Doctoral Thesis

Value-Adaptive Instruction: Improving the Productivity of Civil Discourse and Addressing Bias

Nicholas Diana

June, 2020

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.

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" [59]) 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 completing 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 the type of dialogue that engenders commonality and collective problem-solving.

In contrast to the status quo, we developed and tested the impact of 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. We found that student performance on key civil discourse skills (e.g., value-identification, tribalism reduction) improved with practice. We also demonstrate, for the first time, that an AI-powered educational game that adapts instruction to a student's specific values can be used to measure and, in some cases, reduce the impact of bias when reasoning about political beliefs. In addition to direct applications in civic technology, this new, value-adaptive instruction has implications for systems designed to mitigate bias and augment human reasoning.

Contents

Chapter 1

A New Age of Unproductive Civil Discourse

"Wow, that internet argument completely changed my fundamental belief system," said no one ever.[1]

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) [64]. 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) [74]. 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 [77, 14, 18]. Online communications is also often anonymous, which is most commonly associated with negative behaviors like "flaming" [45], but could also be potentially beneficial to communities whose voices have historically been marginalized in face-to-face communication [2, 6]. 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

[1] This text is taken from a popular internet meme created by someecards.com user stellar3713

society because it exists outside of institutional and economic power structures.[2]

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, and more angry than other places where people discuss politics [20]. The present state of civil discourse may more closely echo the concerns of the so-called "luddite" internet pessimists (e.g., [5]). 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) [62]. In this polarized moment, more than half Americans (53%) find it "stressful" to discuss politics with someone they disagree with [63]. 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 [68, 56, 75]. But recent years have seen the introduction of a new source of filter bubbles: curated news algorithms [60]. 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 [21].

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.

[2] 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 [24, 61, 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 [52, 22, 67, 76], and has been linked to societal problems like the perpetuation of stereotypes [19]. 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

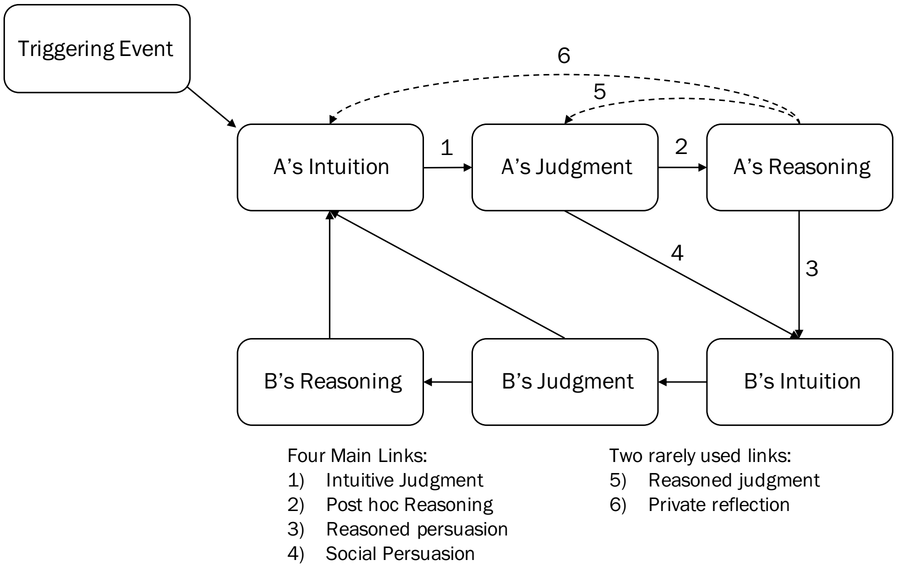


Figure 2-1: The Social Intuitionist Model reproduced from [34], p. 55. We see that moral decisions are first driven by System 1 intuitions, and then justified after the fact by System 2 reasoning.

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 understanding 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..." [70].

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 activity (e.g., voting,

contacting government officials, discussing politics) [8]. 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 driven by 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 [66]. 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 [8]. People who posses 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* [53] 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 [5] – 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 [57], 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 [47], 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 [82] (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 [7] (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 [45]). 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 discussants 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 unguided 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 from 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 success and efforts 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, and 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-adaptive 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 later work.

4.1 Potential Benefits of Value-Adaptive 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-adaptive 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-adaptive 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-adaptive) 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-adaptive instruction: **targeted perspective taking**.

Perspective taking is present in both the C3 and CIRCLE civic education standards we've reviewed above [30, 82], but notably absent from the PA State Standards [57], which make no mention of empathy or perspective taking as important skills for civic life. Value-adaptive 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-adaptive 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 our own views or beliefs (and conversely, more critically when they do not) [76]. 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-adaptive 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 [73], and that these stories are generally more engaging than stories produced by traditional news outlets [72].

The proposed solutions to these problems generally fall into two categories. The first category leverages various machine learning methods [15] to create "fake news detectors." While some of these classifiers are quite sophisticated [86], 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 [54]. 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" [87]. 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 [11], 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[1].

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

[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.

in 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 Inutuitionist 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 rationale 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 [25] as well as more specific political beliefs (e.g., "Climate change is real") [48, 69]. 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-adaptive 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

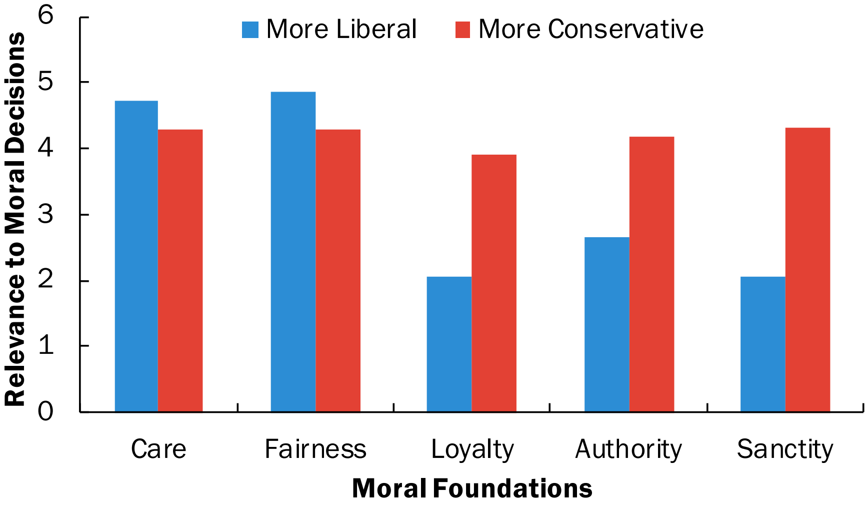


Figure 4-1: Relevance to Moral Decisions by Moral Foundation for more conservative and more liberal participants in one of our studies. These values closely match previously observed values for liberals and conservatives [35], suggesting that our recruitment pool is politically diverse.

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) [16]. 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 [55] 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

Foundation	Related Concepts	Example Vignette
Care & Harm	kindness, gentleness, and nurturance	You see a zoo trainer jabbing a dolphin to get it to entertain his customers.
Fairness & Cheating	justice, rights, and autonomy	You see a runner taking a shortcut on the course during the marathon in order to win.
Loyalty & Betrayal	patriotism and self-sacrifice for the group	You see the US Ambassador joking in Great Britain about the stupidity of Americans.
Authority & Subversion	leadership/followership, deference to authority, and respect for traditions	You see a woman refusing to stand when the judge walks into the courtroom.
Sanctity & Degradation	disgust, contamination, purity, and holiness	You see a man in a bar using his phone to watch people having sex with animals.

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://wwwm.moralfoundations.org), and an example vignette designed to evoke the foundation (adapted from [11]).

