Thesis Proposal

Value-Adaptive Instruction for Promoting More Productive Civil Discourse

Nicholas Diana
July, 2019

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kenneth Koedinger, Co-Chair (Human-Computer Interaction Institute, Carnegie Mellon University)
John Stamper, Co-Chair (Human-Computer Interaction Institute, Carnegie Mellon University)
Jessica Hammer, (Human-Computer Interaction Institute, Carnegie Mellon University)
Sharon Carver, (Department of Psychology, Carnegie Mellon University)
Matthew Easterday, (School of Education and Social Policy, Northwestern University)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Copyright © 2019 Nicholas Diana
Abstract

Civil discourse is our most basic form of civic engagement. In a democracy, it is our best tool for collectively answering a society’s most fundamental question: “What shall we do?” While most of us have no doubt participated in political discussions, engaging in civil discourse that is productive (i.e., dialogue that “fosters democratic goals” [55]) can be substantially more difficult. For example, when both sides of an argument view their beliefs as part of their identity [43], debating the merits of those beliefs without calling into question the merits of the individuals who hold them can be challenging. Ideally, a careful discussant might find common ground, previously obscured by the trappings of tribalism, or, at the very least, foster a mutual understanding of and respect for the values that inform beliefs they do not share. On the other hand, an unskilled discussant, perhaps one only interested in personal attacks or “winning arguments,” will likely only further entrench each party in the views of their political tribe.

Engaging in productive civil discourse is a skill that needs to be explicitly taught, modeled, and practiced in the same way that students are taught and given opportunities to practice skills like finding the length of the hypotenuse or filling in a Punnett square. Unfortunately, civic education often takes a backseat to so-called core subjects like math, science, and English. Learning about civics may be relegated to elective courses, and opportunities to practice civic reasoning skills only afforded to the members of the school’s debate club. And even in these environments where we would expect civic education to be a central focus, the emphasis on argumentation and persuasion, while undoubtedly crucial to civic learning, is insufficient and in some cases counter-productive to kind of dialogue that engenders commonality and collective problem-solving.

In contrast to the status quo, we propose a novel civic education intervention designed to provide students with 1) a better understanding of the values that shape their own beliefs and the beliefs of others, 2) opportunities to practice overcoming the biases that are born out our pre-existing beliefs, 3) an understanding of what makes civil discourse productive and examples of model civil discourse, and 4) opportunities to practice the skills that underpin productive civil discourse. Our proposed intervention is an AI-powered educational game that adapts instruction to a student’s specific values. This value-adaptive instruction is critical because it uniquely allows students to more effectively practice skills like political perspective taking, identifying personal biases, and engaging in discussions that foster democratic goals.
Table of Contents

1 A New Age of Unproductive Civil Discourse 5
  1.1 The Promise of Online Civil Discourse .......................... 5
  1.2 Polarization, Tribalism, and Filter Bubbles ...................... 6

2 Moral Reasoning 7
  2.1 A Brief History of Moral Reasoning ............................. 7
  2.2 Dual-Process Theory .............................................. 8
  2.3 Social Intuitionist Model ......................................... 8

3 Civic Education in America 10
  3.1 Education’s Second-Class Citizen (Whatever That Is) ......... 11
  3.2 Gaps in Civic Education ........................................... 12
    3.2.1 Civic Skills and Dispositions ............................... 13
    3.2.2 The Acknowledgment of Values ............................. 14

4 Prior Work 17
  4.1 Potential Benefits of Value-Sensitive Instruction ............. 17
  4.2 Myside Bias as a Civil Discourse Difficulty Factor .......... 18
    4.2.1 The Critical Evaluation of News Media ..................... 19
  4.3 Estimating and Comparing User and Content Values ............ 20
    4.3.1 Moral Foundations Theory ................................. 20
    4.3.2 Distributed Dictionary Analysis ........................... 21
    4.3.3 Computing Alignment ....................................... 24
  4.4 Study 1: Alignment as a Predictor of Bias ................... 24
    4.4.1 Interaction between Age and Alignment .................... 26
    4.4.2 Performance on Incongruent Problems ..................... 27
  4.5 Study 2: Bias in the Identification of Logical Fallacies .... 28
  4.6 Study 3: Value-adaptive debiasing agent ........................ 28
    4.6.1 Games for Debiasing ....................................... 30
  4.7 Teacher Interviews ............................................... 31
    4.7.1 Focus on Factual Knowledge ................................ 31
    4.7.2 Civics Skills .............................................. 31
    4.7.3 Instigating Conflict ....................................... 31
    4.7.4 Neutral Facilitators ....................................... 31
    4.7.5 Importance of Context/Background Knowledge .............. 31
    4.7.6 Engagement ............................................... 32
    4.7.7 Formal Debates ............................................ 32
5 Proposed Work

5.1 Affordances of Games in Civic Education ........................................ 36
5.2 A Finite Game within an Infinite Game .............................................. 37
5.3 A Brief Overview of the Current Iteration of Our Game ......................... 38
  5.3.1 What the Game Isn’t .............................................................. 39
5.4 The Transformational Framework of Game Design ................................ 40
  5.4.1 High-Level Purpose ............................................................... 40
  5.4.2 Audience and Context ............................................................ 40
  5.4.3 Player Transformations ............................................................ 40
  5.4.4 Barriers ................................................................................. 41
  5.4.5 Domain Concepts ................................................................. 42
  5.4.6 Expert Resources ................................................................. 42
  5.4.7 Assessment Plan ................................................................. 42
5.5 Study 4: Pilot ................................................................................. 43
5.6 Study 5: Full-Scale Experiment ............................................................. 43
  5.6.1 Potential Contributions ............................................................ 44

6 Conclusion ......................................................................................... 45
Chapter 1

A New Age of Unproductive Civil Discourse

“Wow, that internet argument completely changed my fundamental belief system,”
said no one ever.

Where and how we engage in civil discourse is changing. This is in part due to the rapid and widespread adoption of social media. In 2005, the percentage of adults in the United States that use at least one social media platform was 5%; in the intervening 13 years, the number has risen to 69% (in 2018) [60]. This dramatic rise is present across all age groups (37% of Americans ages 65 and up used social media in 2018), but social media is most integrated into the lives of younger Americans. The vast majority of adults (88%) aged 18-29 used social media in 2018, and nearly half (45%) of teens reported being online “almost constantly” [1]. As online communities continue to play a greater role in our lives, more aspects of our offline lives are reflected in those online communities. Social media platforms like Facebook and Twitter often mimic a public forum, where deeply held opinions are expressed and debated. Unfortunately, these changes to our medium for civil discourse have likely affected the quality of that discourse.

1.1 The Promise of Online Civil Discourse

Early internet optimists heralded the then new technology as a revolutionary tool that would bring about a “new Athenian Age of democracy” [29]. Internet evangelists argued that new media would create more trustworthy and impartial news sources, weaken political parties, and drastically reduce the impact of lobbyists and special interest groups (by rendering unrequested advertising ineffective) [69]. This early optimism was supported by computer-mediated communication studies that found that certain features of online communication supported civic engagement ideals. For example, online communication is generally asynchronous, which could provide more time to craft reasoned, thought-out responses [72, 15, 19]. Online communications is also often anonymous, which is most commonly associated with negative behaviors like “flaming” [44], but could also be potentially beneficial to communities whose voices have historically been marginalized in face-to-face communication [2, 7]. For these reasons and others, some have hailed the adoption of social media as the realization of Habermas’ ideal Public Sphere [32], a space for free public discourse and debate about societal issues, accessible to all members of
society because it exists outside of institutional and economic power structures.\(^1\)

In the intervening decades, we have seen the internet realize some of the promises made by early optimists (e.g., access to limitless amounts of information, global communication), but with respect to civil discourse, the impact of new technologies is less clear. Discussions on social media in particular are perceived as less respectful, less likely to come to a conclusion, less civil, less focused on policy debates, and more angry than other places where people discuss politics [21]. The present state of civil discourse may more closely echo the concerns of the so-called “luddite” internet pessimists (e.g., [6]). But while it’s easy to look back from our future vantage point and criticize yesterday’s dreamers, it’s perhaps more productive to ask why their dreams went unrealized (or in the case of trustworthy news sources, became nightmares).

### 1.2 Polarization, Tribalism, and Filter Bubbles

Since 2004, Pew Research has observed a steady increase in political polarization (i.e., a widening of the gap between the ideologies of members of the United States’ two major political parties) [58]. In this polarized moment, more than half Americans (53\%) find it “stressful” to discuss politics with someone they disagree with [59]. This steady increase in political polarization and breakdown in cross-tribe communication may be partially explained by the emergence of filter “bubbles,” or spheres of information that are unrepresentative of the information or perspectives present in the population as a whole.

Historically, filter bubbles have been shaped by social and psychological factors, such as our tendency to associate more frequently with people who share similar viewpoints and our tendency to give our limited attention to news stories that reinforce (rather than challenge) our own worldview [64, 53, 70]. But recent years have seen the introduction of a new source of filter bubbles: curated news algorithms [56]. These algorithms, employed by large technology companies like Google and Facebook, use user data to generate a list of news stories tailored to the interests of each user. Given that user attention is the primary source of profit, the algorithm is incentivized to increase clicks or viewing time. This, coupled with the fact that users are generally more likely to click on or share stories that affirm their worldview (rather than challenge it) means that overtime, a particular user’s news feed will only contain articles that align with their worldview (i.e., articles they are most likely to read). To be fair, this is not unlike filter bubbles driven by social forces, in which the only information we’re exposed to is the information shared by like-minded friends. But one could argue that algorithm-driven bubbles are more insidious. When a polarized friend shares a news story, we might consider what we know about their (potentially biased) perspective when judging the story’s validity. On the other hand, algorithm-curated new sites can give off an air of editorial discretion that conceals their sources of bias, and news feeds found on social media platforms may be perceived as less biased than traditional news outlets with known editorial leanings. A recent user study examining attitudes about news curation algorithms on Facebook found that more than half of participants (62.5\%) were completely unaware of the fact that their news feed is curated by an algorithm [23].

The persistence and power of filter bubbles is, in part, rooted in their ability to exploit a bug in the way we make reasoned judgments, but understanding this reasoning bug first requires a brief history lesson on the academic study of moral reasoning.

\(^1\)In a rare 2010 interview, Habermas himself was skeptical about the promise of the internet as an ideal public sphere, saying the internet “... releases an anarchic wave of highly fragmented circuits of communication that infrequently overlap.”
Chapter 2

Moral Reasoning

“...reason is, and ought only to be, the slave of the passions.”
- David Hume

2.1 A Brief History of Moral Reasoning

Why is reasoning about politics different than, say, reasoning about buying a new appliance? The short answer is that politics is personal and emotional, and microwaves (generally) aren’t. The presence of (and conflict with) emotion when we reason about what is right and wrong (i.e., moral reasoning) has been a subject of discussion for millennia. In his comprehensive book, *The Righteous Mind* [34], Jonathan Haidt provides an excellent overview of the divisive history of moral reasoning, especially as it pertains to his own model (which we also employ). Chronologically, his history begins with Plato, who characterized logic and emotion as two separate souls, one housed in the head, the other in the body, and the latter being far inferior to the former. Haidt calls this the rationalist perspective and defines it as the belief that, “reasoning is the most important and reliable way to obtain moral knowledge.” In other words, moral judgements ought to be left to logic; emotions only taint the process. Haidt argues that the rationalist perspective has dominated discussions of moral reasoning ever since, tracing its influence through Piaget and Kohlberg’s belief that morality in children is self-constructed through experiences (particularly with other children). In contrast to the rationalist perspective, Haidt argues that Hume’s account of logos and pathos is more accurate: “reason is, and ought only to be, the slave of the passions.” The modern re-emergence of Hume’s hypothesis can be attributed to biologist Edward O. Wilson. Haidt [34] summarizes Wilson and Hume’s case against the rationalists in dramatic fashion:

> It seemed clear to Wilson that what the rationalists were *really* doing was generating clever justifications for moral intuitions that were best explained by evolution. Do people believe in human rights because such rights actually exist, like mathematical truths, sitting on a cosmic shelf next to the Pythagorean theorem just waiting to be discovered by Platonic reasoners? Or do people feel revulsion and sympathy when they read accounts of torture, and then invent a story about universal rights to help justify their feelings? (p. 38)

Ultimately, Haidt sides with Hume and Wilson, arguing that moral reasoning is primarily driven by intuitions rather than conscious rational thought. Still, rationality is not completely
absent from Haidt’s model of moral reasoning. Instead, the model captures the symbiotic relationship between two distinct processes.

