Bass's argument.
They decided to change their rating to indicate the same.
Ashley Thurston updated their view rating
Johnnie Helm did not agree with the other opinions.

(a) Round 1 of discussions

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.

(b) Round 2 of discussions

Figure 3: Excerpts from a generated conversation

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.

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.

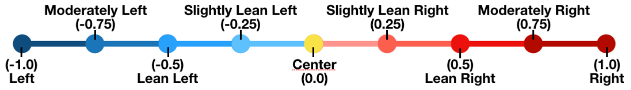


Figure 4: Simplified political scale for each topic discussed

Further, we described both the Media Bias and the Character Attitudes on topics on the spectrum using descriptions for the positions from Allsides.com (AllSides 2018). These descriptions, provided to the survey respondents, have been added to the Appendix at the end.

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 the 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 throughout the conversation rounds (see Fig. 5).

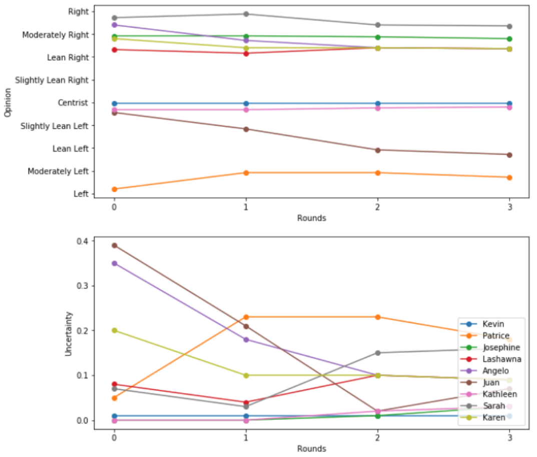


Figure 5: Summary of the change in the opinions of the characters over 3 rounds of discussion.

Study Design

To understand Lyra's effectiveness 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 computers generated conversations between groups of virtual characters with different political ideologies and biases (see Fig. 3) and looked at charts summarizing the rounds (see Fig. 5). They were then asked questions regarding the believability of the conversations, and the intentions of the virtual characters participating in said conversations. They were also asked to rank the persuasiveness, and feelings of membership or inclusiveness with the group for each character. Participants were given the option to enter open text for the conversations for additional feedback or take-aways. Finally, they were asked to fill out a short demographic form.

The survey was distributed online via email lists and social media. The first 25 participants that completed the survey were offered to be paid with an Amazon Gift card. The survey took about one hour to complete.

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 differently, 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.

Method: Qualitative Analysis

With this section, we detail our method for the qualitative analysis of the survey results.

Constructing Queries We chose to use a directed approach to content analysis in both phrasing and analysis of open text queries asked during the survey (Mayring 2004). Our goal was to be able to validate and extend conceptually our theoretical framework and model for opinion dynamics amongst NPCs. Thus, our queries were framed to probe participant predictions and expectations of the conversation, and explore their understanding of the relationships about our variables of interest. Primarily these included the believability of the conversation, the change in the opinions and attitudes of the participant NPCs, the relationship between uncertainty and change in opinions. Keeping this in mind, we asked participants for open text responses to four questions as detailed below:

- What was the most believable part about the conversation described above?
- What was the least believable part about the conversation described above?
- One reasoning question about an NPC per conversation:
  - Why do you think Ashley was so uncertain of their views?
  - Why do you think Ada's uncertainty reduced?
  - Why do you think James's uncertainty increased?
  - What does Juan's change in opinion tell you of their private attitude of the conversation?
  - Why do you think Amy's uncertainty increased after Round 2?
- Any thoughts of take-aways from this conversation that you would like to share

With the reasoning question, we hoped to incite responses to indicate to us the mental model of the respondent, and whether their interpretation and expectations of the change in an NPC's views matched those of our algorithm.

Directed Content Analysis After the survey ended, the open text responses to the question described above were transferred to a Google Sheets document.

The authors of this paper performed initial open coding used content analysis to analyze the data. Both coders were familiar with the underlying theory of the discussion model, and the formulated research questions. This allowed the initial codes noted to have a more structured, directed approach (Hickey and Kipping 1996) described by Mayring (2004) as deductive category development and application. The step model for this analysis has been depicted in Fig. 6 as shown below.

The data was read from start to end to obtain a sense of the whole. First impressions, and thoughts were noted down to capture key concepts based on the variables of interest (Potter and Levine-Donnerstein 1999). Both authors were aware of the variables of interest for the study, namely, to better understand the uncertainty, attitude, opinions and believability of the conversations. This allowed the authors to create two independent codebooks with explicit definitions, examples and coding rules for each deductive category discovered. A few codes from this initial coding scheme (or codebook) have been described below in Table 2.