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 [55] methodology employed in the current study) have demonstrated the effectiveness and efficiency of the method (in terms of computational cost). Mikolov et al. [55] 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(France)*) 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 [27] 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 a word's 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 [55] called *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 *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 *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 *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 *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] 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):

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

Where *rating* is ratings of argument strength, *quality* is argument quality (high or low), *alignment* is the alignment between user and content values, *(1/participantID)* is a random effect for participant, and *(1/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 *alignment* to the impact of quality. Still, we can make a meaningful, subjective comparison between the impact of *alignment* and "quality" (as operationally defined in this context). In this context, the impact of *alignment* on ratings of strength ($\beta = 3.06$, $p < 0.001$) was greater than the impact of *argument quality* ($\beta = 1.33$, $p < 0.01$).

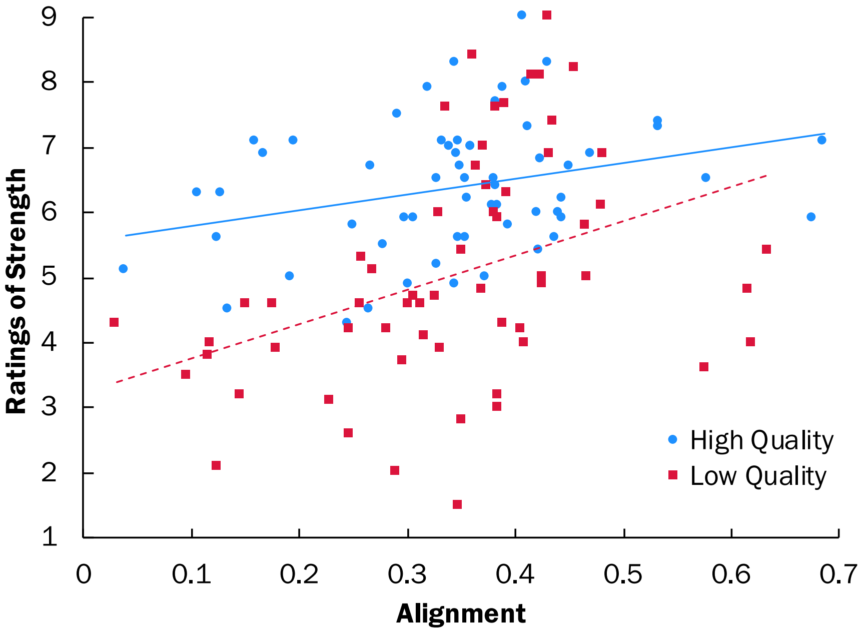


Figure 4-2: Relative impact of alignment on the ratings of high and low quality arguments. Each data point represents the average rating and alignment for all arguments within a category (high or low quality) for one participant. On average, participants rated high quality arguments as stronger than 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 [46]. To test whether this was true of our sample, we built a mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *argument quality* and *alignment*age* as fixed effects (where *alignment*age* is an interaction term). We found that

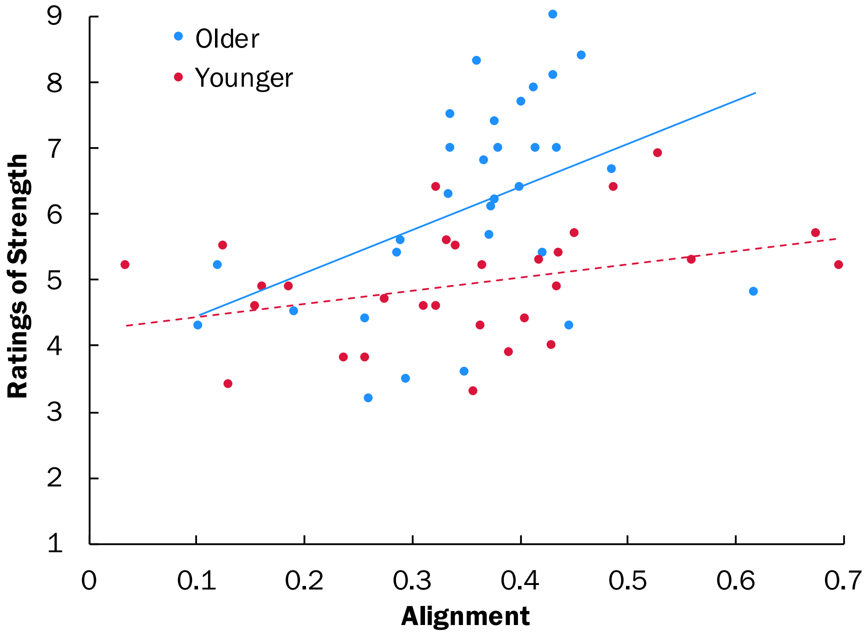


Figure 4-3: The interaction between age and alignment. Each data point represents one participant's average rating and alignment scores. Alignment had a much larger impact on ratings of strength for older participants (participants above the median age) than younger participants. This conforms with previous findings examining the relationship between bias in argument evaluation and age.

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 that 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" [49]. Despite the extensive work identifying and measuring the impact of biases on reasoning [76, 22, 83], work on debiasing is limited [49]. This is likely because developing debiasing interventions that are effective has been shown to be challenging [23].

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 1, all participants were asked to evaluate the relative strength of a 20 politically charged arguments. However, in this experiment, some participants were shown an intervention message

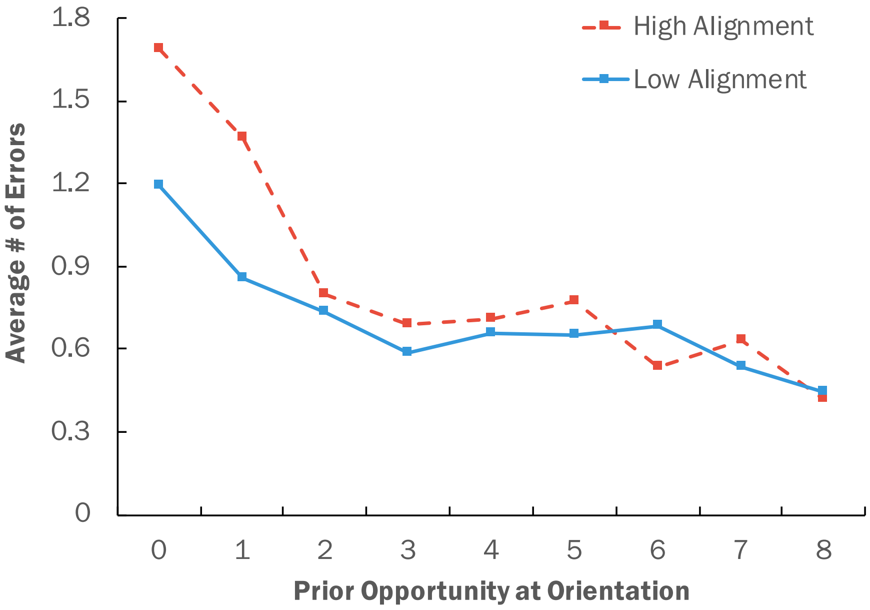


Figure 4-4: We found that users were biased by their beliefs, as evidenced by the differences between early performance on high alignment and low alignment problems. Alignment group was determined by score (1-5) on the self-reported alignment questions (High = 4 or 5, Low = 1 or 2). These results suggest that belief bias impacted performance early in the experiment, but that the effect diminished with practice.

alongside 10 of the arguments. In the adaptive condition, participants were shown an intervention message alongside the 10 arguments that most closely aligned with their own values. In this random condition, participants were shown an 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 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) [65].

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 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 theses 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 of themselves, and to observe consequences of that vulnerability 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 a 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 Abigal 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 [80], aggressive political behavior [79], and racism [3]. Interestingly, prompting analytic thinking can reduce belief in conspiratorial thinking [81].

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 a component 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.

4.8 Conclusions

In these first four 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.

The following chapters detail 1) the educational game that we developed to address these shortcomings, 2) the experiments we designed to test the efficacy of this system, 3) the results of those experiments, and 4) the implications of those results.