### 2.2 Dual-Process Theory

For decades, researchers across a wide array of disciplines have demonstrated that there appears to be two general processes for decision-making [26, 57, 42]. The first process is deliberative, conscious, and rational. When we colloquially use the term “reasoning,” we are generally referring to this kind of thinking. But not all decisions are made as the result of a slow, deliberative process. Some decisions are seemingly automatic, made outside of our conscious-awareness. Consider the how quickly you arrive at the answer to 2+2=? compared to the conscious effort required to answer 38+17=? In the first case, the answer requires no computation because you have learned an association between the problem 2+2 and the answer 4. This kind of thinking, which Kahneman refers to as System 1 thinking, is mentally effortless and typically faster than deliberative thought (see [4] for a potential exception), but it has also been proposed as a likely source of cognitive bias [49, 24, 63, 71], and has been linked to societal problems like the perpetuation of stereotypes [20]. In contrast, the second case (38+17=?) likely requires computation because you have not made any such association between the two sides of that specific equation. Instead of drawing on the memory of a fact, you have to consciously engage a computational procedure in order to arrive at the correct answer.

It can be difficult to acknowledge the role of unconscious processes in our lives, and System 1 thinking is no exception. Consider the momentary fear one experiences after an episode of “Highway Hypnosis,” that is, arriving at a destination with no conscious memory of the numerous and complex decisions required to drive there. In response to these feelings of helplessness, we may take solace in the belief that the truly important decisions, decisions that make us who we are, are not the result of unconscious processes, but rather deliberate, reasoned decisions. Amongst those important, identity-forming decisions are our political beliefs. But while one can no doubt conjure some evidence (e.g., facts and figures) to justify their beliefs, Haidt argues that these are post-hoc justifications for an initial automatic intuition.

### 2.3 Social Intuitionist Model

This model of moral reasoning, called the Social Intuitionist Model, argues that almost all moral reasoning (i.e., decisions about whether something is right or wrong) is done within Kahneman’s System 1 thinking. Rational (or System 2) thinking always comes after an initial intuitive judgment has already been made, and only comes online if we are asked to justify our position, or conversely, to challenge the position of someone we disagree with. In short, we are not, by default, the rational thinkers we think we are. Moreover, when we do make use of our capacity for rational thought, it’s generally to justify a decision we have already made, not to search for the truth.

If we subscribe to the Social Intuitionist Model, then reasoned civil discourse is only happening in spaces of ideological disagreement, where people are required to activate their critical, System 2 thinking in order to challenge the assertions of those they disagree with. In contrast, filter bubbles are ruled by intuitions. Because users only encounter information that supports their own worldview and that information intuitively feels true, they are never required to leave System 1. In fact, they have no incentive to engage in the difficult task of critiquing a viewpoint they likely agree with. It only serves to undermine their own self-identity, and could lead to
alienation from the group [41]. And with no out-group members present to challenge the beliefs perpetuated inside the filter bubble, the result is a public sphere devoid of the System 2 thinking required for the critical examination of policy.

Based on these models, the answer to filter bubbles and tribalism seems clear: if we integrate members who hold opposing viewpoints, then they will use their System 2 thinking to hold each other accountable to truth and logic. Unfortunately, if our goal is to have discussants engage in productive civil discourse, merely activating System 2 is likely not enough. This is because our beliefs are grounded in intuitions (i.e., System 1 thinking), not logic (i.e., System 2). To use Haidt’s metaphor: “...moral reasons are the tail wagged by the intuitive dog...You can’t make a dog happy by forcibly wagging its tail. And you can’t change people’s minds by utterly refuting their arguments.” Instead, we are better served by “elicit[ing] new intuitions, not new rationales.” This is, of course, a much more challenging proposition, as it requires a deeper understand of your own values and the values of others, perspective taking, a commitment to shared goals, and a recognition of our own reasoning biases. In the next chapter, we will demonstrate that each of these requirements for productive civil discourse corresponds to a gap in typical civic education instruction. In subsequent chapters, we show that these gaps are largely due to a shortage of resources (time mostly), and how a hybrid solution could leverage the benefits of technology to supplement teacher instruction in order to provide students with the opportunities to practice the skills required for productive civil discourse.
Chapter 3

Civic Education in America

“I know no safe depository of the ultimate powers of the society but the people themselves; and if we think them not enlightened enough to exercise their control with a wholesome discretion, the remedy is not to take it from them, but to inform their discretion by education.”

- Thomas Jefferson

The founders of the United States rightly understood that if the new republic was to survive, its people must receive a formal education in their rights and responsibilities as citizens. Unlike other nations that are bound together by a common ancestry, religion, or history, the citizens of the United States are bound together only by their belief in a set of common ideals. Being a nation united by ideals allows us to reflect numerous rich cultures, but it also means that the defining quality of American citizenship (i.e., an understanding of those ideals) is not innate. As former Associate Justice of the Supreme Court Sandra Day O’Conner put it, “Knowledge about our government is not handed down through the gene pool. Every generation has to learn it...” [66].

The public school system was established, in part, to serve the civic education demands of a new democracy. This recognition of civic education as a key component in the preservation of democracy is made explicit in many early state constitutions. Consider, for example, the state constitution of Massachusetts which includes the following passage:

“Wisdom and knowledge, diffused generally among the body of the people, being necessary for the preservation of their rights and liberties; and as these depend on spreading the opportunities and advantages of education.”

In a their 2011 report [30], the Center for Information and Research on Civic Learning and Engagement (CIRCLE) listed four of the major “opportunities and advantages of education” that the Massachusetts state constitution alludes to:

1. **Democratic Accountability:** Without civic education, citizens would be unable to recognize when the actions of their elected representatives do not reflect their own political beliefs. Moreover, without civic education, citizens would be unaware of their options for recourse against such unrepresentative representatives. Conversely, citizens who are unaware of the ways in which they can have a positive impact on their government will likely become disillusioned and lack confidence in their self-efficacy as a citizen. There is a strong relationship between civic knowledge and engaging in civic engagement (e.g.,
voting, contacting government officials, discussing politics) [10]. An unengaged citizenry leaves a power vacuum that can be filled by special interest groups, whose clear efficacy further perpetuates disillusion with the system.

2. Public Discourse: The report argues that a public discourse centered around basic facts, core ideals, and civility is “the only means of addressing the root causes of vacuous and sometimes viscous dialogue.” Similarly, the news media’s affinity for superficial and sensationalist stories over substantive, policy-focused stories will only be solved by an educated citizenry’s demand for more meaningful media. This 2011 report was perhaps prescient in its recognition of the danger of “outright falsehoods” in the public square, arguing that such falsehoods are perpetuated through appeals to emotion and ultimately prevent productive civil discourse. The authors argue that, “The only way to escape from these vicious cycles is to educate citizens to think critically and demand facts and evidence from the media and their elected officials.” Based on the work validating the Social Intuitionist model, we believe that even this is an insufficient solution. We will return to this and similar claims later in this document.

3. Civic Equality: The affordances of any government are only available to those who understand its structure and function. Madison once wrote, “Knowledge will forever govern ignorance: And people who mean to be their own Governours, must arm themselves with the power which knowledge gives.” As a democracy, it is our responsibility to ensure that all citizens receive such an education. Civic education enables the productive civic engagement of diverse and often marginalized communities. In this way, civic education is required for true civic equality.

4. A Nation of Immigrants: Civic education is the primary means by which new Americans learn the core ideals that form the bedrock of our national identity. While the education of new immigrants is an important part of civic education in the United States, it should not imply that America’s native-born students have an elevated status with respect to civic knowledge. The authors of the report note that “many native-born Americans would fail to pass the citizenship test.”

This is, of course, and incomplete list, but these core benefits are echoed by many organizations that advocate for civics education, such as the Education Commission of the States, which adds that civic education improves school climate and lowers drop out rates [31].

3.1 Education’s Second-Class Citizen (Whatever That Is)

Given the centrality of civic education to the health of our democracy and the rights of its citizens, the current state of civic education in the United States is worrying. Since the 1960s, civic education has deteriorated into a shell of the pillar of democracy that the founders envisioned. Civic learning in the public education system had been historically supported by three civics courses: Problems of Democracy, Civics, and U.S. Government. The first two of these three courses dealt directly with the rights and responsibilities of citizens, and gave students an opportunity to engage with important, real-world problems. It is unfortunate then, that of these three courses, only the last is still commonly found in high school curricula. In contrast to the other courses that are motivated by a goal of participatory civic engagement, the current iteration of U.S. Government more often focuses on a conceptual understanding of the function of government rather than a practical understanding of avenues of political participation.
There is no clearer indicator of the current attitudes about the importance of civic education in America than its absence from federal and state standardized assessments. For example, as of 2016, only 15 states require that students demonstrate proficiency in civics or social studies in order to graduate from high school [62]. Similarly, assessments of civic education are not required by the federal guidelines laid out in the No Child Left Behind Act. While standardized assessments are not without serious problems, they are also a strong determinant of funding. As such, not including civic education in standardized assessments, inevitably results in resources being allocated away from civics and to those other subjects where learning has been deemed more consequential (in the eyes of state and federal legislators).

Given the relegation of civics education to a second-tier priority in the United States’ public education system, we should not be surprised to find a severe deficit in civic knowledge, and consequently, engagement. The 2010 National Assessment of Educational Progress (NAEP) found that three-quarters of students were less than proficient in civics. An example grade 12 “proficient” level skill measured by the NAEP is the ability to, “define the term ‘melting pot’ and argue if it applies to the U.S.” Only 24% of twelfth-graders, students on the verge of entering society as voting citizens, demonstrated this level of proficiency in civics. The 2017 Annenberg Constitution Day Civics Survey found that only 26% of respondents could name all three branches of government, and 33% of respondents couldn’t name a single branch. A staggering 37% of respondents couldn’t name a single right guaranteed by the First Amendment of the U.S. Constitution. The fact that these deficits in civic knowledge persist beyond high school is further evidence that secondary education is our best and only chance for formal civics instruction.

Ignorance isn’t the only negative consequence of inadequate civic education. Many studies have demonstrated a strong link between civic knowledge and civic engagement [10]. People who possess more civic knowledge tend to be more civically engaged than those who do not. Much of this data is correlational, but some instructional interventions have demonstrated the positive impact of civic education on political participation. For example, the civics education program *Kids Voting USA* [50] has been shown to not only increase civics knowledge, but also increase the alignment between the student’s beliefs and potential voting behaviors. Engaging classroom discussions about current and historical events has been shown to increase argument evaluation skills, concern for the treatment of others, and civic efficacy [6] – all of which are crucial ingredients for productive civic engagement.

Without adequate and equal access to high quality civic education, we fail to provide our citizens with the knowledge and skills required to hold their representatives accountable and engage in productive civil discourse. Perhaps more importantly, a deficit of civic knowledge will likely result in reduced civic engagement, as citizens are ignorant of or lose confidence in their own self-efficacy.

### 3.2 Gaps in Civic Education

It is easy to lay modern day problems like increased tribalism and the perpetuation of false news stories at the foot of an inadequate education in civics, but merely increasing the amount of civics instruction (in its current form) is unlikely to alleviate these problems. What follows is a discussion of what we perceive as three gaps in civic education. Following the enumeration of these gaps will be a discussion of our prior and proposed work that addresses some ways in which these gaps might be bridged.
3.2.1 Civic Skills and Dispositions

The first gap is a lack of emphasis on civic skills and dispositions. As mentioned above, civics curricula tend to be overly focused on teaching conceptual knowledge, usually pertaining to the structure and function of government. Consider, for example, the Pennsylvania State Civics and Government Standards [54], which merely asks students to evaluate techniques to address conflicts rather than asking students to practice resolving civil conflicts themselves.

The KLI Framework.

Our instructional design is informed by the Knowledge Learning Instruction (KLI) framework [46], which helps to illustrate why simply asking students to evaluate civic skills is not the most efficient way to teach those skills. The KLI framework argues that the features of a piece of knowledge (or knowledge component (KC)) dictate what mental processes are at work when learning the knowledge and what kind of instruction best supports that learning.