The authors performed open-coding analysis independently, reading the responses to derive codes (Miles et al. 1994; Morgan 1993). An initial discussion was conducted to discuss, negotiate and merge the codebooks as a formative check for reliability. For instance, the phenomenon coded as "Influence by personal bias of participant" by Coder 1 and "Liberal Open Minded" by Coder 2 were found to be linked and were merged and renamed to "Used Political Stereotype" for the initial coding scheme formed. Additionally, during the discussion, other codes were organized into more meaningful clusters. For instance, the label "Increasing Certainty Expected" was decomposed into two sets of codes, the

Coder	Code	Description	Sample quote matching open code
1	Influence by a Group	Identification of a group influence on NPC opinions	She was swayed by the rest of the group
1	Influence by personal bias of participant	The comment seems to have been influenced by the participant's own personal bias	The centrists didn't change at all; which doesn't seem characteristic of the topic
2	Decreasing Certainty	Observation of characters' decrease in certainty	The fluctuation from high certainty back to uncertainty in a seemingly short time period.
2	Liberal Open Minded	Stating the belief that liberal people are more open minded or right-wingers are less likely to change their minds	That the most liberal person would be the person most open to changing their mind

Table 2: A few codes from the Initial Coding Scheme

Tag	Definition
#NPCMentionedUnprompted	The behavior of a character was noted when they were not mentioned in the question
#ChangedOpinion	Noting that an NPC or a group of NPCs had a change in their opinions
#StandingGround	No change in opinion. The NPC stood their ground.
#SimilarViewsConverge	Noting that NPCs with similar views eventually converge their views
#GroupInfluence	Noting when an NPC is swayed by a group.
#UsedPoliticalAffiliationStereotype	Respondent made a stereotypical judgement about a political affiliation.
#IndividualInfluence	Noting when a character is swayed by other individuals (but the individuals are not identified as a group)
#DecreasingCertainty	Noting that an NPC's certainty in views decreased / uncertainty increased.
#InferFactsFrom	Infer facts (or make assumptions) that are not given to them by us
#CertaintyConvinces	Noting when a character who is more certain in their views has more influence

Table 3: High frequency codes and definitions obtained after qualitative analysis

first, "Increasing Certainty", and "Decreasing Certainty", the second, "Expected" and "Unexpected". This allowed for the second set to be used in conjunction with other tags, to capture the survey respondents expectations, and allow for more in-depth frequency analysis at a later point. Finally, this initial discussion allowed for the discovery of new codes that only one or the other coder had noticed without being biased. For instance, one author noted that several respondents discussed the existence of an "Overton Window", or that respondents noticed that NPCs with higher "Certainty Convinced Others", while the other author was interested in how respondents displayed an "Emotional Response" to the conversations they read, or that responses could be tagged to indicate whether the "Clustering" of NPCs during the discussion phase was found to be "Believable". This resulted in the creation of an initial coding scheme with 34 codes identified and described with their usage and examples.

Validating Initial Coding Scheme To further establish rigour, reliability, reduce the coding scheme's discriminant capability – that is, reducing coding errors – and to validate this initial codebook, two additional independent coders were recruited. These new coders were unfamiliar with the project and were given a description of the research to explain the purpose of the same. Next, we discussed the Initial Coding Scheme developed, examples of the discovered codes, and what each one meant. Based on this discussion some definitions and examples were further clarified. Next, we selected 15% of the survey data at random from the Google Sheets. This was paired with the discussions and questions that the text was in response to and given to the new coders.

Measure	Agreement Value
Fleiss kappa	0.9099
Cohen kappa	0.9121
alpha	0.9012

Table 4: Table 2 Interrater agreement

Each of the new coders coded the subset of data given to them using the codes in the Initial Coding Scheme developed. They were encouraged to create new codes if they felt that the existing scheme did not fully represent the data. One of the authors also re-coded the same segment using the Codebook provided. Next, we compared results and discussed problems where there were discrepancies in the codes used by the authors and the coders. Adjustments were made after negotiations, and some new codes were discovered, and old codes modified. For instance, the addition of "Opinion Changed Despite Certainty" to depict situations where survey respondents found that an NPC's opinion changed despite their certainty and "Meta Discussion" to capture discussions and feedback about the study design or survey were added. The final Thematic Coding included 44 codes. The Intercoder reliability was calculated using NLTK's Fleiss' Kappa and Krippendorffs alpha statistical measures to assess the reliability of agreement amongst

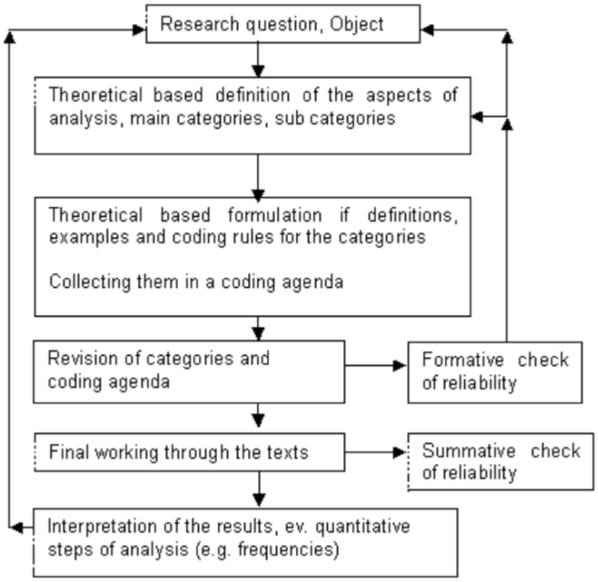

Research question, Object
Theoretical based definition of the aspects of analysis, main categories, sub categories
Theoretical based formulation if definitions, examples and coding rules for the categories
Collecting them in a coding agenda
Revision of categories and coding agenda
Formative check of reliability
Final working through the texts
Summative check of reliability
Interpretation of the results, ev. quantitative steps of analysis (e.g. frequencies)