Chapter 5

Game Design and Development

"I liked the way the game was made and how it reminded me of how old games were made, like gameboys."
- A student making me feel 1000 years old

Given the rising popularity of educational games in other domains (e.g., math and science), it might be natural to look towards new instructional technologies as a potential solution to our civil discourse challenges. However, games in the civic education domain also have unique features that make implementing meaningful games in this space particularly challenging. In this chapter we will discuss:

1. some of the affordances of games in civics education (i.e., why an educational game may be the best medium for meeting our instructional goals),

2. some of the unique challenges of designing instruction in this space,

3. our development process, including how we address these challenges, and finally

4. the goals and mechanics of the educational game we ultimately deployed in classrooms.

A brief disclaimer regarding terminology: I use the terms "foundation" and "value" interchangeably to refer to Haidt's Moral Foundations. Strictly speaking, these two things may not be equivalent, but the term "value" proved to be more intuitively understandable and, thus, efficient while working with students (rather than introducing a new term). In this context, the noun "value" and the verb "value" are almost tautologically linked, where one's value is something one values. For this reason, I make an effort to only use the word "value" to mean 1) one of Haidt's moral foundations or 2) the act of deeming something important.

5.1 Games in Civic Education: Affordances and Challenges

Many of the features of instruction that are considered best-practices in civics education align with features of educational game environments. 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 [10]. Teachers know this, which is why simulations of civic events (e.g., mock debates, elections, trials) are so prevalent in schools. The immersive nature of games makes them well-suited for these kinds of simulations [30]. For example, individual play may reduce the negative effects of social pressure (e.g., shyness, tribalism, feigned apathy). Additionally, 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]. Finally, games also have some features that make them a more ideal platform for simulations of civic events. Games are well-structured, allowing both for scaffolding opportunities 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 (e.g., state testing), games allow us to measure participatory skills, not just conceptual knowledge.

However, games in the civic education space also present with unique challenges that designers must overcome if they hope to have a meaningful impact on a player's civic knowledge, skills, or attitudes. We will discuss two of these challenges in detail: the ill-defined nature of knowledge and skills in the civics domain and the infinite nature of problem-solving in civics.

5.1.1 Civics as an Ill-Defined Domain

Problems in civics often do not have the clear, verifiable answers that are more commonly found in well-defined domains (e.g., physics, chemistry) [51, 50]. But because instructional technologies generally require some degree of definition, educational games in the civics domain tend to be shallow, focused on teaching students conceptual knowledge (generally about the structure of government) rather than the skills they need to become productive and engaged citizens.

Ill-defined domains have a number of features that distinguish them from their well-defined counterparts, but for the purposes of this discussion, we will focus on two. First, ill-defined domains often require students to understand when and how to apply *open-textured concepts* (i.e., concepts that require the user's interpretation and judgment in order to be applied correctly) [85, 50]. For example, imagine an author sets out to write a short story. How and when to incorporate an element like *suspense* is dependent on context and the author's judgment.

Second, ill-defined problems are, in general, difficult to divide into independent subproblems, or may require that subproblems be solved in parallel [50]. Our author cannot subdivide his story into a beginning, middle, and end in such a way that they can work on each independently without impacting the others. Coherent narratives require that all three parts be developed essentially in parallel.

Lynch and colleagues [50] suggest that an essential part of the ill-defined problem-solving process is framing or recharacterizing the ill-defined problem to make it more tractable. In this light, the mechanics of the game discussed in this chapter are one instance of the recharacterization of the domain of civil discourse. Importantly, as with any recharacterization of an ill-defined domain, this treatment of civil discourse is subject to debate [50]. We will discuss how the game presented in this chapter either represents or adapts to these ill-defined features in the *Game Content* and *Gameplay* sections.

5.1.2 A Finite Game within an Infinite Game

In his 1987 book *Finite and Infinite Games* [9], 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...continu[ing] into the indefinite future" [13]. 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 is 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 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 (but not always), games are driven by 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 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. We describe how we address each of the above three features and attempt to strike a balance between authenticity and enjoyment in the subsection *Addressing the Challenging Features of Civics Education.*

5.2 Development Process

To guide the development of our game, we used the Transformational Framework [17], which, as the name suggests, is designed to support the creation of games that change or transform the player in some meaningful way. The specific transformational goal of the current game was to make players able to engage in civil discourse that is more productive (i.e., discourse that fosters democratic goals [59]).

One of the key principles of this framework is a reliance on expert guidance during the iterative development process. Input from experts (civics educators) allows us to: 1) Validate the high-level needs that we had gleaned from national Social Studies curricula, and 2) Adjust our system, if necessary, to any unforeseen challenges present in our local community (i.e., the community in which the system will be deployed). To gather input from expert instructors, we conducted a series of structured interviews (see Section 4.7). The data from these interviews were organized via affinity diagramming. As previous noted, many of the recurring themes present in these expert interviews echoed the high-level needs present in national curricula [82, 30] (e.g., not enough civic skill training, the importance of perspective taking), but the interviews offered new insights as well. For example, many of the teachers emphasized the importance of relevance in keeping students engaged – a principle that guided the design of much of our game content.

5.2.1 Iterative Designs

The game described at the end of this chapter was preceded by a number of previous iterations. Here we will describe some of the key points in this iterative design process, and how they relate to or informed the ultimate design.

Before setting out to design the first iteration of the game, it was important to specify the primary goals of the game we intended to develop. These goals were shaped by our interviews with expert teachers and surveys of established civics curricula, as well as our review of the relevant literature. Based on these data sources, we developed a rough set of concepts, skills, and dispositions that we hoped players would understand, acquire, or develop while playing our game. Once our approximate goals were specified, we began mapping the relationships between these knowledge components. Figure 5-1 shows an early attempt at this mapping. These connections represent the foundation of future game mechanics.

Developing a context that allows players to practice these numerous skills in a coherent way (i.e., a way that preserves their authenticity and interconnectedness) proved daunting.