In the KLI framework, types of knowledge components can be distinguished from one another by the contexts or conditions in which they are applied and the response given when they are applied. For some KCs, both the application conditions and response are constant (i.e., there is a 1 to 1 mapping between the condition and the response). This is true for facts and associations (e.g., "What is the Mandarin symbol for house?" or "What is formula for a circle’s area?"). For other KCs, the condition is variable, but the response is constant (i.e., there is a many to one mapping between possible conditions and the response). This is true for category-recognition knowledge. For example, one can be asked to solve for the area of a circle given any number of diameters (i.e., variable conditions), but in each case the response (recalling the formula for calculating a circle’s area) remains the same. Finally, some KCs have both variable conditions and variable responses. For example, if one is asked to find the area of an irregular shape X, one first has to find the area of the regular shapes that make up X. In this case, both the condition and response is variable, so the knowledge component is instead a principle (i.e., “The area of an irregular shape is the sum of the area of the regular component shapes that make it up”).

Consider, for example, being asked to teach someone the following three knowledge components:

1. The Mandarin symbol for house.
2. How to recognize a tumor on an x-ray.
3. What internal validity means.

The first KC only requires that we learn an association between two facts (the symbol and its label). This kind of KC is learned through memory and fluency-building processes. In contrast, the second KC requires that we learn to recognize patterns. This kind of KC is learned through induction and refinement, when a student is given many examples and counter-examples. The third KC requires something more than the previous two: understanding and sense making. The key distinguishing feature of this learning process is its reliance on verbal instruction.

These three types of learning processes (memory, induction, and sense making) are not necessarily independent. Instead, as knowledge components become more complex, they tend to be cumulative. Consider again the task of finding the area of an irregular shape. Executing the principle ("The area of an irregular shape is the sum of the area of the regular component shapes that make it up") requires that one remembers the principle (memory), knows the mathematical rules that govern the principle (induction), and then finally, understands the
principle well enough to solve for the area. Simpler KCs, on the other hand, tend to be less cumulative. For example, you cannot explain why the letter “A” makes the “ah” sound, it is just an association that needs to be learned. The key takeaway is that the kind of instruction that is most supportive of learning is dependent on the kind of knowledge component being taught. If our goal is for students to acquire the skills they need to be productive citizens, then that learning is better supported by opportunities to practice those skills rather than simply asking students to “evaluate” them.

In contrast to the PA State Standards, the College, Career, and Civic Life (C3) Framework for Social Studies State Standards [76] (which is explicitly designed to increase the rigor of state standards), recognizes the importance of practice when learning a skill. It states that students should, “not only study how others participate, but also...practice participating and taking informed actions themselves.” We cannot rightfully expect students to be productive participants in our democracy without providing them with opportunities to practice productive participation. Civics is, in some sense, the applied, practical branch of the Social Studies department, and civics instruction similarly needs to be applied and practical.

3.2.2 The Acknowledgment of Values

The second and third gaps likely result from widespread adherence to the classic, rationalist model of human reasoning. Take, for example, the quote from Thomas Jefferson at the beginning of this chapter. While it’s possible that Jefferson is using a broader definition of “education,” it certainly seems that Jefferson would share the opinion that the way to solve civic crises like the prevalence of fake news, is simply to make citizens better critical thinkers. In this view, the only barrier to correct action is ignorance. All problems can be solved by providing more information and then reasoning about it. This perspective is, perhaps, especially appealing to academics (like Jefferson) because it makes modeling phenomena (like democracies) easier. But these models are, of course, insufficient. They fail to account for the effects of less-than-rational factors like bias and tribalism. If Jefferson was right, disagreements could be solved with facts and figures. This is not to say that Jefferson’s model is entirely wrong. An educated citizenry is still paramount to a functioning democracy, but if our education is going to be effective, we must acknowledge the impact of less rational factors on political decision making.

A good modern example of the continuing influence of the rationalist perspective in civic education is the “classroom debate.” The classroom debate usually involves two students (or two teams of students) who are each assigned to research and subsequently argue for one side of an issue. These classroom debates are excellent instructional tools in a number of ways. First, they give students the opportunity to practice civic skills such as the ability to critically evaluate arguments or defend a position. Second, this practice usually occurs under the supervision of the instructor, a (generally) neutral facilitator, which allows the students to get real-time feedback. For instance, if a student engages in an Ad Hominem argument, the instructor can highlight how this violates a norm. In addition to establishing an environment of mutual respect, teachers can also promote the expectation that public discourse is a truth-finding exercise. Researching an issue isn’t about validating your own particular beliefs, but rather about figuring out what’s true. Finally, classroom debates can center around current issues. Discussing current events is known to be especially engaging for students [8] (though teachers may be hesitant to use current events due to concerns about potential parental backlash).

While it is clear that small doses of facilitated conflict have been associated with gains in a number of skills relevant to civil discourse, classic classroom debates fail to teach many of the skills required for productive civil discourse (i.e., civil discourse that fosters democratic goals).
The shortcomings are most easily demonstrated if we consider the implicit goal of a debate: to persuade your audience (or your opponent) that your position is the correct one and/or that your opponent’s position is (necessarily) the incorrect one. In Jefferson’s perfectly rational utopia, a debate would resolve the ignorance that was the root of any disagreement; by the end of a well-reasoned debate, there would be consensus about what to do next.

However, real-world informal debates are much different. Political discourse is generally not facilitated by a neutral party. Mutual respect is not always an established norm (particularly in anonymous online discourse [44]). In contrast to the truth-seeking orientation of classroom debates, real-world debates are more likely to be motivated by a need to validate one’s own opinion or “win” the argument. Worse, political discourse often results in discusants becoming even more entrenched in their political tribes. Productive civil discourse results in actionable next steps that respect the values of all parties. Outside of Jefferson’s rational utopia, these kinds of political discussions are at best unproductive and at worst destructive.

One method for extending the positive benefits of traditional classroom debates is by using “constructive controversy” [40, 39]. Proponents of constructive controversy argue that while un-guided conflict typically has destructive consequences, guided intellectual conflict can be used as an instructional tool to energize and engage students. While the constructive controversy procedure shares many of the same stages as traditional classroom debates, the authors make a few meaningful distinctions between the two. Debates typically end after students have presented their side of an issue (and perhaps a winner is declared). This may leave students more divided and entrenched at the conclusion of the debate. In contrast, constructive controversy adds two additional steps to the process. First, it requires students to reverse perspectives, doing their best to present the best case for the opposing position. The explicit goal of this perspective-taking exercise is to, “strive to see the issue form both perspectives simultaneously” (p. 41). The final step requires students to “drop all advocacy” and integrate both sides of the issue into a joint position. This step concludes with a conceptual assessment that tests each student’s knowledge about each side of the issue, as well as a procedural self-assessment which allows students to judge how well the group functioned. Finally, and importantly, the group’s efforts and success are celebrated.

The constructive controversy procedure is a massive step forward toward instruction that improves the productivity of civil discourse. It integrates perspective taking elements, facilitates collaboration that ultimately produces an integrated perspective, assesses the productivity of discourse directly, and redefines success, tying it to productive collaboration rather than winning a competition (i.e., a debate). But constructive controversy is still rooted in the rationalist perspective and, as a result, missing a key ingredient for productive discourse: an understanding of the values that inform our beliefs. Take, for example, the perspective taking component of the constructive conflict procedure, which requires students to “sincerely and forcefully” present the opposing side of an issue. Imagine one group of students has just argued for not prosecuting illegal immigrants and is now tasked with arguing for the prosecution of illegal immigrants. These students may reiterate the facts and figures of their peers, but Haidt argues that we cannot truly understand a belief until we can feel, intuitively, that it is true. In other words, sincere perspective taking cannot come from reiterating their opponents case against illegal immigration. Instead, it requires that they understand and actually develop an intuition that “violating the law is morally wrong,” and that they feel that intuition alongside perhaps their own intuition that “caring for less fortunate people is morally right.” By teaching students to identify and empathize with the values that inform beliefs, we 1) make them more effective perspective takers, and 2) give them a valuable tool for engaging in productive civil discourse outside of the formal, structured discussion procedures of a civics classroom.
Acknowledging the values that underly our beliefs and the beliefs of others helps us to address the second gap in civics education: the lack of an explicit emphasis on the knowledge skills, dispositions that combat tribalism. Tribalism allows us to displace our own values in favor of preserving our group identity. Some models of moral reasoning argue that the preservation of social bonds is an incredibly strong determinant of beliefs [41]. They argue that one’s beliefs about immigration may not be grounded in their personal values, but rather a desire to “fit in” with their social group (who may have an established or assumed belief). An adequate civic education should fight against tribalism in two ways. First, it should give students an opportunity to understand what they value in the absence of the social influence of their tribe. Second, it should produce citizens capable of placing shared goals (i.e., the common good) ahead of tribalistic or even individualistic goals.

Acknowledging values also allows us to address the third gap in civics education: a failure to recognize the role of biases in our perceptions and judgements of civic arguments. We believe that these biases play a significant role in the propagation of false information, the development of filter bubbles, and the further entrenchment of citizens into their political tribes. What follows is a deeper discussion of these biases, their impact on informal reasoning, and our work showing how we can use technology to measure bias and aid debiasing instruction.
Chapter 4

Prior Work

“Remember that what you believe will depend very much on what you are.”
- Noah Porter

In the previous chapters, we’ve discussed the negative consequences of tribalism, its root causes, and what we view as three critical gaps in the civic education space:

1. Civics education is narrowly focused on the knowledge about government, rather than the skills and dispositions needed to engage productively as citizens.

2. Fighting tribalism requires an understanding of and respect for values of people who you disagree with.

3. Combatting powerful biases requires instruction that is sensitive to student-specific values

In this chapter, we will present our prior work, which lays the foundation for addressing these identified gaps. First, we will describe in greater detail some of the affordances of instruction that adapts to the specific values of each student. Following that is a discussion of the primary benefit of value-sensitive instruction: more effective debiasing instruction. We will describe the successes and failures we encountered in our effort to estimate user values, estimate values latent in text, and use the relationship between those two sets of values to provide targeted interventions designed to reduce the biased assessment of politically-charged arguments. Finally, we will describe the main lessons learned from a series of structured interviews, and how those lessons inform our proposed work.

4.1 Potential Benefits of Value-Sensitive Instruction

Instruction that is sensitive to each student’s values allows us to measure learning and adapt instruction in new, important ways. It is worth noting that while value-sensitive instruction is new to intelligent tutoring systems, expert teachers have long been adapting instruction to the values of individual students. Imagine, for example, a classroom full of students engaging in civil discourse about a topic, facilitated by an expert teacher. Now imagine a student is particularly entrenched in a certain viewpoint. The normally neutral instructor might temporarily assume the position of the opposing viewpoint in order to challenge the student to reason more deeply about their own beliefs. If we view this interaction through the lens of the social intuitionist model, we might guess that the student might not have ever moved past the intuition-based
System 1 thinking had the teacher not asked them to justify their beliefs (which requires System 2 thinking).

This example illustrates the first benefit of value-sensitive instruction: targeted justification requests. Asking a student to justify their beliefs requires that they engage System 2 thinking. Moreover, Haidt [34] argues that people are unlikely to engage in reasoning about their beliefs unless they are prompted to justify them. Challenging the opposing side to provide reasons and evidence for their beliefs is also central to the efficacy of models like Constructive Conflict Theory.

However, this specific kind of adaptivity is absent in educational technology. Consider the state-of-the-art unadaptive tutoring system. These unintelligent tutoring systems (i.e., systems that are not value-sensitive) must instead ask students to justify a diverse set of beliefs. If the set of beliefs is diverse enough, students will inevitably be asked to justify a belief that they happen to hold. But unintelligent systems are unable to distinguish these specific belief-alignment events from other instances. If the ability to justify your own beliefs is a skill we are interested in measuring, it is crucial that intelligent tutoring systems have some prior knowledge of the student’s beliefs.

Recall that the Social Intuitionist Model suggests that our ability to justify our beliefs with evidence, while important for civil discourse, is disconnected from belief formation and revision. Students who are prompted to justify their beliefs are unlikely to reconsider their position and change their mind. Justification is post-hoc in nature. As such, the dialogue is much more likely to move towards a discovery of shared values and actionable solutions if the discussants focus on the real reasons they believe what they believe: their foundational values.

As we’ve discussed above, when our unintelligent tutoring system asks students to justify arguments, it cannot distinguish between arguments that align with the student’s beliefs and arguments that do not. We’ve noted that being able to know when a student is justifying their own beliefs is essential for measuring their ability to use evidence to support their arguments. However, it is equally, if not more important, to know when student is justifying beliefs that are not their own. This requires students to engage in the second key benefit of value-sensitive instruction: targeted perspective taking.