Figure 6: Step model of deductive category application (Mayring 2004)

the three coders (Fleiss 1971; Loper and Bird 2002). Additionally, Cohen's Kappa was calculated amongst the raters and the author (Cohen 1960) for the random 15% of the data analyzed. The results have been shown in Table 4. The coders were found to have *Almost Perfect Agreement*. Thus, the coding scheme was finalized, and used by the author to code the rest of the data before further frequency analysis was performed.

Thematic Codes Some of the thematic codes that occurred with higher frequency in our analysis have been described in Table 3 above. The remaining codes with their descriptions and sample survey respondent quotes tagged with the thematic codes can be accessed in the appendix to this paper.

Analysis of 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?

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. 7 shows how liberal and conservative respondents rated the believability of the conversations. We found that the responses from Conservatives and Liberals were not significant ($p > 0.05$) in the believability rating of Discussion 4.

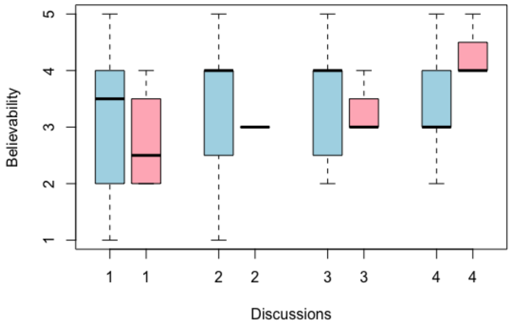


Figure 7: Perceived believability rating of the 4 generated discussions by Liberals and Conservatives (per Pew Research Political Typology results)

We hypothesized that the personal biases of the participants on the topics discussed by the NPCs would impact their believability ratings in the groups where those issues were discussed. To test this, we asked participants to *"Rate their views on a 5 point Likert Scale ranging from 1 (Strongly Left-wing) to 5 (Strongly Right-wing)"* on the topics of Gun Control, Legal Immigration and Illegal Immigration. These were topics discussed by the NPCs to see if their perspectives on a particular topic affected their suspension of disbelief in the generated discussions.

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 experience and interpretation of the conversation. In a discussion generated with smaller variations in the views of the NPCs, one participant mentioned that *"[the fact that] people [would be] swayed by the other participants [wasn't] likely [to happen] with [discussions on] gun-control."* Another participant pointed out that the NPC, Juan's *"views on gun-control aligned with liberal views."*

Finally, we asked respondents to select all the political descriptors from a hand-generated list that they identified with. We ran a linear regression model and found that the political identifiers were not significant ($p > 0.05$). 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."* 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."*

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

The results of this rating has been summarized in Fig. 8 below.

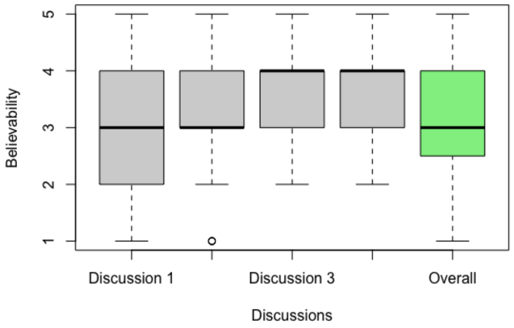


Figure 8: Four box plots showing the perceived believability rating across all 4 conversations as well as the overall believability.

We ran the Friedman test 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 among the 4 discussions analyzed 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)"* with another surprised that *"Shirley was not influenced by the other two in any way."*

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 (GVF) 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.

For the second question, participants were free to choose from a list of groupings that the algorithm evaluated as the highest score for each possible value for number of clusters, or they could enter their own clustering if they disagreed with the choices given to them.

Table 5: Describes respondents' agreement with Lyra's clustering results and the highest rated clusters.

Discussions	Model Clustering	Best Respondent Clustering
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

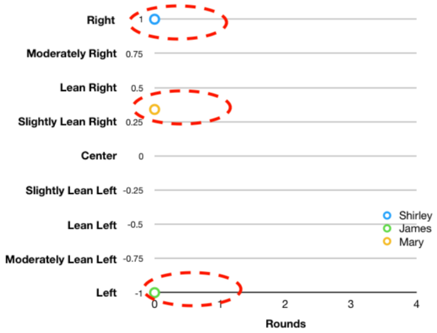


Figure 9: 57.14% of respondents agreed with our algorithm, and clustered the NPCs depicted here into 3 clusters (shown by the dashed red circles), one per NPC, with each NPC disagreeing with the other opinions proffered.