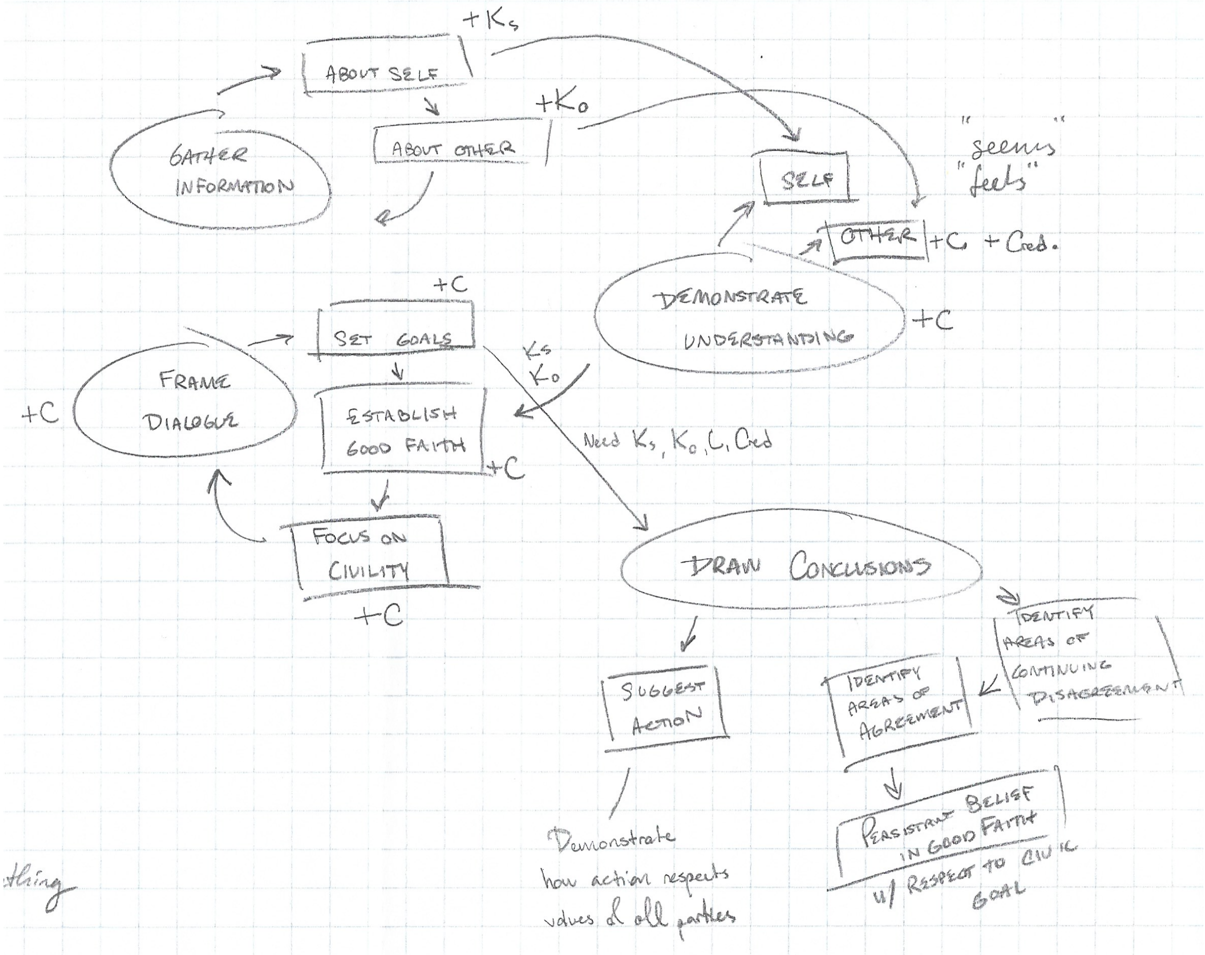
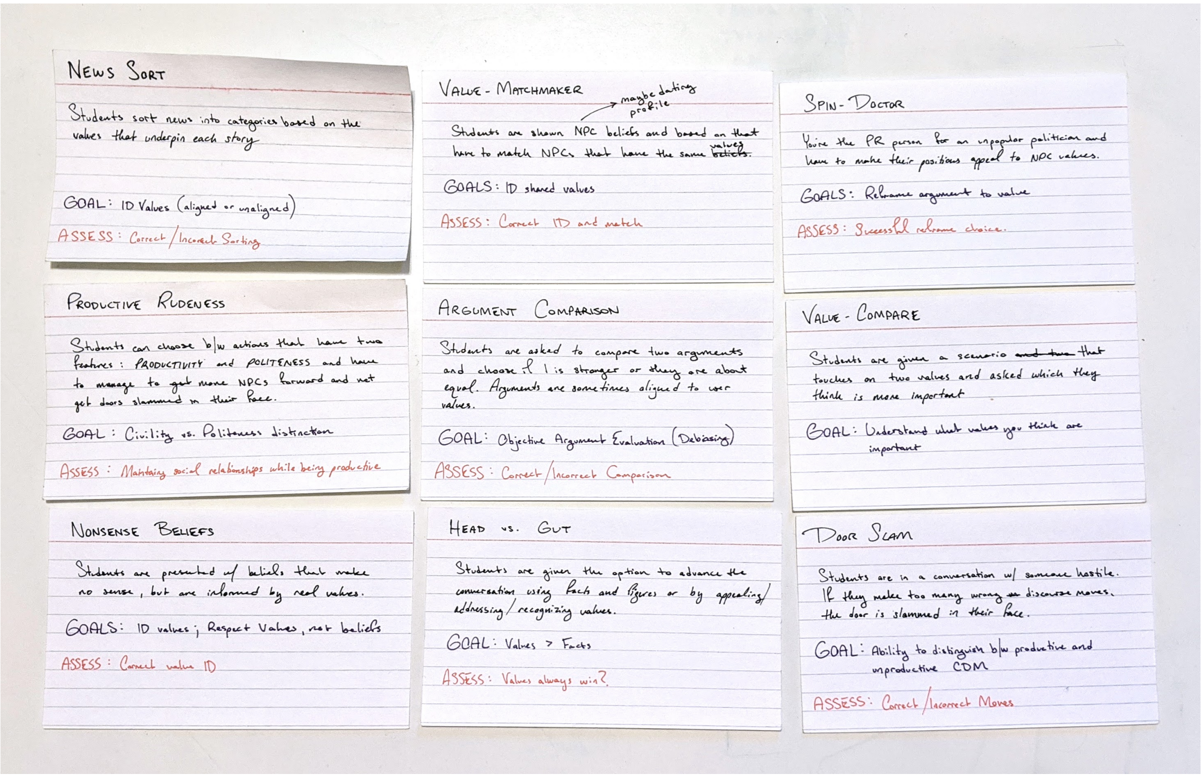


Figure 5-1: An early attempt at mapping the relationships between target concepts, skills, and dispositions. This mapping forms the foundation for many game mechanics.



News Sort

Students sort news into categories based on the values that underpin each story

GOAL: ID Values (aligned or unaligned)

ASSESS: Correct/Incorrect Sorting

Value - Matchmaker

maybe dating profile

Students are shown NPC beliefs and based on that have to match NPCs that have the same ~~beliefs~~ values.

GOALS: ID shared values

ASSESS: Correct ID and match

Spin - Doctor

You're the PR person for an unpopular politician and have to make their positions appeal to NPC values.

GOALS: Reframe argument to value

ASSESS: Successful reframe choice.

Productive Rudeness

Students can choose b/w actions that have two features: PRODUCTIVITY and POLITENESS and have to manage to get more NPCs forward and not get doors slammed in their face.

GOAL: Civility vs. Politeness distinction

ASSESS: Maintaing social relationships while being productive

Argument Comparison

Students are asked to compare two arguments and choose if 1 is stronger or they are about equal. Arguments are sometimes aligned to user values.

GOAL: Objective Argument Evaluation (Debiasing)

ASSESS: Correct/Incorrect Comparison

Value - Compare

Students are given a scenario ~~and two~~ that touches on two values and asked which they think is more important

GOAL: Understand what values you think are important

Nonsense Beliefs

Students are presented w/ beliefs that make no sense, but are informed by real values.

GOALS: ID values; Respect Values, not beliefs

ASSESS: Correct value ID

Head vs. Gut

Students are given the option to advance the conversation using facts and figures or by appealing/addressing/recognizing values.

GOAL: Values > Facts

ASSESS: Values always win?

Door Slam

Students are in a conversation w/ someone hostile. If they make too many wrong ~~as~~ discourse moves, the door is slammed in their face.

GOAL: Ability to distinguish b/w productive and unproductive CDM

ASSESS: Correct/Incorrect Moves

Figure 5-2: The initial output of the rapid-prototyping process. Each prototype was designed to be focused on a specific learning objective.

This is likely a consequence of civics' ill-definedness. Recall that, while well-defined problems can generally be divided into independent subproblems, ill-defined problems generally cannot. Instead, subtasks in ill-defined domains are often dependent on one another, requiring them to be solved in parallel.

Our solution was to generate a series of rapid prototypes, with each prototype using a specific subtask as its primary focus. This process afforded us two main benefits:

1. It allowed us to recharacterize specific subtasks narrowly, amplifying their unique features and critical dependencies. Provided each prototype authentically represents its specific subgoal, this process could illuminate how these individual prototypes could fit together as modular components of a more comprehensive whole.

2. The narrow focus of these prototypes also inspires the kinds of creative and meaningful mechanics that might not fit into the coherent story of a more comprehensive system.

Figure 5-2 shows the initial output of this rapid-prototyping process. Several of the mechanics outputted by this process remained integrated into the ultimate design. An additional side-effect of this process is the opportunity to pursue unused mechanics as directions for future work.