Perspective taking is present in both the C3 and CIRCLE civic education standards we’ve reviewed above [30, 76], but notably absent from the PA State Standards [54], which make no mention of empathy or perspective taking as important skills for civic life. Value-sensitive instruction is one way to bridge the gap between the aspirational C3 standards and the insufficient state standards.

### 4.2 Myside Bias as a Civil Discourse Difficulty Factor

The third, and primary benefit of value-sensitive instruction is its potential ability to measure myside bias. Myside bias is the formal name for our tendency to evaluate arguments more favorably when they align with your own views or beliefs (and conversely, more critically when they do not) [71]. As we’ve mentioned previously, filter bubbles exploit this weakness in human reasoning to protect themselves from any critical thought that might pop the bubble. A related domain where myside bias may play an even greater role is in the acceptance of false or misleading news stories as reliable and valid pieces of news. News media has a direct relationship to civil discourse, where it functions as the fodder of discussion. But the consumption of media can also be framed as a sort of discourse itself, in which the media creator and the media consumer are the discussants. In this frame, the news media presents their argument (in the form of a
news story), and the media consumer must think critically about the argument to determine its validity and value. Unlike civil discourse between two people, this is a one-turn interaction, but this framing can be useful for modeling the potential impact of myside bias on civil discourse, and how value-sensitive systems might help mitigate myside bias in reasoning about news media specifically, and in civil discourse more broadly.

### 4.2.1 The Critical Evaluation of News Media

The rise of social media has been accompanied by a rise in smaller, decentralized media sources. One clear negative consequence of this democratization of media has been an increase in access to unreliable or misleading news stories. Despite their lack of credibility and veracity, these stories are persuasive and appealing. Some estimates suggest that Americans fall for fake news headlines approximately 75% of the time [68], and that these stories are generally more engaging than stories produced by traditional news outlets [67].

The proposed solutions to these problems generally fall into two categories. The first category leverages various machine learning methods [16] to create “fake news detectors.” While some of these classifiers are quite sophisticated [78], these detectors tend to limit their scope to the detection of stories that are patently false. More nuanced instances of stories that are merely misleading are generally beyond the purview (and perhaps ability) of these systems [51]. Moreover, even if accurate classification was possible, one might question whether it is in our best interest to delegate this task to machines, potentially allowing our own ability to critically evaluate media sources to languish in the process.

In contrast to the content-driven detectors, other solutions focus on improving the critical thinking skills of the media consumers themselves. There is certainly evidence of a deficit in this regard. A recent study of students in middle school, high school, and college summarized the student’s “civic online reasoning” (e.g. evaluating arguments, recognizing spin) as simply, “bleak” [79]. Non-detector solutions tend to focus on strengthening these kinds civic reasoning skills. For example, Factitious is a game created by the American University Game Lab [37] that is marketed as a way to test the player’s ability to distinguish fake and real news stories, but along the way teaches the player to identify features like reliable sources and neutral language.

While the detectors focus on the media content itself (hoping to fill the role of editor in the new democratized news space), the civic education solutions focus instead on the media consumers, with the hope that better critical thinkers might be more or less immune to the appeal of misleading content. Both of these approaches unfortunately tend to neglect the dynamic relationship between the media content and the media consumer. Consider the following actual fake news headline:

“Pope Francis Shocks World, Endorses Donald Trump for President”

If you happen to be a religious Trump supporter, this story may seem plausible. After all, if you, a person of faith, have found reason to endorse him, why shouldn’t another person of faith. This headline confirms what you already believe to be true. However, if you are not a Trump supporter, this headline might raise several red flags. It is in that wave of skepticism that you may dart your eyes to the URL in order to check the credibility of the source. Because this headline runs counter to your beliefs, you go searching for evidence to disprove it. In either case, the degree of critical thought that is brought to bear on the content is, at least to some extent, dependent on the values and beliefs of the reader.

Misleading and false news stories can exploit this vulnerability by designing stories that strongly align with the prior-held beliefs of the target audience. Because the reader wants to
believe the story is true (to affirm their reality), System 2’s critical reasoning skills are never engaged. The bias literature suggests that overcoming the strength of this intuitive appeal may require more than detecting falsehoods or training consumers to be more critical. Solutions that ignore the dynamic relationship between the user’s beliefs and the beliefs latent in the misleading media content are perhaps ignoring the very feature that makes the target content so powerful.

### 4.3 Estimating and Comparing User and Content Values

Accurately capturing user beliefs is a daunting challenge. Each user likely possesses countless individual beliefs, and new beliefs are constantly being created in response to their current political context. Consider the following scenario:

Sam secretly voted against his wife in a local beauty pageant. Is Sam a good husband?

We might expect most people to answer, “no,” but what this exercise is really meant to illustrate is just how easily and quickly we can generate completely new beliefs (i.e., “I believe that Sam is a bad husband”). Beliefs are too specific and numerous to incorporate into a student model. Instead, we measure the foundational values that theoretically inform our beliefs. For example, I couldn’t possibly know if you, the reader, would think Sam is a good husband, but given the fact that most people value loyalty, and that voting against your wife is incongruent with that value, it is safe to assume that most people would consider Sam a bad husband. In other words, if we have some knowledge about a user’s values, we can use those general values to estimate the user’s more specific beliefs.

#### 4.3.1 Moral Foundations Theory

Moral Foundations Theory [35] argues that our moral decision making is rooted in a small set of foundational values (Care, Fairness, Loyalty, Authority, and Sanctity). The above scenario about Sam and the beauty pageant is adapted from a larger set of empirically validated Moral Foundation Vignettes [13], which are short scenarios designed to evoke a specific moral foundation. See Table 4.1 for more information about the five well-established moral foundations.

Research has demonstrated that different subsets of the population weight these five foundational values differently in their moral decision-making. For example, American liberals tend to weight Care and Fairness much more strongly than the other three foundations. Haidt notes that this emphasis on care and fairness matches the relatively limited scope of most Western philosopher’s accounts of morality [34]. In contrast to American liberals, American conservatives tend to have a more even distribution of weights across the five foundations, with generally less weight placed on Care and Fairness than liberals, but more weight placed on Authority, Loyalty, and Sanctity. The values of American conservatives, Haidt argues, more closely match those seen in non-Western traditions. They are less individualistic and more collectivist, have a greater respect for traditions, and are more motivated by ideas like spiritual purity. Haidt argues that differences in beliefs and opinions are just manifestations of more fundamental differences in

---

[1] The authors of Moral Foundations Theory do not claim to have discovered all of the potential foundational values that shape our moral judgments, but the validity of these five is supported by a large amount of empirical research. Recently, the authors have proposed an additional sixth foundation, Liberty/Oppression, which may be incorporated into future iterations of the theory.
the relative importance of foundational values. Consider, for example, the issue of illegal immigration. Through the lens of Moral Foundations Theory, we might expect American liberals, motivated by the Care foundation, to be more lenient towards an impoverished immigrant, particularly if they are fleeing violence. To an American liberal, allowing an immigrant to break the law in exchange for their well-being is the more moral thing to do. Conversely, American conservatives, motivated by the Authority foundation, will likely be less lenient towards illegal immigration, which, by definition, violates the law. To an American conservative, upholding the rule of law is critical to the preservation of our civilization’s social contract, and not allowing cracks to form in that contract is the more moral choice.

While these are the more respectable motivations for views on illegal immigration, more unsavory appeals are often made to other foundations. For example, anti-immigration messaging is often laced with a contemptible subtext that suggests that immigrants are dirty and will destroy sacred American values. This subtext is both repulsive and powerful, as it strongly evokes the Sanctity foundation (which is associated with concepts like cleanliness, purity, and desecration of the sacred). The intuitive response evoked by this kind of messaging can (and ought to) be challenged, but the Social Intuitionist Model suggests that any internal challenge is unlikely. If the intuition aligns with one’s worldview, they are unlikely to leave System 1 thinking, and if System 2 is engaged, it may only be used to justify the initial intuitive response. Recall that our intuitions and rational are most commonly challenged by another person (who usually disagrees with you). Perspective taking allows us to more accurately simulate the opponent’s challenges in our head, and hopefully, hold our own intuitions accountable.

The relative importance we give to each moral foundation when making moral judgments can be estimated using the Moral Foundations Theory Questionnaire (MFQ). This 30-item questionnaire developed by the authors of Moral Foundations Theory asks participants to respond to Likert-scale items relevant to each of the five foundations. For example, participants are asked to indicate the degree to which they agree with the following statement: “Respect for authority is something all children need to learn” (which is obviously relevant to the Authority foundation. The final output of the questionnaire is a vector of five scores that indicate the relative importance of the five moral foundations to the participant’s moral decision making. These moral foundations have been empirically shown to be highly predictive of both general voting behavior [27] as well as more specific political beliefs (e.g., “Climate change is real”) [47, 65]. Ultimately, we are interested in constructing a model that relates the values latent in text to the values and beliefs of an individual person. This vector of five scores represents the human side of that relationship. Moral Foundations Theory allows us to approximate beliefs in a theory-driven, context-general way.

### 4.3.2 Distributed Dictionary Analysis

We discussed how we can use Moral Foundations Theory to derive a measure of user values (as a proxy for beliefs), but what a value-sensitive system really needs to know is how the user values relate to content’s implicit values. To measure this alignment between user and content values, the system must also be able to estimate the values latent in the text the user is reading. Historically, this has been done by developing a large list of words that are semantically similar to the target concept (i.e., a dictionary), and then counting the number of times a word in that dictionary appears in the text you are examining. This solution has several drawbacks that make it a less than ideal choice for our context.

First, this approach is only effective for analyzing large bodies of text. This is because smaller bodies of text (e.g., news headlines, tweets, etc.) may be highly relevant to the target
Figure 4-1: Relevance to Moral Decisions by Moral Foundation for more conservative and more liberal participants in one of our studies. These values closely match previously observed values for liberals and conservatives [35], suggesting that our recruitment pool is politically diverse.

concept of interest, but nevertheless happen to not contain any of the terms in a target concept’s dictionary. One potential solution to this scaling problem is to increase the variety of words in the dictionary, which increases the chance of relevant dictionary terms appearing in smaller bodies of text. This solution also has major drawbacks. As the size of the dictionary increases, we would expect the semantic distance from the core meaning of the target concept to increase as well. Adding more terms to the dictionary increases breadth, but causes the meaning of the target concept to become less precise.

Another potential limitation of any methodology that requires the manual creation of a concept dictionary is obsolescence. While some (perhaps most) concepts are relatively static (semantically), concepts that are intrinsically tied to our culture (such as those related to political discourse), may be more semantically dynamic. For example, if we wanted to create a dictionary for the concept “evil,” we might include a word like “wicked” in the dictionary. While this would be a perfectly reasonable choice throughout most of history, it would likely conflict with the positive connotation that has entered the vernacular in the past decade (or since the 1960’s if you’re from New England) [17]. One solution to this so-called lexical drift is to adopt a more data-driven approach, where the meaning of words is linked to their colloquial usage in a real-world, contemporary text corpus.

Distributed Representations do just that. In contrast to word-frequency methods, distributed representations [52] estimate the meaning of words by comparing the numerous, varied contexts that the word appears in within a large text corpus. These models are rooted in the distributional hypothesis, which states that words that appear in similar contexts likely share some semantic features. For example, consider the following two sentences:

“The apple she picked was juicy.”
“The orange she picked was juicy.”

Given that the two concepts (apple and orange) appear in such similar contexts, apples and oranges likely share some properties (e.g., both are juicy and pickable). Other properties, like texture for example, are not shared, but we would expect words like “smooth” to appear more
<table>
<thead>
<tr>
<th>Foundation</th>
<th>Related Concepts</th>
<th>Example Vignette</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care &amp; Harm</td>
<td>kindness, gentleness, and nurturance</td>
<td>You see a zoo trainer jabbing a dolphin to get it to entertain his customers.</td>
</tr>
<tr>
<td>Fairness &amp; Cheating</td>
<td>justice, rights, and autonomy</td>
<td>You see a runner taking a shortcut on the course during the marathon in order to win.</td>
</tr>
<tr>
<td>Loyalty &amp; Betrayal</td>
<td>patriotism and self-sacrifice for the group</td>
<td>You see the US Ambassador joking in Great Britain about the stupidity of Americans.</td>
</tr>
<tr>
<td>Authority &amp; Subversion</td>
<td>leadership/followership, deference to authority, and respect for traditions</td>
<td>You see a woman refusing to stand when the judge walks into the courtroom.</td>
</tr>
<tr>
<td>Sanctity &amp; Degradation</td>
<td>disgust, contamination, purity, and holiness</td>
<td>You see a man in a bar using his phone to watch people having sex with animals.</td>
</tr>
</tbody>
</table>

Table 4.1: The five well-established foundations of Moral Foundations Theory, some key concepts that are related to each foundation (adapted from the framework’s website, http://www.moralfoundations.org), and an example vignette designed to evoke the foundation (adapted from [13]).

often in the context of apples and “bumpy” to appear more often in the context of oranges. When given the many, diverse contexts provided by the text corpus, the model is able use the relationships between a target concept and the contexts in which it appears to approximate the meaning of the concept. While the notion of distributed representations has existed for some time [36], recent implementations (such as the Word2Vec [52] methodology employed in the current study) have demonstrated the effectiveness and efficiency of the method (in terms of computational cost). Mikolov et al. [52] compared their modern method to the other state-of-the-art methods at the time (e.g., feed-forward and recurrent neural network language models) by asking the models to solve simple semantic questions about the analogical relationships between sets of words. For example, a sample semantic question might be: “France is to Paris, as Germany is to ___ ___ ___” where the answer is “Berlin.” They found that their skip-gram model out-performed all other models when answering these kinds of semantic questions.