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 5. 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. 10. 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. 10. 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 favourably.

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

Respondents were asked to rate how believable the change in the opinions of the conversationalists was through all the rounds of discussion for each conversation. They were provided with a Likert scale ranging from 1-Not believable at all, to 5-Very believable. Overall, the four conversations had a mean believability rating of 3.3 out of 5. We then qualitatively analyzed their open-text responses to our discussion questions. We report some of the more interesting responses and results from our qualitative analysis below.

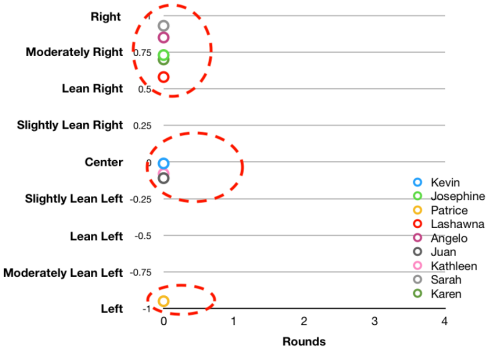


Figure 10: Respondents clustered the opinions depicted here into 3 clusters (shown by the dashed red circles).

Most Believable When asked what the most believable part about the conversation was, respondents had varied responses. The most frequently mentioned themes from their responses have been summarized in Table 6 below.

Theme	**Frequency**
NPC mentioned unprompted	23
NPCs standing ground	18
Similar views converging	12
Influence from groups	10
Used political affiliation stereotype	9
Influence by an individual	8
Polarization	8

Table 6: Frequently occurring codes in response to what respondents found most believable

Breaking down the *#NPCMentionedUnprompted* tag further by the discussion we find that survey respondents interpreted the change in the characters views in the way our algorithm performs it finding that to be the most believable part of the conversation. For instance, in Discussion 1, noting an NPC sticking to their convictions (with a prevalent *#StandingGround* tag), *“Helga started at Left; moved to centrist and then closed at left.”* In Discussion 1 and 4, we also see that influence exerted by other NPCs was accurately recognized (i.e. coded by the *#IndividualInfluence* tag) with statements such as *“Amy was swayed by William”* or *“Ada and Johnnie matching their views,”* being the most believable part of the conversations. Next, in Discussions 2 and 3 the *#GroupInfluence* exerted on NPCs was noted, with respondents pointing out it seemed as though an NPC changed their mind only because they seemed outnumbered or due to peer pressure stating, *“The fact that James had not changed drastically on his political opinion but has opened up his opinion to uncertainty seems believable since he is outnumbered in the group,”* or that *“Lashawna swaying slightly more conservative because she had a very convincing and large group and this would easily move her to similar opinion.”* Additionally, respondents discussed the *#Polarization* of views in Discussion 1 with statements such as, *“That over time and rounds of arguments consensus develops around two poles of thought; even though within the poles there’s a range of opinion/degree of certainty,”* and that *#SimilarViewsConverge* stating, *“No drastic changes in views but groups did come closer to same opinion on both sides.”*

Finally, an interesting observation was that a lot of times people allowed their bias to affect their judgement and *#UsedPoliticalAffiliationStereotypes* while discussing the most believable part of the discussion summary they had just read. This despite the fact that our analysis of RQ1 found these biases did not affect their rating of the believability of the conversation. Respondents described how in Discussion 1 *“The consistency with which the Right Opined people stuck to their stand.”* With Discussion 2, participants thought it was to be expected that *“that the most liberal person would be the person most open to changing their mind,”* while with Discussion 3 they found *“that the centrists didn’t change their opinion much”* was very believable as was the fact that *“the five on the right [were] sticking together.”* Similarly, with Discussion 4 they expected and found believable that *“Lefts found common ground and reached equilibrium.”*

Least Believable When asked what the least believable part about the conversation was, respondents had varied responses. The most frequently mentioned themes from their responses have been summarized in Table 7 below.

Theme	**Frequency**
NPC mentioned unprompted	44
Changed Opinion	19
Decreasing Certainty	11
NPCs standing ground	10
Believable	6
Influenced by Article	6

Table 7: Frequently occurring codes in response to what respondents found least believable

It is heartening to note that 5 of the 21 respondents found Discussion 2 to be entirely *#Believable* and could not describe the least believable part of the conversation stating,*“I find it believable”* as their responses.