The next step was integrating these mechanics into a coherent system. We developed a series of paper prototypes to test our designs. The first iteration (see Figure 5-3) required players to manage six factors that impact an NPC's willingness to engage in productive civil discourse. These factors included concepts like *Approval of [the proposed] Plan*, *Sense of Community*, and

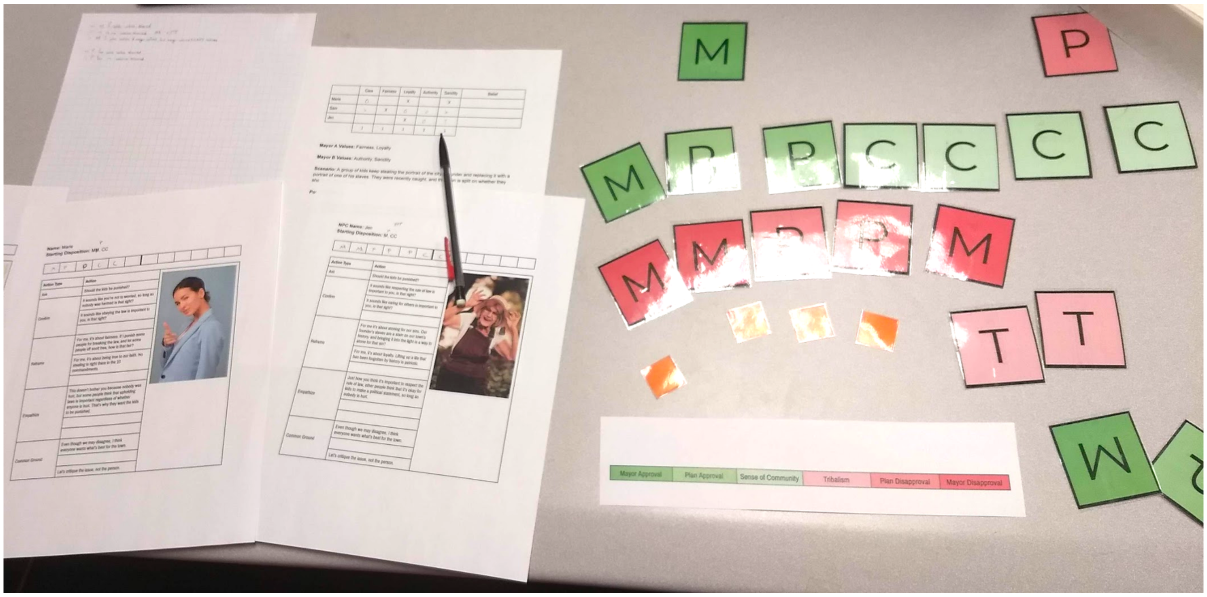


Figure 5-3: The first paper-prototype iteration. An excess of resource-management resulted in a functional, but mostly joyless experience.

Tribalism. Figure 5-4 shows a graph of the relationships between these six factors. This resource-management heavy gameplay resulted in a game that was functional, but not enjoyable. One playtester summarized his experience with a charitable, "It was logically sound."

The six factor model was abandoned for a simpler model, in which a variety of actions can increase or lower just one factor: *Tribalism*. Figure 5-5 shows a later iteration of the game, where the types of interactions players experience begin to resemble the kinds of interactions seen the ultimate design. This iteration required that players not only identify an NPC's value, but also identify the kind of discourse move that would be most effective in persuading them (i.e., *Empathize*, *Reframe*, or *Respect*). Ultimately, this second identification task tended to restrict player choices (which was not enjoyable) and divert attention away from the more crucial value-identification task. We eliminated this source of extraneous load by reconfiguring the *Empathize* and *Respect* actions as additional options the player could choose to employ if they wish. This elevated the *Reframe* action to the core mechanic of the game, with the *Reframe* action ultimately becoming the *Persuade* action.

5.2.2 Learning Objectives

Based on the insights gained from our structured interviews, we generated a small set of primary learning objectives for our game. Upon completion of the game:

1. Students should understand and be able to identify the values that motivate specific beliefs.

2. Students should be able to identify shared values and use those shared values to bridge ideological divides.

3. Students should understand the barriers to productive civil discourse (e.g., tribalism, misinformation) and the discourse moves that can help discussants overcome those barriers.

4. When given a scenario, students should be able to choose the most productive civil discourse move.

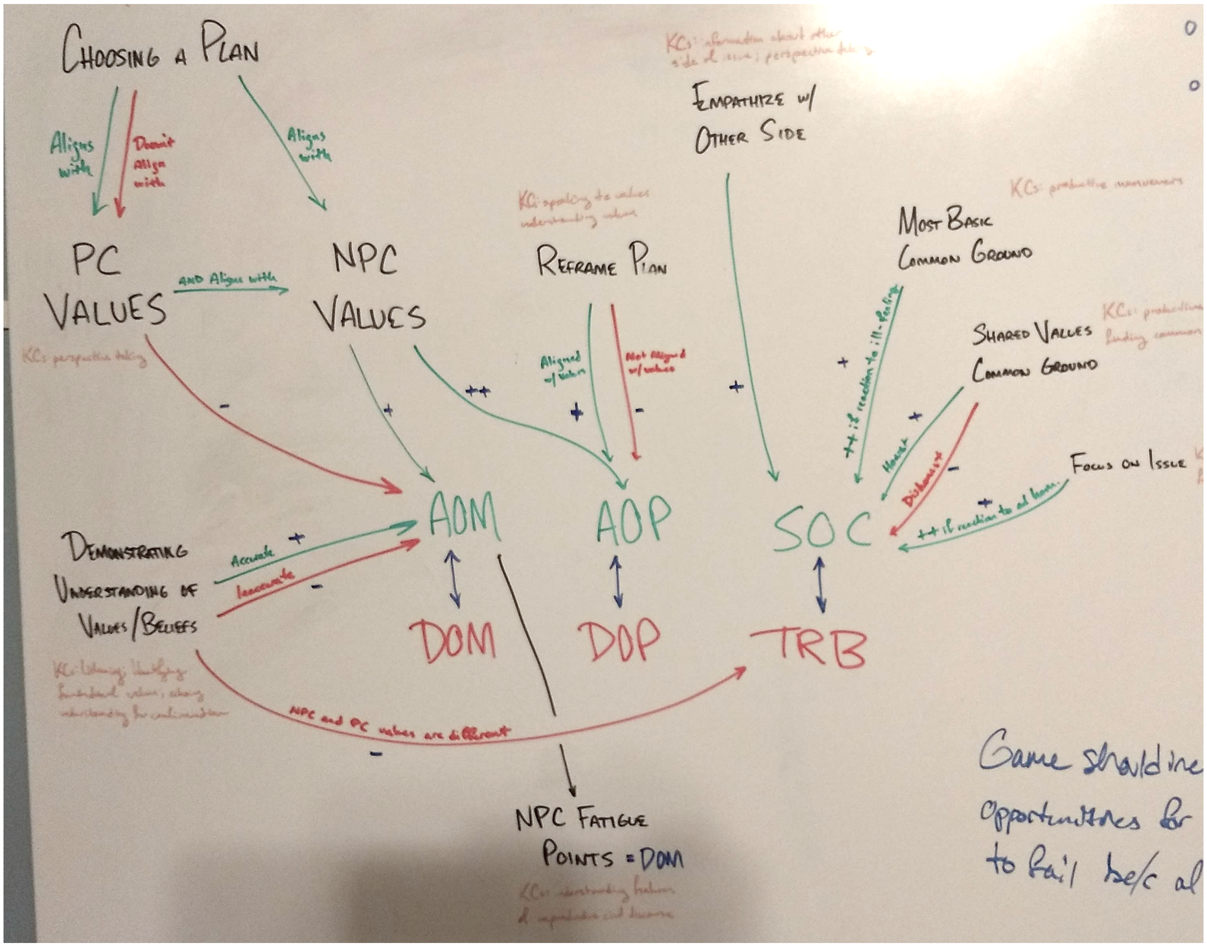

Choosing a Plan
Aligns with
Doesn't Align with
Aligns with
PC Values
AND Aligns with
NPC Values
KCs: perspective taking
-
+
++
KCs: speaking to values understanding values
Reframe Plan
Aligned w/ values
Not Aligned w/ values
+
-
KCs: information about other side of issue; perspective taking
Empathize w/ Other Side
+
KCs: productive maneuvers
Most Basic Common Ground
+
++ if reaction to ill-feeling
Shared Values Common Ground
KCs: productive finding common
+
Honest
Dishonest
-
Focus on Issue
+
++ if reaction to ad hom.
AOM
AOP
SOC
DOM
DOP
TRB
Demonstrating Understanding of Values/Beliefs
Accurate
+
Inaccurate
-
KCs: listening; identifying fundamental values, showing understanding for communication
NPC and PC values are different
-
NPC Fatigue Points = DOM
KCs: understanding features of unproductive civil discourse
Game should inc
opportunities for
to fail b/c of

Figure 5-4: Relationships between the six key factors present in the first iteration of the game. AOM=Approval of Mayor (the player's character), AOP=Approval of Plan, SOC=Sense of Community, DOM=Disapproval of Mayor, DOP=Disapproval of Plan, TRB=Tribalism.