The distributed representation of a word is simply that word’s location in a low-dimensional (10-10,000 dimensions) space. This location can be represented as a vector, which allows us to compute the semantic distance between two concepts using cosine similarity. Mikolov and colleagues found that these questions can be answered using distributed representations (i.e., vectors) by computing the difference between the vector representations of the first set (vector(Paris)-vector(Paris)) and adding the vector representation of one of the concepts in the second set (+vector(Germany)). In essence, the resulting vector representation (X) contains features of the concept “Germany” as well as features of the concept “Paris” that are left after we took all of the “France” features out of it. Vector X exists as a concept in the low-dimensional semantic space, so we can determine the correctness of the model’s answer to this question by
using cosine similarity to find the nearest (i.e., most similar) concept to vector \( X \). If the closest concept to vector \( X \) is the concept “Berlin,” then the model has answered the question correctly.

Garten and colleagues [28] extended this work in distributed representations to incorporate concept dictionaries. A distributed dictionary representation is computed by simply averaging the distributed representations of all the words in the dictionary. The result is a point in the semantic space that amplifies the shared, core features of each of the component dictionary terms. Because we are ultimately using an abstract representation of a concept, our dictionaries can be highly focused, including only the most relevant terms. Distributed dictionary representations mitigate the two major drawbacks of word-frequency methods. First, because the method calculates the semantic distance between the body of text and the target concept, it does not require that any of the dictionary terms actually be present in the text. This allows for the effective analysis of small bodies of text. Second, because the distributed representations are built using a text-corpus, the estimated meaning of words will be true to the words’ contemporary meaning, so long as the text-corpus is contemporary. In our analysis, we use the pre-trained Google News corpus (approximately 100 billion words) Word2Vec model\(^2\), and a Python implementation of Word2Vec [52] called \texttt{gensim}.

### 4.3.3 Computing Alignment

Having established a theory-driven method for estimating student values and a data-driven method for estimating the values latent in content, the next step is establishing a method for relating these values to one another. We call the extent to which the student’s values align with the values latent in the content \textit{Alignment}. Recall that the output of the Moral Foundations Theory Questionnaire is a vector of five values, representing how relevant each foundation is to a specific student’s moral decision making. After we have this estimation of student values, we compute the cosine similarity between the text content and each of five moral foundation concept dictionaries using the distributed dictionary representation analysis described above. This also results in a vector of five values (one per foundation) that reflect the values latent in the text content. To generate an \textit{Alignment} score, we simply compute the cosine similarity between the student’s vector of values and the text’s vector of values. The resulting score is a number from 0 to 1 that reflects the extent to which the student and text values align. Finally, we use a normalized log-transformation to correct for skew. Alignment should be computed for each student/content combination.

### 4.4 Study 1: Alignment as a Predictor of Bias

If this measure of alignment is a sufficiently good estimate of values (both user values and the values latent in text), we would expect that \textit{alignment} would be predictive of myside bias. We expect to see more bias in situations where the user’s values align with the values in the text (i.e., high alignment). To test if alignment is indeed predictive of bias, we conducted an argument evaluation experiment. We hypothesized that, when asked to evaluate the strength of politically charged arguments, participants would rate arguments as stronger when there was a high degree of alignment between the participant’s values and the argument’s values.

Sixty (\( n=60 \)) participants were recruited from the online participation platform \textit{Prolific} and asked to read and rate the strength of 20 arguments on a nine-point Likert scale (1=Very Weak; 9=Very Strong). Each argument had three key features. First, each argument was designed to

\(^2\text{The pre-trained Google News model can be found here: https://code.google.com/p/word2vec/}\)
evoke a specific moral foundation. For example, the following argument was designed to evoke the Authority foundation:

Greenville School District requires students to address all adults as “Sir” or “Ma’am” and their students always score higher on state tests than ours. Instilling a strong respect for authority for their teachers helps students learn.

Regardless of the argument’s actual strength, we would expect that if a participant believes that respecting authority is important, this argument will resonate with them. Each of the five foundations is the focus of an argument four times, for a total of 20 arguments.

The second key feature is the relative quality of an argument. This is a categorical feature with two levels, high quality and low quality. The above argument is an example of a low quality argument. In contrast, consider the following argument:

The number of suspensions at Redbridge School District has been slowly increasing for the past 5 years. Last year they added three police officers to their staff and saw a 10% decline in suspensions. The presence of a strong authority figure reduces bad behavior.

While this argument is certainly not airtight, it has several attributes that make it a relatively higher quality argument. First, it shows the reversal of a long-term trend, in contrast to the low quality argument where no temporal context is established. Second, it uses concrete figures that are relative to the norm, as opposed to the low quality argument which uses vague terms like “higher” to quantify changes. In general, high quality arguments include information that can be used to rule out some alternative explanations. Low quality arguments leave open the possibility of alternative explanations. Of the 20 arguments, half are high quality and half are low quality.

The third key feature is congruence with the target foundation. A potential limitation of the distributed dictionary representation methodology (described below) is that statement representations are formed using the representations of single words. This means that, while this methodology should have no problem knowing that the word “son” in the context of the word “king” likely refers to the concept “prince,” it will likely have more difficulty identifying the cultural nuances between statements like “God is good” and “God is dead.” The congruence feature is designed to test the robustness of this methodology’s ability to adapt to these kinds of unfavorable circumstances. Consider again the two previous example arguments. Both arguments 1) use language that evokes the authority foundation, and 2) are supportive of that foundation. In contrast, consider the following argument:

Woodford School District doesn’t allow teachers to reprimand students, and last year they had fewer detentions than our district. Students behave better when they’re treated like equals instead of children.

While this argument also evokes the Authority foundation, this example argues against an increased respect for authority. We would expect that participants that value authority will be more skeptical of the claims in this argument, because they violate their intuitions. Whether the model’s representation of the values latent in the argument is nuanced enough to make the distinction between incongruent and congruent arguments is an open question. Again, half (10) of the arguments are congruent, half incongruent.
We used the following mixed effects model to determine the impact of alignment on ratings of strength (i.e., the impact of bias):

\[
\text{rating} \sim \text{quality} + \text{alignment} + (1|\text{participantID}) + (1|\text{argumentID}) \tag{4.1}
\]

Where \text{rating} is ratings of argument strength, \text{quality} is argument quality (high or low), \text{alignment} is the alignment between user and content values, \((1|\text{participantID})\) is a random effect for participant, and \((1|\text{argumentID})\) is a random effect for problem. It should be noted that although participants on average rated high quality arguments as significantly stronger \((t(59) = 8.07, p < .001)\) than low quality arguments \((M = 5.06, SD = 1.72)\) (suggesting some categorical validity), the labels “high” and “low” are very much subjective labels. As such, we cannot objectively compare the impact of \text{alignment} to the impact of quality. Still, we can make a meaningful, subjective comparison between the impact of \text{alignment} and “quality” (as operationally defined in this context). In this context, the impact of \text{alignment} on ratings of strength \((\beta = 3.06, p < 0.001)\) was greater than the impact of \text{argument quality} \((\beta = 1.33, p < 0.01)\).

Figure 4-2: Relative impact of alignment on the ratings of high and low quality arguments. Each data point represents the average rating and alignment for all arguments within a category (high or low quality) for one participant. On average, participants rated high quality arguments as stronger than low quality arguments. The ratings of both types of arguments were associated with alignment scores.

4.4.1 Interaction between Age and Alignment

Previous research suggests that, because reliance on heuristic reasoning increases with age, older adults may be more likely to exhibit biases in everyday reasoning [45]. To test whether this was true of our sample, we built a mixed effects model with \text{participant} and \text{argument ID} as random effects, ratings of strength as the outcome variable, and \text{argument quality} and
alignment*age as fixed effects (where alignment*age is an interaction term). We found that there was a significant interaction between alignment and age ($\beta = 15.01, p < 0.001$), such that alignment’s impact increases as age increases. This finding aligns with previous research. Additionally, this alignment*age interaction model had a better fit (AIC=5033.05) than the previous model built without the interaction term (AIC=5058.63).

### 4.4.2 Performance on Incongruent Problems

A potential limitation of this particular NLP method is its reliance on the semantic relationships between isolated words. A robust methodology should be able to accurately determine the valence of an argument that may contain several words related to a foundation, but nonetheless is incongruent with the beliefs of someone who values that foundation. To test the robustness of our method, we built another iteration of the above, best performing mixed effects model (including the alignment*age interaction), but selected only incongruent arguments (previously both congruent and incongruent problems were used). The impact of alignment on ratings of incongruent arguments also appears to be dependent on age, as the interaction term alignment*age was again a significant predictor of ratings of argument strength ($\beta = 15.01, p < 0.001$). To examine this relationship further, we divided the sample into two groups (older and younger) along the mean age, and then calculated the correlation between participants’ mean ratings and mean alignment for each group. While we found a significant correlation between ratings and alignment in the older group ($r = 0.26, p < 0.001$), we found no such correlation in the younger group (see Figure 4-3).

These results suggest we can estimate when a user might be susceptible to myside bias
by relating theory-driven estimates of user values to data-driven estimates of text values. As we’ve mentioned above, this has obvious implications for instruction aimed at reducing bias, but knowing when a user might be susceptible to bias is only useful if we can then give targeted interventions that reduce bias.

### 4.5 Study 2: Bias in the Identification of Logical Fallacies

The process of reducing or eliminating the impact of bias on reasoning is referred to as debiasing. Debiasing has been hailed by some as “among psychology’s most enduring legacies to the promotion of human welfare” [48]. Despite the extensive work identifying and measuring the impact of biases on reasoning [71, 24, 77], work on debiasing is limited [48]. This is likely because developing debiasing interventions that are effective has been shown to be challenging [25].

Despite these challenges, our previous work in the domain of informal logical fallacies suggested that an effective intervention may only require the priming of the user’s System 2, critical thinking faculties. In a series of two experiments, we found that participants had more difficulty identifying informal logical fallacies in arguments that aligned with or supported their own political beliefs. However, we also found that the effect of bias diminished with practice (see Figure 4-4).

There are two possible interpretations for the diminishing impact of belief bias with practice. First, it is possible that an improved understanding of the logical fallacies makes the fallacious features of an argument more salient. If this is the case, then reducing belief bias is a matter of better training in argument evaluation. However, it is also possible that it’s not learning that is reducing belief bias, but rather that some typically dormant critical thinking faculties are coming online (as the task requires them) and overpowering the influence of belief bias. This interpretation seems to support the main assertion of Haidt’s Social Intuitionist Model of moral reasoning [33], which argues that everyday moral reasoning happens quickly and is primarily based on intuitions (as opposed to a rational assessment of the argument). Rationalization enters into the model after a moral decision has been reached, to justify the decision (or conversely, to undermine an opposing position). With respect to the current experiment, it is possible that the belief bias effect seen early in the experiment is evidence of an intuitions-based moral reasoning, and performance improves as participants discover that the task requires rational reasoning. If this interpretation is correct, then performance on the earlier problems is representative of how we typically evaluate everyday arguments (i.e., in the absence of heightened critical thinking). Moreover, the difference observed between the High Alignment and Low Alignment groups on these early problems suggests that being susceptible to belief bias may be the typical case.

If this second hypothesis is true, then mitigating belief bias in everyday reasoning may not simply be a matter of better training in argument evaluation. Instead, systems designed to combat our susceptibility to weak arguments or misleading news stories should place a greater emphasis on understanding the user’s beliefs and how those beliefs 1) relate to the beliefs present in the content they are consuming, and 2) impact their judgment of that content’s validity.