Analysing Table 7, and the *#NPCMentionedUnprompted* tag, we initially find it was of almost double the frequency as it’s an occurrence in Table 6 with 44 of the 68 responses in this section specifically calling out individual NPC behaviours as not believable or unexpected. Of Discussion 1, respondents discussed how they did not believe Helga should not have been influenced by the article (i.e. *#ArticleInfluence*) as much as they were, further triggering the change in the views of Ashley and Ada. In Discussion 2, they brought up that they found it unbelievable that *“James (someone who was extreme left) was swayed by [the Centrist] article”* as much as they were. Two participants pointed out that it was the most believable and the least believable fact that Shirley *#StandingGround* was unbelievable stating, *“Shirley had no uncertainty in their views”* and *“not influenced by the other two in any way.”* Of Discussion 3 modelled in Fig. 5, the *#ChangedOpinion* of Juan was pointed out as showing similarities to human conversations with one respondent 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. Finally with Discussion 4, participants found Kenneth's *#ChangedOpinion* unbelievable, stating that they didn't think that *"Kenneth wasn't persuaded much at all; and shifting to the right seemed weird,"* or that *"William would be so persuasive [towards Kenneth] with such fluctuating levels of uncertainty"* was unexpected. While the *#DecreasingCertainty* tag was found to be frequent, there was no consensus amongst respondents on how this affected the least believable part of the discussions with the tag occurrence being sparsely distributed.

Reasoning Queries When asked to reason about an NPCs change in opinion, certainty or attitudes, respondents had varied responses. The most frequently mentioned themes from their responses have been summarized in Table 8 below.

Theme	Frequency
Individual Influence	19
NPC mentioned unprompted	15
Opinion Attitude Difference	12
Infer Facts not provided	11
Group Influence	10
Certainty Convinces	10
Lacking Support	8
Emotions Attributed	7

Table 8: Frequently occurring codes in response to the reasoning questions asked

#IndividualInfluence by a fellow conversationalist NPC was denoted as the major factor influencing the change in the uncertainty of Ada (in Discussion 1) and Amy (in Discussion 4), with quotes such as *"She was uncertain to begin with and her groupmate; who was the most knowledgeable (ie if no of prior articles read is an indicator of knowledge); was also wavering her convictions,"* or *"The influence of William's arguments [swayed her]."* 7 of 21 respondents blamed William (*#NPCMentionedUnprompted*) in Discussion 4 for Amy's uncertainty with statements such as, *"I think they were aware of their drift in position and how convinced they were by William's arguments."*

Respondents in Discussion 3 concurred that it was an awareness of an *#OpinionAttitudeDifference* that caused Juan's change in opinion. They quoted, *"He didn't want to seem biased externally so wanted to be portrayed as a centrist; but was privately left-leaning,"* and concurring that *"their view was probably more left-leaning than they initially realized."*

Additionally, for Discussions 1 and 2, both discussions with a smaller group of conversationalists, respondents pointed out that the certainty of the other NPCs helped sway opinions (i.e. *#CertaintyConvinces* stating, *"You must assume this is because of Johnnie's certainty"* or *"The opposition members confidence and articulation was strong."* We believe this smaller number of conversationalists is what influenced both discussions' NPCs to be tagged as *#LackingSupport*. Respondents discussed NPCs having the *"feeling of being marginalized,"* and that they seemed to be a *"lack of support from like-minded people."*

In discussions 2 and 4, both discussions of a longer duration, respondents pointed out that *#GroupInfluence* was a factor in changing the NPCs views. Respondents stated how the *"opposition had convincing arguments or [that there was a] tendency to want to agree with the majority,"* and that there was a tendency for an NPC to cave on their views since they would associate them with *"temporary bias because of peer-pressure in a group of majority conflicting opinions."* Interestingly, in discussions of shorter duration (i.e. Discussions 1 and 3) respondents were more likely to *#InferFactsFrom* the study that was not initially provided to them. They made statements about how the NPC must *"support for innovation and reform strongly"*, or seemed to *"value [the] Rights and Interests"* of the other conversationalists more. These conversations also tended to have stronger *#EmotionsAttributed* to the constituent NPCs with respondents attempting to articulate the emotional distress of the conversationalists saying, *"Changing one's political identity on an issue isn't an easy task and can result in much internal conflict and therefore high uncertainty"* or blaming the *"feeling of being marginalized"*, or that an NPCs *"competitiveness seemed to be declining"* or that an NPC didn't seem to *"care for the well-being"* of the rest of the population.

Discussion

With this section, we revisit the goals of our paper and discuss each. We also discuss our major findings, along with our future work plans.

G1: 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 as 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 and match the mental model of the reader. In the future, these could be used to produce natural language dialogue utterances.

G2: 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.

G3: To extract insights from the study to inform future research

We see these insights in our section on *Analysis of Results*. 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. An interesting point to note is that readers expected NPCs to stand ground and not change their mind in many cases, claiming that this added to the believability of the discussions. Additionally, some 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"*), while others attributed emotions to the NPCs involved discussing NPC competitiveness, or caring for the well-being of the population, or the NPCs support for reform.