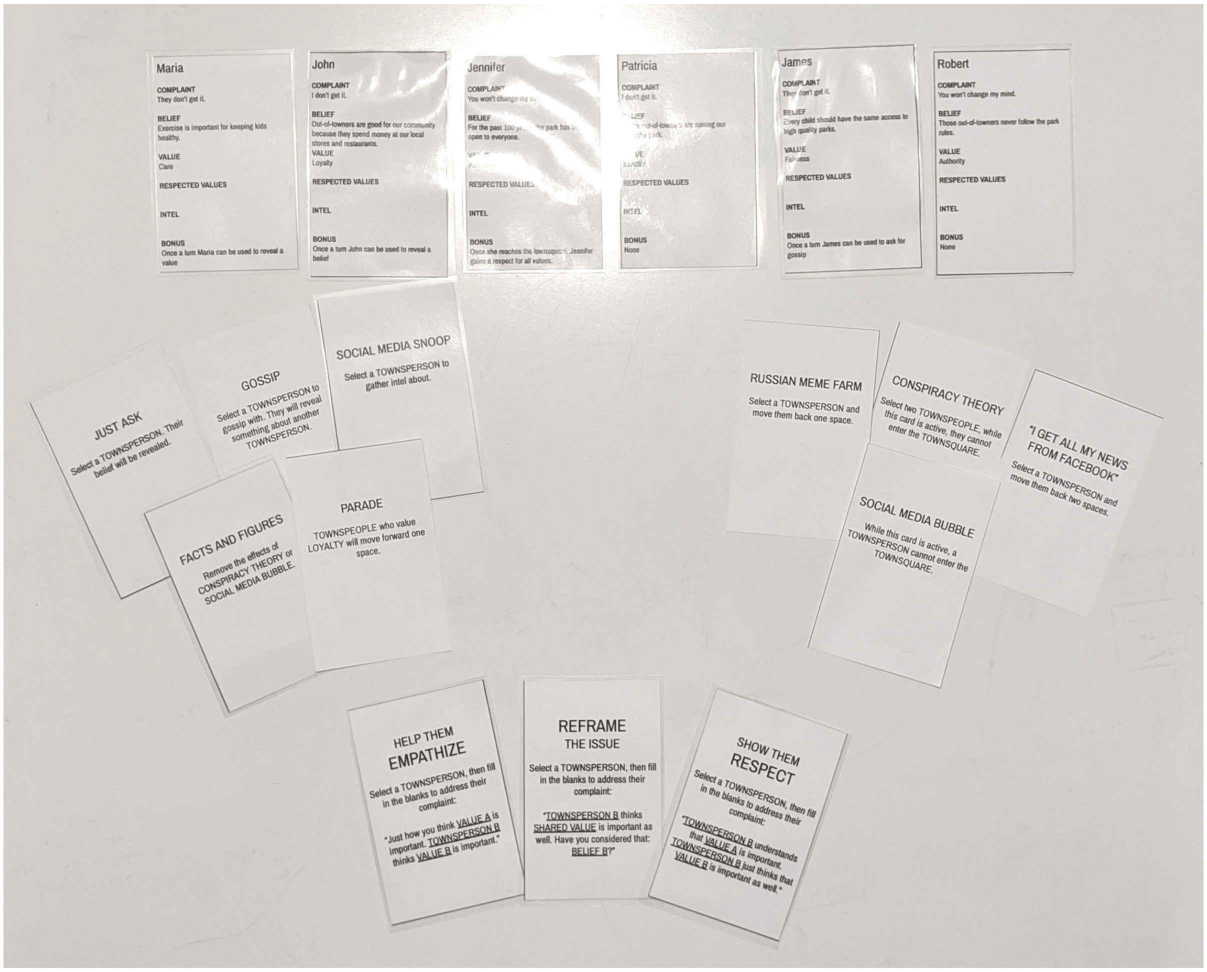
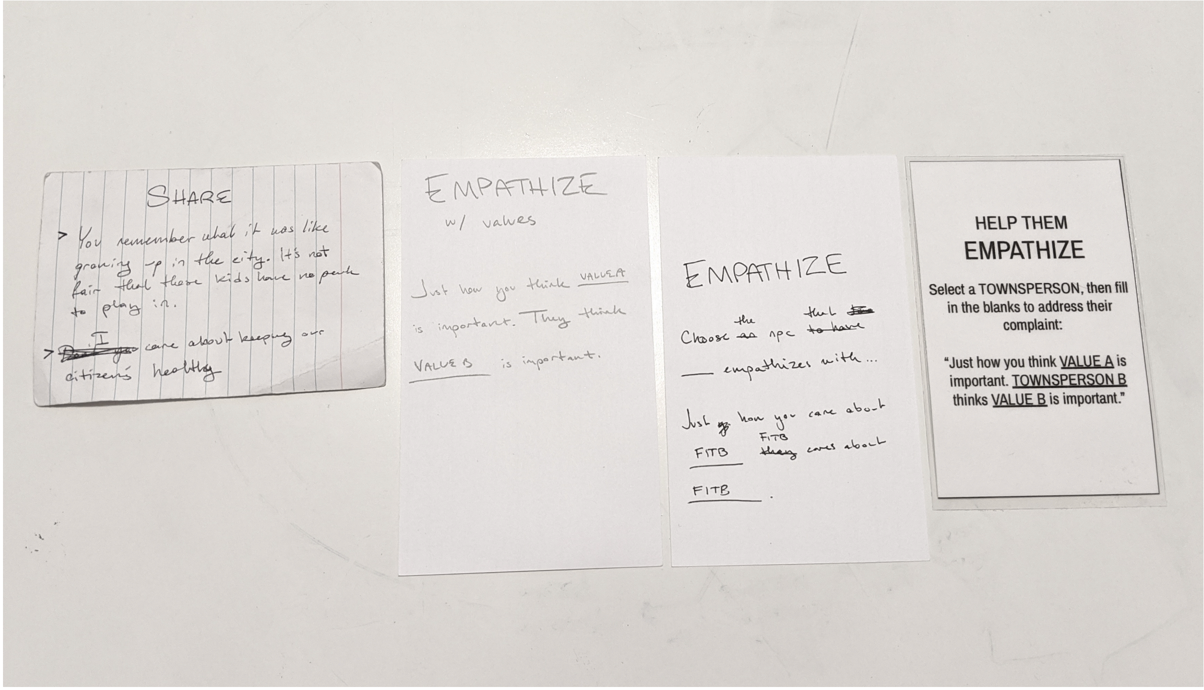