### 4.6 Study 3: Value-adaptive debiasing agent

Emboldened by our success in measuring alignment, we returned to the same experiment paradigm outlined in Study 1, but added two additional intervention conditions. As in Study
Figure 4-4: We found that users were biased by their beliefs, as evidenced by the differences between early performance on high alignment and low alignment problems. Alignment group was determined by score (1-5) on the self-reported alignment questions (High = 4 or 5, Low = 1 or 2). These results suggest that belief bias impacted performance early in the experiment, but that the effect diminished with practice.

1, all participants were asked to evaluate the relative strength of a 20 politically charged arguments. However, in this experiment, some participants were shown the following intervention message alongside 10 of the arguments. In the adaptive condition, participants were shown this intervention message alongside the 10 arguments that most closely aligned with their own values. In this random condition, participants were shown this intervention message alongside 10 random arguments. Including both an adaptive and random condition allowed us distinguish between effects of the intervention and effects of the adaptivity. Finally, one third of participants were assigned to the control condition, in which the intervention message never appeared.

We found no significant differences between groups. Participants who saw the intervention were no less biased than those who did not. We re-ran the experiment several times, varying in the intervention message along a number of what the literature suggests might be key features. For example, we thought that perhaps participants needed to be made aware of bias to appreciate the intervention, so we presented this more explicit message:

Warning: this argument might align with your values (and bias your response)!
Think about this argument carefully, then click here when you’re ready to respond.

Again, we saw no difference between conditions. Results from one iteration of the experiment suggested that the bias effect was strongest for arguments the participant likely disagrees with. In other words, people weren’t rating arguments they agree with. In our most recent iteration, we reduced the number of arguments in which the intervention appears (5 rather than 10). We hypothesized that, in labeling 10 out of 20 arguments as high-alignment, we may be inadvertently selecting arguments that the participant does not agree with, which undermines the credibility of
the intervention system. By only selecting the top 5 arguments, we give the intervention a better chance at being perceived as credible. We also added questions to the post-test questionnaire that directly address the efficacy and perceived credibility of the intervention (which of course should have been included in the first iteration).

As with the previous studies, we found no clear differences between conditions. However, the post-test questions did provide some new insights. First, only half of the participants who saw the intervention message said that it changed how they rated the argument. Worse, these supposed changes in ratings don’t show up in the data. When we included an interaction term between alignment and a binary variable indicating whether the participant said the intervention changed their rating as a feature in a regression model predicting ratings of argument strength, it was not a significant predictor. In other words, participants who said the intervention changed their opinion were no less biased than those who said it didn’t. Furthermore, half of the participants said it had no effect on their ratings. Open-ended questions that asked for reflections on the intervention suggest that some participants may have had a negative reaction to the notion that they might be biased. This points to the presence of another phenomenon, bias blindspot, which is our inability to see or accept our own biases (but ability to see them in others) [61].

The second insight from the post-test questions was more promising. When participants in the adaptive condition were asked, “Do you think the warning message generally showed up on for the problems you disagreed with the most?” exactly half (11) said yes and half (11) said no. However, when participants in the random condition were asked the same question, the vast majority (13) said no with only 5 participants saying yes. These differences in frequencies were not significant in a Chi-square test, but they are certainly trending in a direction that supports the ability of our model to detect arguments that people disagree with. Still, a success rate of 50% (as seen in the adaptive condition) is nothing to be proud of.

Despite any small insights we gained about the efficacy and perception of our model, the main takeaway from these experiments is still: Our interventions don’t work. And while it’s possible that we just haven’t yet discovered the exact recipe for an intervention that could reduce bias, it’s also possible that our difficulty in discovering it suggests that another solutions may be more effective and robust. An effective solution likely requires that users “buy into” 1) the power of bias, 2) the fact that they may be susceptible to bias, and 3) that a computer agent might be able to help them recognize when they are susceptible to bias. This may, in turn, require learning, vulnerability, and trust – none of which are likely to happen in the short timeframe of an experiment.

4.6.1 Games for Debiasing

However, each of those prerequisites (learning, vulnerability, and trust) can be fostered inside an educational game. Games can also scaffold personal vulnerability by allowing players to make their characters vulnerable instead, allowing them to simulate vulnerability and observe its consequences from a safe emotional distance. Finally, games operate with a level of trust in the faithful execution of the game’s rules. Game mechanics are expected to be consistent and predictable (or at least predictably random), so integrating our model into a game as a mechanic might lend it credibility, which is essential to its success. The model might also benefit from the lower stakes of a game. Telling a person that a machine can predict their beliefs is probably unsettling, perhaps even causing the intervention to backfire (i.e., users respond in defiance against the computer agent in an effort to assert their autonomy). Players may be more willing to change a biased choice in enclosed, fictional world of a game than admit that they are biased in real life.
4.7 Teacher Interviews

To inform the design of our educational game, we conducted a series of semi-structured interviews with five local Social Studies teachers. After the interviews were conducted, we used an affinity diagram to group together similar thoughts in an effort to glean insights from the data. What follows is a summary of the key insights from our interviews. Many of these insights echo the concerns laid out in the CIRCLE report, but a few provide evidence for gaps and opportunities not addressed by the standards we reviewed above.

4.7.1 Focus on Factual Knowledge

When asked about the primary learning objectives of their courses (with respect to civics), most teachers focused on factual knowledge about the system and function of our government. Common topics included the bill of rights, the three branches, and the concept of checks and balances.

4.7.2 Civics Skills

In addition to teaching factual civics knowledge, students also have the chance to practice some civic skills. For example, every teacher interviewed incorporates discussions of current events into their classrooms. Some of the stated goals of these discussions included: exposing students to dialogue, giving students a platform to voice their opinions, and allowing students to “make up their own mind” about an issue. In addition to classroom discussions, 3 of the 5 teachers mentioned having students fill out mock or real voter registration forms that aren’t submitted.

4.7.3 Instigating Conflict

When asked about their role as a facilitator of classroom discussions, several teachers mentioned that they will “prod” or “instigate” students as a way to challenge or engage them. One teacher described an environment in which students yell at him in disagreement as “perfect.” Teachers also play the role of devil’s advocate or advocate for a minority position if that perspective isn’t being represented, arguing that even if they disagree with the viewpoint, students deserve to hear it and it’s their “job to say it.”

4.7.4 Neutral Facilitators

Remaining a neutral facilitator was one of the few themes common throughout all five interviews. Many of the teachers took pride in their neutrality, pointing to the fact that students are unable to identify their political affiliation, or contrast themselves against other “non-history” teachers they see as pushing their own beliefs and ideas onto students. One teacher took a firm stand against this kind of influence stating, “It’s not my job to flip their opinion.” Still teachers admit that teaching without bias is “tough,” and speak about neutrality more like an aspiration rather than a goal that has been met.

4.7.5 Importance of Context/Background Knowledge

Aside from occasionally instigating conflict and playing devil’s advocate, teachers primarily see themselves as sources of information in classroom discussions. They provide the context and background information for an issue and “make sure students understand what they’re saying.”
One teacher even said that creating well-informed students was his “primary goal.” This seems to support sentiments often seen in civics standards which essentially argue that civil discourse is impossible without sufficient background knowledge.

4.7.6 Engagement

Another common challenge was engagement (or lack thereof). Teachers lamented that it’s easy in informal discussions to have students “sit on the sidelines.” Students who hold minority opinions are often too afraid to speak up in classrooms that are otherwise politically homogenous. To combat these problems, one teacher had a personal goal of directly engaging with every student at least once each week. Another teacher used real-time anonymous online chatrooms which allow shy students to voice their opinion. Several teachers mentioned that engagement is highest when the issues are relevant to the student’s lives (e.g., student’s rights, treatment of rights in popular movies, etc.). When asked about why they felt engaging students in discussions was so important, one teacher said, “How can we expect them to stay engaged as adults if we don’t engage them now.”

4.7.7 Formal Debates

One teachers solution to the inconsistent engagement of informal discussions is to make the discussions more structured. These formal debates usually involve smaller groups of students (3-12 students opposed to classes of 20-30). Practitioners claimed that formal debates fostered deeper discussion and, surprisingly, more empathy with the other side. When probed about the latter, the teacher attributed the increase in empathy to formal debates being a more efficient form of communication that gives students more of a platform to explain their side. In at least one case, formal debates are also more engaging (defined narrowly) simply because each student was required to be engaged in discussion (or at least was being held accountable by their peers to participate).

4.7.8 Ed Tech in Classroom

A few teachers mentioned using educational technology into their classroom. The two main categories of tools were tools that help facilitate discussion, like those mentioned above, and educational games, like those offered by iCivics. With respect to the educational game, teachers complained about the games being too focused on factual knowledge.

4.7.9 Unknown Beliefs and Values

Several teachers alluded to students’ lack of knowledge about their own beliefs and values. Student beliefs primarily come from their parents or guardians, and are not informed by a meaningful understanding of an issue. In an effort to prompt reflection on any changes to their beliefs throughout a class, one teacher has students write down their initial response to some issue, and then, after the issue has been covered, write 4 sentences on whether or not their beliefs changed. Other teachers have students take online surveys like the one offered by isidewith.com, which match responses to political figures or parties. Interestingly, when their survey results don’t match their assumed political identity it results in a state of “denial” rather than self-reflection.
4.7.10  Perspective Taking

In one district, students have a formal opportunity to practice perspective taking. Each class holds a mock constitutional convention, in which each student is asked to assume the role of a real-life convention attendee. During the mock convention, students are required to argue the position of their character and convey their viewpoint to the other attendees. One teacher encapsulated the exercise, stating that asking students to play a role, “gives personality to policy.” Teachers also recalled that, periodically, they would ask a student to play the role of an attendee that held opinions that directly oppose the known opinions of the student. In these cases, the students generally maintain their own beliefs, but finish the exercise with a better understanding of the opposing side.

Despite the fact that perspective taking rarely appeared in formal instructional activities, some teachers thought of perspective taking as an important prerequisite to productive civil discourse. One teacher recalled an Abigail Adam’s quote hanging in their classroom: “I’ve always felt that a person’s intelligence is directly reflected by the number of conflicting points of view he can entertain simultaneously on the same topic.”

4.7.11  Compromise

Even in cases where students are explicitly taught that our nation is built on compromises and “things only get done” with compromise (such as during the mock Constitutional Convention), students are not assessed on their ability to reach a compromise. To be fair, students do not seem to be assessed on their ability to “win arguments” either. Instead, the relative productivity of the discourse seems to be, at least academically, inconsequential.

4.7.12  Other Simulations

In addition to mock Constitutional Conventions, schools also hold mock trials and mock elections. These simulations of civic events are more elaborate than asking students to raise their hands. Students act as poll workers, hand out flyers, discuss party platforms, and (as mentioned above) register to vote. Simulations allow students to practice civic engagement, activism, and the exercise of their civic responsibilities, all in a scaffolded environment in which teachers can model and impose worthwhile values. For example, one teacher forces students to “pinky-swear” that they will register to vote when they come of age.

4.7.13  Tribalism in the Classroom

When asked about the presence of political tribalism in their classrooms, teachers were split. For example, one teacher claimed that, "students don’t have disdain" for students they disagree with, while another teacher said that in his after school political club he sees “disdain for the other side.” Another teacher said that students of the same political orientation “sit together.” Some teachers pointed to social pressures to identify with a particular group (which I would consider a loose definition of tribalism). For example, when asked to take a survey of political viewpoints (like the ones described above), some students worked together to intentionally game the survey so that it would classify them as Trump supporters. In this case, students were intentionally ignoring their own beliefs and values in an effort to remain loyal (or be perceived as loyal) to a group. I can’t think of a better example of textbook tribalism.

Differences in tribalism may result from an unclear understanding of what tribalism entails (a few interviewees asked for clarification or initially thought I was asking about cliques). It
may also, as one teacher suggested, be due to differences in school sizes, with smaller schools experiencing less tribalism than, say, a high school being fed students from three middle schools who don’t know each other. One teacher noted that students are often unable to believe the other side could hold a particular opinion. When asked what has been most effective at combatting that kind of disbelief he said simply, “engaging in conversation,” noting that personal stories are most effective.