Togelius et al. (2013) discuss how game believability is a critical subcomponent of the player experience. It can be linked to a stream of player emotions triggered by events occurring during interaction but also related to cognitive and behavioural process during gameplay. They continue to describe how games with believable elements can elicit emotions in the player. Additionally, several authors argue that the appearance of human intelligence or human-likeness adds value to a computer-controlled character and thus to the quality of gameplay (Togelius et al. 2013; Champandard 2003; Bateman and Boon 2005). We believe that evidence described above of these emotional responses elicited in the player, and the emotions and humanity ascribed to our NPCs can be taken as further evidence of the believability of our system. We suggest this shows that despite the simplicity of our chosen discussion template readers are primed to imagine complex layers of interactions between the NPCs. Our simulation was able to invite users to use their imagination and provide to them a more immersive and compelling narrative effect.

Modeling Social Influence and Simulation

In their responses, respondents pointed out some interesting features of social dynamics that we did not intentionally simulate, attributing changes to these social phenomena. One of those was the existence of an Overton Window in some of the discussions. They pointed out when an NPC changed their mind *"because she was an outlier, and had the most extreme view"* or in another case how *"everyone else expressed a more rightward view; making Ashley's view appear more extreme left that it actually was."*

The participants also pointed out when *#Polarization* seemed to be occurring with the groups clustering away from the centre. One participant stated that *"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,"* while another pointed out that this type of polarization could lead to the feeling that NPCs were *#LackingSupport*, feeling marginalized or as though they were outliers with the participants tending to *"cluster away from centrism."*

Most interestingly, with our analysis of the most believable part of the conversation, we noticed an interesting pattern of how readers discussed *#GroupInfluence*, *#Polarization* and how *#SimilarViewsConverge*. Survey respondents spoke of this as a matter of fact, pointing out how individual members ceded to peer pressure or conformed stating they were *"outnumbered"*, or that an NPC *"was in the minority so probably felt uncertain"*, or how *"deliberation within a group is important and with the right convincing you can change someone's mind"* and that *"there is some power in group mentality"*. We believe these observed patterns further strengthen and support our hypothesis on social rules. Beliefs and attitudes that go against a group's values are looked down upon unfavourably by the members of the group. This would allow with our approach, for the group and cultural rules to emerge organically through the course of interaction with their members.

While these findings were unexpected, we are heartened that our simulation can model and generate these recognizable social phenomena. We believe this further adds to the immersion and believability of the characters.

Conclusion and Future Work

We believe that this research can be an interesting tool not only to increase agent believability but also to evaluate inter-character relationships. Accounting for co-location of the characters in the discussion model allows us to contextualize the interaction and the cultural significance of the conversations. For instance, NPCs of Indian origin, congregating together during a social event can be modelled to discuss views on cricket or rating the latest Bollywood movie they saw. Similarly, NPCs at an office party may share opinions on the work ethics of a colleague leading to a manager being swayed to promote the same.

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 imagine our simulation to work in conjunction with one such as PromWeek (McCoy et al. 2011), where modifiers reflecting the intentions of the participants could be inferred from their expressed opinion and used to generate a dialogue. For instance, [Moderately Left, Gun Control, Legislation] could produce a statement from a virtual character regarding increasing legislation to combat gun violence or implementing strict background checks. Similarly, [Right, Gun Control, Arming] could produce a statement regarding arming teachers in schools or bolstering security in public spaces. and ascribe to the NPCs human-like characteristics of emotion, competitiveness, and intention. Additionally, the discussion mechanic could be used to see how these views of the characters themselves can convince or persuade other NPCs.

In conclusion, 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. Agents with the Lyra model would be able to simulate and model opinions on any aspect of the virtual world they inhabit and then sway one another's mind on the same. We believe our system affords agents the opportunity to reflect on not just their knowledge, but also the certainty of their views. The qualitative analysis of our results shows that Lyra is able to replicate social mechanics such as individual or group influence, feelings of being marginalized, and conformity. We believe reproducing these behaviours in NPCs would improve a player's interactive experience.

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.

Bateman, C., and Boon, R. 2005. *21st Century Game Design (Game Development Series)*. Charles River Media, Inc.

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)*.

Budak, C.; Goel, S.; and Rao, J. M. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80(S1):250–271.

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.

Champandard, A. J. 2003. *AI game development: Synthetic creatures with learning and reactive behaviors*. New Riders.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

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.

Entman, R. M. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57(1):163–173.

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

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

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.

Hickey, G., and Kipping, C. 1996. A multi-stage approach to the coding of data from open-ended questions. *Nurse researcher* 4(1):81–91.

IBM Research. 2018. Ai and bias - ibm research.

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.

Lee, H. 1960. *To Kill a Mockingbird*. J. B. Lippincott & Co.

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.

Loper, E., and Bird, S. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

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.

Mayring, P. 2004. Qualitative content analysis. *A companion to qualitative research* 1:159–176.

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.

Miles, M. B.; Huberman, A. M.; Huberman, M. A.; and Huberman, M. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

Morgan, D. L. 1993. Qualitative content analysis: a guide to paths not taken. *Qualitative health research* 3(1):112–121.

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?

Potter, W. J., and Levine-Donnerstein, D. 1999. Rethinking validity and reliability in content analysis.

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.