Maria
COMPLAINT
They don't get it.
BELIEF
Exercise is important for keeping kids healthy.
VALUE
Care
RESPECTED VALUES
INTEL
BONUS
Once a turn Maria can be used to reveal a value
John
COMPLAINT
I don't get it.
BELIEF
Out-of-towners are good for our community because they spend money at our local stores and restaurants.
VALUE
Loyalty
RESPECTED VALUES
INTEL
BONUS
Once a turn John can be used to reveal a belief
Jennifer
COMPLAINT
You won't change my
BELIEF
For the past 100 the park has open to everyone.
RESPECTED VALUES
INTEL
BONUS
Once she reaches the townsquare, Jennifer gains a respect for all values.
Patricia
COMPLAINT
I don't get it.
out-of-towners are ruining our park.
Sanctity
RESPECTED VALUES
INTEL
BONUS
None
James
COMPLAINT
They don't get it.
BELIEF
Every child should have the same access to high quality parks.
VALUE
Fairness
RESPECTED VALUES
INTEL
BONUS
Once a turn James can be used to ask for gossip
Robert
COMPLAINT
You won't change my mind.
BELIEF
Those out-of-towners never follow the park rules.
VALUE
Authority
RESPECTED VALUES
INTEL
BONUS
None
JUST ASK
Select a TOWNSPERSON. Their belief will be revealed.
GOSSIP
Select a TOWNSPERSON to gossip with. They will reveal something about another TOWNSPERSON.
SOCIAL MEDIA SNOOP
Select a TOWNSPERSON to gather intel about.
FACTS AND FIGURES
Remove the effects of CONSPIRACY THEORY or SOCIAL MEDIA BUBBLE.
PARADE
TOWNSPEOPLE who value LOYALTY will move forward one space.
RUSSIAN MEME FARM
Select a TOWNSPERSON and move them back one space.
CONSPIRACY THEORY
Select two TOWNSPEOPLE, while this card is active, they cannot enter the TOWNSQUARE.
"I GET ALL MY NEWS FROM FACEBOOK"
Select a TOWNSPERSON and move them back two spaces.
SOCIAL MEDIA BUBBLE
While this card is active, a TOWNSPERSON cannot enter the TOWNSQUARE.
HELP THEM
EMPATHIZE
Select a TOWNSPERSON, then fill in the blanks to address their complaint:
"Just how you think VALUE A is important, TOWNSPERSON B thinks VALUE B is important."
REFRAME
THE ISSUE
Select a TOWNSPERSON, then fill in the blanks to address their complaint:
"TOWNSPERSON B thinks SHARED VALUE is important as well. Have you considered that: BELIEF B?"
SHOW THEM
RESPECT
Select a TOWNSPERSON, then fill in the blanks to address their complaint:
"TOWNSPERSON B understands that VALUE A is important. TOWNSPERSON B just thinks that VALUE B is important as well."

Figure 5-5: A later paper-prototype of the game. This iteration shares many of the same types of interactions as the ultimate version, however the core mechanic of this game was unnecessarily complicated by an additional identification task.



SHARE

> You remember what it was like growing up in the city. It's not fair that these kids have no park to play in.

> I care about keeping our citizens healthy

EMPATHIZE

w/ values

Just how you think VALUE A is important. They think VALUE B is important.

EMPATHIZE

Choose ~~an~~ the npc that ~~to have~~

___ empathizes with...

Just how you care about FITB FITB ~~they~~ cares about FITB.

HELP THEM

EMPATHIZE

Select a TOWNSPERSON, then fill in the blanks to address their complaint:

"Just how you think VALUE A is important. TOWNSPERSON B thinks VALUE B is important."

Figure 5-6: A progression of the *Empathize* action across various paper prototype iterations.

These learning objectives can be thought of as easily communicable descriptions of what are likely constellations of more specific knowledge components. Towards that end, we target the specific knowledge components listed in Table 5.1.

The instruction and opportunities to practice these skills were scaffolded according to the instructional plans depicted in Figures 5-7 and 5-8.

5.2.3 What the Game Isn't

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

1. **It is not about changing beliefs.** Our goal is not to give students tricks that 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. **It 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. Instead, this game is about finding shared values and using those values as the foundation of an actionable solution.

3. **It does not confuse civility with politeness.** Instead, we subscribe to Papacharissi's [59] 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.

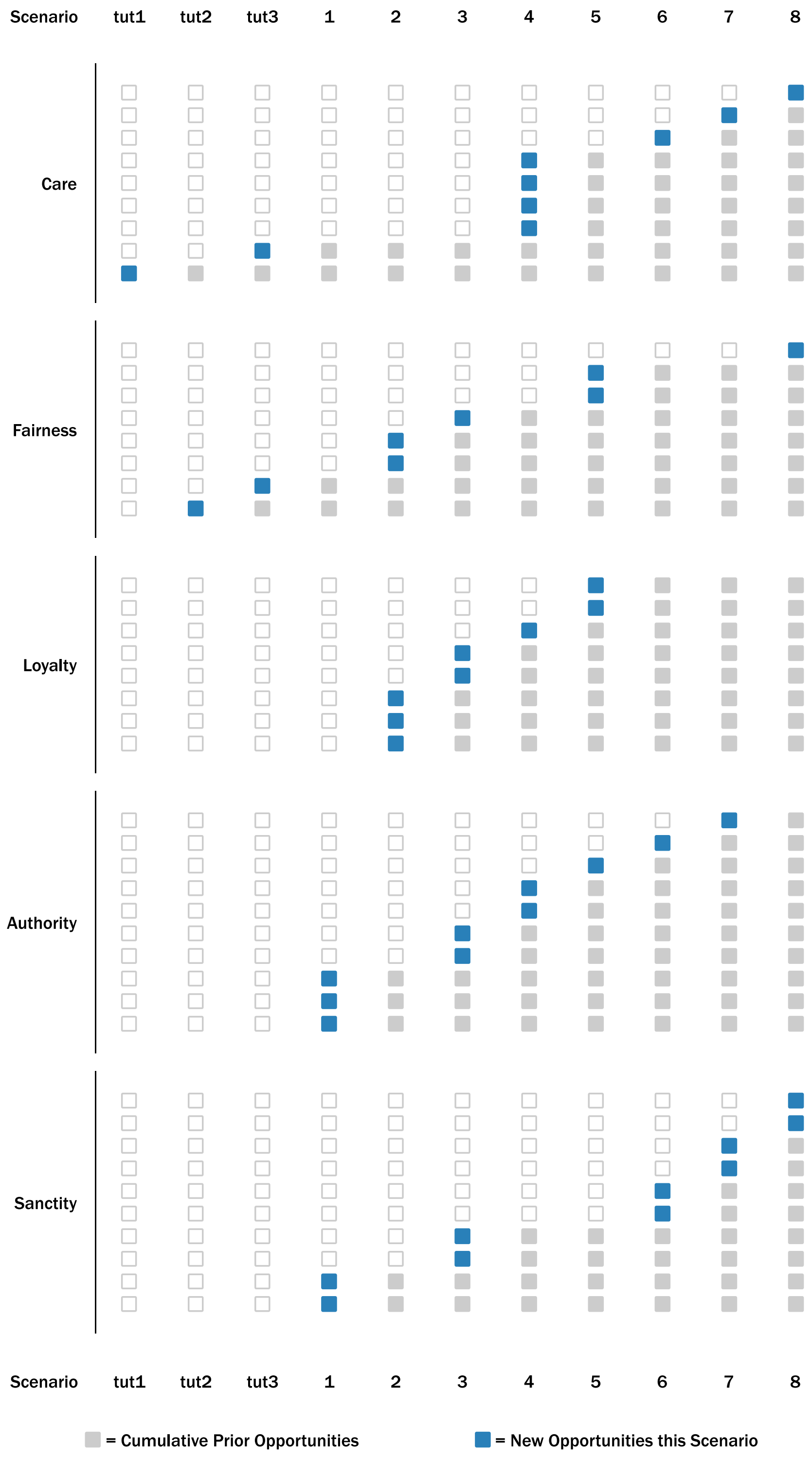


Figure 5-7: The number of opportunities students have to practice identifying each of the five moral foundations across all scenarios. Solid blue squares indicate new opportunities experienced during that scenario. Solid gray squares indicate the cumulative number of prior opportunities experienced up to this point.