4.7.14 Impact of Real World Events Creeping In

While most students may not be engaged enough to exhibit the kind of tribalism that is on the rise in the real-world, that does not mean they are unaffected by our current political climate. More than one teacher described how their district handled students wanting to wear “Make America Great Again” paraphernalia on school trips to the nation’s capital. Given a recent, similar situation that made national headlines, subjecting the students involved to a massive amount of negative attention, the districts were apprehensive about the request. Ultimately, one district called the parents and dissuaded them from wearing the apparel. The other district, on the other hand, thought it would be a violation of free speech to interfere.

Teachers also mentioned the increase of some troubling behaviors in recent years. One teacher with more than a decade of experience said, unequivocally, that the amount of racist comments and behaviors have increased in the past few years. He likened the students to the raptors in Jurassic Park, testing the fences to see how much they can get away with. Several teachers mentioned that many more students now believe in conspiracy theories and attribute the increase to the popularity of conspiracy theory videos on youtube. One teacher said that he can no longer assume that students believe the Holocaust happened and is “bothered” by what students are willing to say, even if it’s a joke.

Conspiratorial thinking is large area of study within the domain of informal reasoning. Studies have shown that conspiratorial thinking is associated with various negative outcomes including less egalitarian human rights attitudes [74], aggressive political behavior [73], and racism [3]. Interestingly, prompting analytic thinking can reduce belief in conspiratorial thinking [75].

In general, these cases show that civic education doesn’t happen in a bubble. The current political context seeps into the lives of students in new, unexpected ways, and teachers, parents, and school administrators guide students through these real world civics challenges.

4.7.15 Media Literacy

One of the ways in which curricula could support the fight against real world problems like the increased belief in conspiracy theories is by improving media literacy. In general, media literacy in classrooms of the teachers interviewed was addressed in passing. Media consumption was rarely an explicitly included in the course syllabus, with teachers instead covering the topic informally throughout the year. Some teachers mentioned that the topic was instead covered in a different class (e.g., English or Library). Still teachers recognized the importance of media literacy, with one school incorporating a unit on confirmation bias for the first time this past year. Another teacher mentioned that he spends time throughout the year making the distinction between fact and opinion more clear to his students, or asking students to determine if a website has an agenda.
4.7.16 Bias

Aside from the above mentioned unit on confirmation bias, only one other teacher mentioned bias explicitly. Describing the broad learning objectives of his classroom discussions, he said that he explains to his students that everyone is biased and that it’s okay to be biased, “but don’t use that to make irrational decisions.” In general, classroom discussions seemed to resemble the Jeffersonian ideal: rational, information communication tools, designed to be “as objective as possible,” and focused on “policy not people.” All of which are, to reiterate, vital components of productive civil discourse. But outside of the controlled, facilitated environment of the classroom, productive civil discourse requires a working understanding of the values that inform and bias our judgements.

One comment illustrates this distinction clearly. In discussing his role as a facilitator of classroom discussions, one teacher said that it’s his job to, “make sure students strive toward the common goal of understanding.” Again, this is a worthwhile goal, but to have it be the ultimate goal suggests that more information (e.g., facts and figures) is enough to foster productive civil discourse – a position not supported by the literature. In contrast, we argue that students should instead use that information, along with their understanding of their own values and the values of others to practice identifying and moving productively toward a common goal. If we end the discourse process before students are given the chance to practice the skills that make civil discourse ultimately productive, we cannot expect students to know how to engage in productive civil discourse as citizens.

4.7.17 Civil Discourse Ideals

The lack of practice engaging in discourse intended to be productive is likely due to limitations of the classroom. The teachers themselves had a intuitive and deep understanding of what makes for productive discourse, mentioning features like: “more listening than talking,” “respecting another’s beliefs,” and “appreciating why they feel that way.” One teacher succinctly captured the entire process, first summarizing the Jeffersonian piece as, “one’s ability to convince someone else to join them,” but then extending beyond understanding to include “empathiz[ing]” with the other side, and “com[ing] together on an issue to try to make things better.” We believe that these latter two skills, empathizing and moving toward a shared goal, are not adequately supported by current civic education curricula, and that systems designed to support these skills will increase the productivity of civil discourse and decrease political tribalism.
Chapter 5

Proposed Work

In the previous chapters, we have discussed how tribalism threatens to undermine productive civil discourse. We have discussed how that tribalism is bolstered by misunderstandings about how we form our own beliefs and view the beliefs of others. Finally, we have discussed our prior work, which lays the foundation for addressing three shortcomings of the current state of civil discourse instruction:

1. A lack of practice on civil discourse skills in general.
2. A lack of focus on the political perspective taking skills needed to engage in civil discourse that is productive.
3. A lack of adaptive interventions that recognize the dynamic relationship between user and content values, and how that relationship impacts how we evaluate political arguments.

What follows is a description of our proposed research agenda. This research agenda is designed to inform and assess an adaptive educational game that addresses each of these gaps in the civic education space directly. Our proposed work can be coarsely divided into three stages: 1) development of the system, 2) a small-scale pilot, and 3) the full-scale experiment. Each of these stages will be discussed in detail below.

5.1 Affordances of Games in Civic Education

Many of the features of instruction that are considered best-practices in civics education are more easily implemented in an educational game. For example, experts recommend that civic education should be focused on practical skills, and that interactive learning combined with traditional lectures in civics is more effective than either by itself [12]. Teachers know this, which is why simulations of civic events (mock debates, elections, trials) are so prevalent in schools. The immersive nature of games makes them well-suited for these kinds of simulations [30]. Individual play may reduce the negative effects of social pressure (e.g., shyness, tribalism, feigned apathy). Investing players in the outcomes of their characters may increase feelings of relevance, and thus, engagement [5]. Allowing players to practice in a safe environment, free of the potential social ramifications of saying the wrong thing, may allow players to develop more confidence in their own civic skills, which itself has been shown to be correlated with civic action [38, 38]. Finally, games also have some features that make them a more ideal platform for simulations of civic events. Games are well-structured, allowing for both opportunities for
scaffolding as well as the ability to measure student behavior in ways that are not possible in a classroom of 30 students. Moreover, unlike large-scale assessments, games allow us to measure participatory skills, not just conceptual knowledge.

5.2 A Finite Game within an Infinite Game

In his 1987 book *Finite and Infinite Games* [11], James Carse argues that many human activities can be thought of as one of at least two categories of games: finite and infinite. In a nutshell, the goal of a finite game is to win the game, while the goal of an infinite game is to simply continue playing. For example, a professional tennis match is a finite game – the goal is to win. In contrast, the game of “seeing how many times we can hit the tennis ball back and forth” is an infinite game – the goal is to keep the game going for as long as possible. Carse extends his dichotomy well beyond what most people might consider as games. Getting a job, for example, is a finite game. The interview process ends with a clear winner. Keeping a job, on the other hand, is an infinite game (at least until retirement).

Carse describes a number of distinguishing features of finite and infinite games, only one of which is relevant for this discussion. Carse states that the rules of a finite game define it. They cannot change during the game because then players would be playing a different game. In contrast, the rules of an infinite game are more malleable, and can be changed if it looks like one person is going to win. He writes:

> If the rules of a finite game are the contractual terms by which the players can agree who has won, the rules of an infinite game are the contractual terms by which the players agree to continue playing.

Carse likens the rules of an infinite game to a living language’s grammar, which is used to ensure that discourse continues to flow smoothly. In contrast, Carse likens the rules of a finite game to the rules of a debate, which are designed to “[bring] the speech of another person to an end.”

Consider the following features of a deliberative democracy, as outlined by political philosopher Joshua Cohen, who writes that a deliberative democracy is “ongoing...continuing into the indefinite future” [14]. He writes that the members share “a commitment to coordinating their activities within institutions that make deliberation possible and according to norms that they arrive at through their deliberation” (p. 346). In other words, deliberative democracy is an infinite process, it governed by rules, members follow those rules to make deliberation possible, and the rules themselves are the result of the process (i.e., malleable). Democratic societies can, in this sense, be thought of as infinite games. The goal of a democratic society is its persistence and preservation. Threats to that preservation often come in the form of one player or one group of players amassing too much power, threatening to end the game of democracy (or at least hindering the function of democracy as as a grammar of government). When these threats arise, the players can agree to change rules (i.e., laws and norms) to ensure that the game persists. Consider, for example, the 22nd Amendment (limiting the president to two full terms) which was a rule-change made to prevent one player from ending the game of democracy (i.e., becoming a dictator).

Carse’s dichotomy is useful in understanding the role of our proposed educational game in its broader context. Our game is finite, but ultimately we hope to teach the skills and dispositions required to perpetuate an infinite game. Consider the following features of the infinite game of democracy:
1. Democracy, especially at the federal level, is a massively cooperative activity. This means that (at least in an ideologically pure democracy) the power of each individual citizen can seem small or even insignificant.

2. A heterogenous and passionate electorate necessarily results in slow progress. President Obama once likened it to steering an “Ocean liner two degrees North or South so that 10 years from now, we’re in a very different place than we were.”

3. Democracy requires that the players are motivated by their shared goals. Civil discourse (a necessary ingredient of democracy) is more productive when players can ignore the appeal of individualistic or tribalistic goals like “winning the argument,” and instead remind themselves (and each other) of their common purpose.

These features – collectivism, feelings of powerlessness, and near-indiscernable progress – aren’t exactly the ingredients for a blockbuster game. More commonly, games center around competition, where you play as a character who is in some way more powerful than yourself, and where progress is clearly defined. A key challenge of our work will be providing students with an honest expectations about the more frustrating features of a real-world democracy without producing jaded or cynical students. Simulations provide one solution to this challenge because the window of time between action and consequence can be controlled. We can adjust the window so that the gratification is delayed, but not so delayed that we’re asking students to wait 10 years to experience the fruits of their labor. In addition to making progress more visible, this allows students to feel the weight and power of their actions as individual citizens.

In contrast to the first two items, the collectivist (versus competitive) nature of productive civil discourse cannot simply be solved my reducing scale. Instead, we put players in competition with tribalism itself, situating players in a Player vs. Entropy scenario. We instantiate this conflict in two ways: First, a key metric of gameplay is a variable called Sense of Community which is in a direct inverse relationship to another variable called Tribalism. Players can take actions to increase the former, which decrease the latter. Second, the antagonists of the game are all designed to be avatars of unproductive civil discourse. They provide an anti-model of discourse by increasing Tribalism and decreasing Sense of Community (in service of their own agenda).

5.3 A Brief Overview of the Current Iteration of Our Game

In order to delve deep into the framework that will guide the development of our educational game, it may be helpful to briefly describe the basic mechanics and goals, at least as they exist in the current iteration. In our game, you play the role of a medieval peasant living in a village suffering from a series of problems. To solve these problems, you must convince your fellow villagers to unanimously elect you as their new leader (this particular feudal society is surprisingly democratic). Getting elected unanimously requires that players maximize three key metrics: Approval of Me (AOM), Approval of Plan (AOP), and Sense of Community (SOC). These metrics are cumulative, and players must get or keep the sum total of these three measures above a certain threshold to ensure the townsperson will vote for them.

1Using a feudal society as the setting serves two purposes: 1) it brings levity to what might otherwise be emotionally heavy content, and 2) it distances players from well-worn issues, forcing them to rely on their own values to inform their beliefs rather than adopting the pre-established beliefs of their tribe.
Throughout the game, if players make productive civil discourse moves, these metrics go up. If they make unproductive moves, the measures go down. Examples of productive civil discourse moves include:

1. Correctly identifying the values that motivate the beliefs of a non-player character (NPC), and confirming your understanding with them. (AOM++)
2. Reframing a proposed plan to appeal to an NPCs values (AOP++)
3. Attempting to get an NPC to empathize with the opposing viewpoint (SOC++)
4. Identifying shared values between two opposing NPCs (SOC++)

Each of these productive moves is dependent on choosing the correct statement (i.e., the statement that respects or appeals to the values of the NPC). A failure to identify or respect the values of an NPC will result in a decrease in one or more of the key metrics. At the same time, your opponent, the incumbent leader is actively trying to increase tribalism and division within the community (to prevent a consensus that would kick them out of their job).

5.3.1 What the Game Isn’t

Because people often come with preconceived notions of what the goal of a political discussion should be, it is important to not only state the goals of our game, but clarify some things that are explicitly not goals.

1. **Our game is not about persuasion.** Our goal is not to make students more persuasive, or to give them tricks they can use to fool someone into agreeing with them. Instead, our goal is to help discussants understand and empathize with the other side in an effort to combat tribalistic thinking.

2. **Our game is not about finding middle ground.** Finding the middle ground is not an effective compromise solution, as it slowly pushes each party to become more polarized (i.e., they can secure more elements of their position if they start much farther from the middle than the opposing side). Instead, this game is about finding shared values and using those values as the foundation of an actionable solution.