Togelius, J.; Yannakakis, G. N.; Karakovskiy, S.; and Shaker, N. 2013. Assessing believability. In *Believable bots*. Springer. 215–230.

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.

Appendices

Thematic Codes

Name	Description	Sample respondent quote tagged with code
#AgreeWithMajority	The respondent tries to infer why a particular change/trend is seen	Left-leaning with centrist views; hence more prone to go with the dominant public opinions
#ArticleInfluence	Noting when a character is swayed by the article	Amy's uncertainty dropping after reading the centrist article and the first round of arguments
#Believable	What they found believable or realistic	That people tended to cluster into ideological groups [was the most believable]
#CausalInference	The respondent causally linked events to explain something	Lashawna swaying slightly more conservative because she had a very convincing and large group and this would easily move her to similar opinion.
#CertaintyConvinces	Noting when a character who is more certain in their views has more influence	That the right leaning group - one of whose members was very certain about his opinion - did budge in their convictions
#ChangedDespiteCertainty	Noting that an NPC opinion changed even though they were certain of their views	I think people get stronger/ more confident in their views after discussing/arguing about them; not weaker.
#ChangedOpinion	Noting that an NPC or a group of NPCs had a change in their opinion(s).	James was swayed by Mary or Shirley
#ClusteringBelievable	Noting that the groups formed in a believable/expected manner	People tended to cluster into ideological groups.
#ClusteringNotBelievable	Noting that the group formation or clustering of opinions was not believable	Mary thinking the group's opinion didn't match their internal attitude was not believable
#DecreasingCertainty	NPC certainty in views decreased / uncertainty increased	The fluctuation from high certainty back to uncertainty in a seemingly short time period
#Disagreement	NPCs disagreeing with each other	Nobody found consensus
#EmotionalResponse	The respondent had an emotional response to the information	It was believable but depressing that nobody ultimately changed their mind at the end of Round 3.
#EmotionsAttributed	Attributing emotions to the NPCs as the reason for an observed behaviour	His competitiveness is declining
#Expected	Was expected. Can be used with other tags. For instance, #IncreasingCertainty #Unexpected	People changing their opinions [was expected]
#ExtremeChangesOpinion	Noting that an NPC had an extreme change in opinion. This tag implies the #ChangedOpinion tag	The extreme left/right fluctuating opinions
#GroupInfluence	Noting when a character is swayed by a group	The fact that James had not changed drastically on his political opinion but has opened up his opinion to uncertainty seems believable since he is out numbered in the group.
#GroupsStandGround	Noting that it's harder to convince groups than individuals, or that groups did not change their mind. #GroupsStandGround =>#StandingGround	The unchanging minds of majority
#IdentifyingSimilarGroups	Identifying a group of people with similar views	People tended to cluster into ideological groups.
#IncreasingCertainty	Noting that an NPC's certainty in views increased or that their uncertainty decreased	Amy's uncertainty dropping after reading the centrist article and the first round of arguments
#IndividualInfluence	Noting when a character is swayed by 1 or 2 individuals (but the individuals are not identified as a group)	Norma and Edward swayed each other
#InferFactsFrom	Infer facts (or make assumptions) that are not given to them by us	The centrists didn't change at all; which doesn't seem characteristic of the topic

Name	Description	Sample respondent quote tagged with code
#LackingSupport	Noting a character is alone/lacking vocal support from other participants	Ashley doesn't fit well with leftist views
#LimitedKnowledge	Explaining/assuming that the effects are related to not knowing enough about a topic	Due to limited knowledge and reading of the matter
#Meta	Talked about the study design; #NotEnoughInfo =>#Meta; #InferFactsFrom =>#Meta	There are too many variables here for me to get a good read on my feelings about these metrics
#MiddleGround	Groups finding a middle ground. Discussion of agreement, or consensus	The other two participants were closer together; so he tried to form consensus in the middle; further from his certain attitude.
#NoAnswer	Participant did not respond or had no meaningful response	I have no idea........
#NotBelievable	What they found not believable or unrealistic	The change in Johnnie stand
#NotEnoughInfo	Noting that the respondent doesn't have enough information to answer a question	...there's also a level of speculation here with limited information on specifics.
#NPCMentionedUnprompted	The behavior of a character was noted when they were not mentioned in the question	Helga started at Left; moved to centrist and then closed at left
#OpinionAttitudeAlignment	Noting that an NPC opinion must have changed since it must not have been aligned with their internal attitude	That he was open to reasoning and reaffirmed his slight left bias
#OpinionAttitudeDifference	Noting the difference between outwardly expressed opinion and internally held attitude	... temporary bias because of peer-pressure in a group of majority conflicting opinions
#OvertonWindow	Noting that a character's views seem more extreme in contrast to others	Because she was an outlier/had the most "extreme" view to the left.
#Polarization	Discussion of groups clustering away from the center	The participants tended to cluster away from centrism.
#PoliticalIdentity	Noting when a character's politics seems to be a part of their identity	Group formations seems coherent with each member's affiliation
#ReceivingSupport	Noting when a character's views are supported by others in the group	Because of the support she saw
#ReinforcedViews	The NPC reinforced their own views	The right-leaning opinions solidified and remained unchanged.
#SimilarViewsConverge	Stating that characters with similar views initially will converge toward one another	People seemed to be swayed by people who were politically similar to themselves
#StandingGround	No change in opinion. The NPC stood their ground	The centrist not changing their opinion
#UncertainMindsChange	Noting when someone who is uncertain changes their mind more readily.	His high level of uncertainty coupled with his moderate stance indicates that Juan felt under informed on the topic.
#UncertaintyStatic	Uncertainty Did Not Change	Shirley's uncertainty [remaining the same]
#Unexpected	Was unexpected. Can be used with other tags. For instance, #IncreasingCertainty #Expected	Even though I didn't expect it; Kenneth's rightward turn is believable.
#UsedPoliticalAffiliationStereotype	Respondent made a stereotypical judgement about a political affiliation.	Ada is a typical right-winger and is looking for viewpoints to confirm her own bias; rather than be convinced by others

Political Scale - Survey Definitions

We wanted to ensure each respondent had a familiarity with the US Political system, and various perspectives associated with different topics in the generated discussions. We provided them each survey respondent with definitions for various terminology on bias and attitudes that they might encounter in the survey. All definitions were taken from AllSides (AllSides 2018) to ensure that the author's bias did was not taken into account

Additionally, respondents were asked to rate their familiarity and positions on the same in order for us to analyse this data for Research Question 1, namely, does the measure of the believability of the generated conversations depend on the personal political biases of the respondents?

Media Bias

This section includes definitions of the concepts of media bias, and media bias.

Media Bias: Media bias is the bias or perceived bias of journalists and news producers within the mass media in the selection of many events and stories that are reported and how they are covered.

Bias: (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.

Media Bias Ratings: Media bias ratings allow us to easily look at a news story or issue from different perspectives. The bias ratings in our dataset are obtained from AllSides.com using a combination of blind bias surveys, editorial reviews, third-party research, independent research, and community votes to calculate media bias of the information.

Left Bias: Left bias is the most liberal media bias rating on the political spectrum. Views with a Left media bias rating are most likely to show favor for government services (food stamps, social security, Medicare, student-loans, unemployment benefits, healthcare, education, etc.), federal laws to protect consumers, the environment, and equal rights. To read more we recommend the following site: https://www.allsides.com/media-bias/left

Center Bias: The view does not predictably show opinions favoring either end of the political spectrum conservative or liberal. A Center view either doesn't show much bias at all, or its bias leans to the left and right equally at different times. It may also mean the author does a good job of portraying both sides equally. It's important to note that sometimes, a conversationalist with a Center rating may miss important perspectives leaving out valid arguments from the left or right. To read more we recommend the following site: https://www.allsides.com/media-bias/center

Right Bias: A Right bias is the most conservative rating on the political spectrum. Some of these sources may be considered "right-wing news." Views with a Right media bias rating are most likely to show favor for decreasing government involvement in economic issues, decreasing federal regulations in general, giving more power to state laws, decreasing government spending, except for defense spending, etc. To read more we recommend the following site: https://www.allsides.com/media-bias/right

NPC views on Topics of Discussion

This section includes descriptions of what typical left-wing and right-wing perspectives on gun control, gun rights, illegal immigration and legal immigration could be.

Left-wing Gun Control and Gun Rights Views - Left-wing views typically argue for more legislation to combat gun violence. Some of these measures include increased background checks, closing the gun show loophole, as well as banning assault weapons and bump stocks.

Right-wing Gun Control and Gun Rights Views - Right-wing advocates typically argue against increasing government regulation and focus on rigorously enforcing current legislation, improving mental health awareness, and bolstering security in public spaces, such as arming teachers in schools.

Left-wing Illegal Immigration Views: - Left-wing views on illegal immigration typically argue that offering a path to citizenship for the 11 million illegal immigrants in the U.S. would lead to higher wages, job growth, and increased tax revenue. They argue it is morally unacceptable to deport undocumented immigrants, particularly since the U.S. was founded by a nation of immigrants, and that they add to the countrys diverse culture.

Right-wing Illegal Immigration Views: - Right-wing views on Illegal Immigration typically argue that these immigrants take jobs from native workers and unlawfully take advantage of government assistance programs, offsetting the taxes they pay. They emphasize the order of law, saying that anyone who wants to come to the U.S. must apply via the pathways we currently have in place.

Left-wing Legal Immigration Views: - Left-wing views on legal immigration typically argue that low-skilled immigration individuals are taking jobs that native-born Americans do not want, filling a major gap in the workforce

Right-wing Legal Immigration Views: - Right-wing views on Legal Immigration Right-wing views generally argue that low-skilled immigrants decrease opportunity for American citizens and depress wages. Many also emphasize that lower education levels and language barriers prevent these individuals from effectively assimilating.