3. **Our game does not confuse civility with politeness.** Instead, we subscribe to Papacharissi’s [55] view of civility. Namely, that the purpose of civility is not simply to ensure that the flow of conversation is smooth, but, more importantly, it is to *foster democratic goals*. Moreover, discourse governed only by politeness will likely work against our instructional goals by silencing marginalized opinions in an effort to minimize conversational friction.

4. **Our game is not a replacement for in-person discussions.** Instead, we hope to augment the productivity of those discussions by giving students the opportunity to practice discourse skills in a safe, scaffolded, and adaptive environment. Our game is a batting cage; in-person discourse is the game of baseball we’re hoping to support.
5.4 The Transformational Framework of Game Design

We will guide the design and iterative development of our game with Sabrina Culyba’s Transformational Framework [18], which aims to aid in the design of games that are meant to change the player in some way. Culyba contrasts this to games intended to be entertaining, arguing that entertainment games focus on the player within the world of the game, where transformational games focus on the player “beyond the game” (p. 8). In entertainment games, players get better at solving the unique problems and challenges of the game. In transformational games, the knowledge, skills, and dispositions used to solve in-game problems should, ideally, transfer to the solving of real-world problems. Finally, for a game to be transformational, the changes to player behaviors and attitudes must persist beyond the span of the game. For example, it may be scientifically interesting if we can increase perspective taking behaviors within the game, but if we fail to see students employ those skills in real-world discussions, then the transformational goals were not met.

The Transformational Framework consists of eight interdependent nodes. Some of which this document has already addressed in some form. What follows is a discussion of the remaining nodes and how our proposed work will incorporate them.

5.4.1 High-Level Purpose

The motivation for this work is to combat the rising tribalism and polarization of American politics, which hinders productive civil discourse, and consequently, the efficacy and security of our democracy.

5.4.2 Audience and Context

Our primary audience will be late high school students. These students are on the verge of graduating and assuming the rights and responsibilities of citizenship, and the public education system is our last best chance for changing how they engage in civil discourse for the rest of their lives. For context on the state of civic education in America, see section 3, *Civic Education in America*.

5.4.3 Player Transformations

After playing our game, we hope that players are more able and willing to engage in productive civil discourse (i.e., discourse that fosters democratic goals). More specifically, our game targets a host of conceptual, procedural, and dispositional knowledge components.

**Conceptual Knowledge**

Students should understand:

**KC 1**: Each of the five foundational values

**KC 2**: The values they personally think are more important

**KC 3**: The values they personally think are less important

**KC 4**: Why speaking to someone’s values can be more effective than using facts and figures

**KC 5**: What makes civil discourse productive
KC 6: The difference between civility and politeness

KC 7: The difference between respecting values and respecting beliefs

Procedural Knowledge
Students should be able to:

KC 8: Identify the values that inform a belief they share
KC 9: Identify the values that inform a belief they don’t share, or that they oppose
KC 10: Communicate their own beliefs
KC 11: Communicate how their own values inform their beliefs
KC 12: Listen to the beliefs and/or values of others
KC 13: Confirm their understanding of perceived values and beliefs
KC 14: Identify shared values
KC 15: Reframe an argument in a way that better speaks to someone’s values

Dispositions
Students should be:

KC 16: Open to considering different points of view, and exercising what philosopher Laurie Calhoun calls “epistemological humility” [9].

KC 17: Empathetic towards the foundational roots of different beliefs
KC 18: Invested in shared goals (i.e., the common good)

5.4.4 Barriers
Before designing a game that is transformational, the framework asks us to first answer the question, “Why aren’t players already transformed?” Each of the gaps in civic education identified in the civic education standards or in the teacher interviews implies a potential barrier. Here are some of the key barriers to our transformations:

B1: Engagement and Practice: Teacher interviews revealed that regular participation (e.g., formal debates) is the most effective way to keep students engaged, but classroom time is so limited that one benchmark of successful engagement was participation in a discussion at least once a week. Mastering a skill, especially a complex skill, often requires many practice opportunities. An educational game can provide each student with many opportunities for practice in parallel with other students. Then, when students return to a classroom-wide discussion, they already have already had some practice engaging in productive discourse, and don’t have to focus on learning the skills of discourse from short bursts of participation, days apart.

B2: Social Pressure: Teachers mentioned that students’ expressed beliefs (or even their willingness to speak at all) can be driven by social pressures from their peers. Almost every teacher mentioned that one of their roles as a facilitator was to create an environment in which students felt safe to express their opinions, but students likely understand
that any publicly expressed belief could have social consequences. This is particularly true for, say, a student with a minority opinion in a classroom that is almost entirely ideologically homogenous. In contrast, our game is designed to be private. This not only allows players to come to terms with their personal values in private, but it also gives them the freedom to fail at civil discourse skills in a way that doesn’t have repercussions for their social life outside of the classroom. It is our hope that the freedom to fail will make practicing civil discourse less stressful and more fun. It is of course, strange to have a game about discourse that is played individually, but again, this game is meant to be used in conjunction with real-world classroom discussions. There are also positive benefits to social pressure that shouldn’t be ignored, the absence of which likely contribute to polarization (e.g., desire to save face). We mitigate these concerns by actively disincentivizing tribalism.

B3: Relevance: In contrast to national concerns like the deficit or increasing tensions with a trade partner, games are more relevant to players because players, almost by definition, have a stake in the outcome. Still, being more confident in student interest will require more user testing.

5.4.5 Domain Concepts

As this is an educational game, our transformational goals center around learning many of the domain concepts. In addition to those knowledge components, our game will be sure to include any required background knowledge necessary to engage in productive discourse (a major theme of teacher interviews). That said, many of the issues explored in the educational game are novel and intentionally designed to be removed from the kind of issues that might show up on a party platform. This is for two reasons: 1) we want students to practice forming beliefs based on their values, not what they heard their parents say, and 2) there’s neither the time to delve deeply into complex issues nor the need to, as teachers already do an excellent job exploring these often sensitive topics.

5.4.6 Expert Resources

In addition to the primary purpose of teacher interviews (i.e., gaining insights and validating needs), one ulterior motive was to begin building a network of engaged practitioners who we could call upon for feedback on future iterations. When asked if they would be willing to be a resource for expertise and feedback in the future, each of the five teachers interviewed said they would be willing to act in that role. For a discussion of the efforts we’ve already made to leverage the expertise of teachers, see Section 4.7, Teacher Interviews.

5.4.7 Assessment Plan

Our assessment plan consists of two components. First, our in-game assessment will include the measurement of each of the conceptual and procedural knowledge components listed above. We will measure performance on these KCs by designing logging student interaction-level data as students complete instructional activities that require students demonstrate particular knowledge components. Second, our assessment plan includes an out-of-game assessment designed to capture growth on both the conceptual and procedural skills as well as the dispositional skills (which can be difficult to measure in a game environment). To measure these ill-defined knowledge components, we will conduct a verbal protocol analysis of classroom discussion data.
5.5 Study 4: Pilot

Before conducting a full-scale implementation of the game, we will conduct a small pilot study. The primary purpose of the pilot is to test the integrity of our experimental procedure, so as to reduce the possibility of data loss during the full scale experiment. On this front, we will pay special attention to the following critical areas:

1. **Data Logging**: Data logging will be checked for completeness, accuracy, and successful integration with LearnSphere. Some simple analyses will be run to test data quality.

2. **Student Engagement**: Student log files should show patterns associated with active engagement rather than random clicking or behaviors designed to game the system. We will also address engagement and enjoyment directly with open- and close-ended questions on the post-test questionnaire.

3. **Usability**: Students will be asked to write down any UI-related bugs, frustrations, or suggestions. Any friction caused by the interface may cause extraneous processing that interferes with learning.

4. **Learning**: Although the pilot experiment will likely not be large enough to confidently draw conclusions from the data, we would expect student responses, as measured by closed- and open-ended post-test questions, to be positive when asked questions like, “Did you learn something you didn’t know before? If so, what was it?” We would also expect student explanations of their learning to roughly align with our learning objectives. Finally, we will pay special attention to the difficulty of the pre-test questions, making sure that they are appropriately leveled as to not produce a ceiling or floor effect.

Approximately fifteen (n=15) high school age students will be recruited for the pilot experiment from local high schools. Before playing the game, students will be asked to complete a pre-test, in which they are asked questions similar to in-game scenarios, such as, “What would be the best response to this person’s concerns?” After the pre-test, students will be asked to play the game for the duration of two, one-hour sessions. At the conclusion of the second session, students will be asked to complete a short, post-test questionnaire designed to assess the how effective the game was at teaching key learning objectives and how enjoyable the game was to play. At the conclusion of this pilot experiment, we hope to have identified and fixed any complications that may interfere with high-fidelity data collection in the full-scale experiment.

5.6 Study 5: Full-Scale Experiment

The full-scale experiment will closely follow the format of the pilot experiment, with some key additions. We will partner with two local civics teachers who teach at least two sections of the same civics course. All four classes will take a pre-test similar to the one described in the pilot experiment. One of each teachers classes will be randomly assigned to the treatment group, which will play the educational game over the course of two, one-hour sessions. During this time, the other class (i.e., the control group) will receive normal, business-as-usual instruction. After the two sessions, both groups will take a post-test to assess any changes in performance, as measures of enjoyment and engagement. The post-test for students in the treatment condition will also include questions designed to measure game efficacy.
In the days following the two sessions of instruction (either via game or typical classroom procedure), students will be asked to engage in a formal classroom discussion of an issue, following the five steps of the Constructive Controversy format. Classroom audio data will be collected during the course of this activity. At the conclusion of the final step of the process, the audio data will be anonymized (to blind whether the audio was from the treatment or control condition), and then analyzed via verbal protocol analysis. The protocol analysis will primarily be searching for markers of productive civil discourse, especially as they pertain to the skills and dispositions outlined in Section 5.4.3.

Following from the above reviewed prior work, we hypothesize that:

**H1:** In a verbal protocol analysis of discussion data, students who play the game will engage in civil discourse that is more productive than students who do not.

**H2:** Students who play the game will be more adept at responding to civil discourse scenarios than students who do not.

**H3:** Log data from the game will show clear learning gains on each of the conceptual and procedural knowledge components outlined in Section 5.4.3.

**H4:** Alignment between student and content values will be a significant predictor of performance on argument evaluation tasks.

**H5:** The impact of alignment on performance will decrease with practice within the game.

**H6:** Students who play the game will be exhibit less bias on post-test argument evaluation tasks than students who do not.

### 5.6.1 Potential Contributions

This results of this experiment could potentially contribute to a number of important domains of study. First, successfully reducing bias as a result of value-adaptive instruction not only demonstrates the accuracy of theory- and data-driven estimates student and content values, but also demonstrates the importance of incorporating estimates of values in learner models (at least in relevant domains). Second, successful learning outcomes would further demonstrate the applied use of Haidt’s Moral Foundations Theory as an instructional tool for improving discourse skills, which, to our knowledge, has only been done once before [22]. Finally and more broadly, we hope that, by offering a novel method for improving the productivity of civil discourse, we can make a more practical contribution to the common good.
Chapter 6

Conclusion

In this document, we have described a novel approach to improving the productivity of civil discourse. This approach includes responding to three gaps in status quo civic education: 1) a lack of focus on civil discourse skills and dispositions, 2) a deficit in instruction designed to fight against tribalism, and 3) the absence of any acknowledgement of the role of values in informing our beliefs and biasing our reasoning. To address these gaps, we are developing an educational game that offers students: 1) opportunities to practice civil discourse skills in a scaffolded, low-consequence environment, 2) instruction intentionally designed to help students understand their own values and the values of others in a space devoid of the social pressure to conform to the established beliefs of the tribe, and 3) adaptive instruction that uses a model of the interaction between student and content values to estimate and reduce myside bias.

Figure 6-1: Broad overview of the project’s stages
Bibliography


[37] Bob Hone, Joyce Rice, Chas Brown, and Maggie Farley. Factitious. 2018.

[38] Myiah J Hutchens and William P Eveland Jr. The long-term impact of high school civics curricula on political knowledge, democratic attitudes and civic behaviors: A multi-level model of direct and mediated effects through communication. circle working paper # 65. *Center for Information and Research on Civic Learning and Engagement (CIRCLE)*, 2009.


[58] Pew Research Center. The partisan divide on political values grows even wider. 2017.


[67] Craig Silverman. This analysis shows how viral fake election news stories outperformed real news on facebook, Nov 2016.