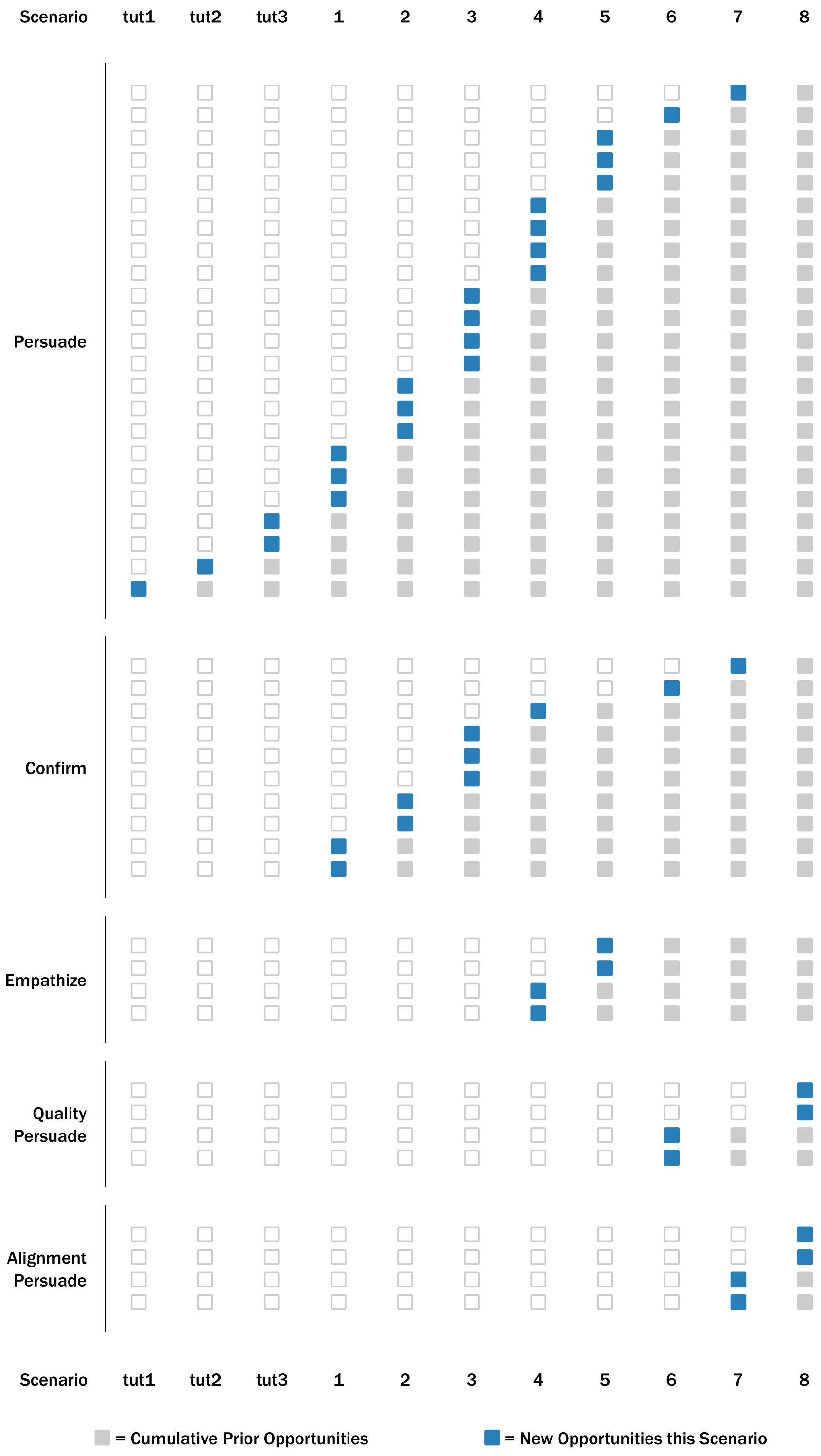


Figure 5-8: The number of opportunities students have to practice the five main task types across all scenarios. Solid blue squares indicate new opportunities experienced during that scenario. Solid gray squares indicate the cumulative number of prior opportunities experienced up to this point.

ID	KC	Description
KC1	identifyCare	User can identify the CARE foundation in a given text
KC2	identifyFairness	User can identify the FAIRNESS foundation in a given text
KC3	identifyLoyalty	User can identify the LOYALTY foundation in a given text
KC4	identifyAuthority	User can identify the AUTHORITY foundation in a given text
KC5	identifySanctity	User can identify the SANCTITY foundation in a given text
KC6	persuade	User can identify which argument from the opposing side that appeals to a particular value
KC7	convoReset	User can correctly use a *Conversation Reset* to lower tribalism
KC8	pointAgree	User can correctly use a *Point of Agreement* to lower tribalism
KC9	quality	User chooses high-quality arguments over low-quality arguments
KC10	alignQuality	User chooses low-quality aligned arguments over high-quality unaligned arguments

Table 5.1: Knowledge Components targeted by our instructional design.

4. **It 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.

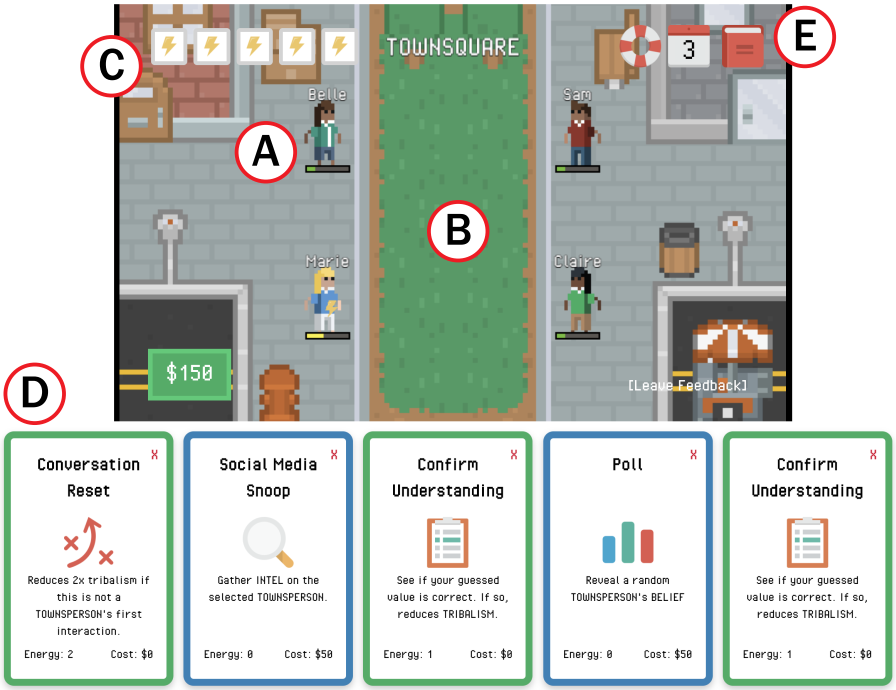
5.3 Design Elements

This section describes some of the intentional design choices that shaped the goals, mechanics, and content of the game. But first, it may be helpful to have a working knowledge of the premise and primary objective of the game:

> In *Persuasion Invasion*, the world is under attack by a species of aliens invaders. However, because these aliens are pacifists, they don't conquer planets through warfare. Instead, the aliens conquer planets by slowly sowing discord and division in small communities. When the community is unable to come together to solve even the most basic problems – that's when they strike! Players assume the role of a government agent, and are tasked with bringing communities together. They do this by employing civil discourse moves that reduce tribalism and by identifying shared values across political lines.

5.3.1 Alien Invasion

Using an alien invasion as the overarching threat serves a number of purposes. First, the kinds of activities that the aliens engage in throughout the game (e.g., sowing division, spreading disinformation) help to establish the alien invasion as a soft metaphor for the threat of foreign



Conversation Reset
Reduces 2x tribalism if this is not a TOWNSPERSON's first interaction.
Energy: 2 Cost: $0

Social Media Snoop
Gather INTEL on the selected TOWNSPERSON.
Energy: 0 Cost: $50

Confirm Understanding
See if your guessed value is correct. If so, reduces TRIBALISM.
Energy: 1 Cost: $0

Poll
Reveal a random TOWNSPERSON's BELIEF
Energy: 0 Cost: $50

Confirm Understanding
See if your guessed value is correct. If so, reduces TRIBALISM.
Energy: 1 Cost: $0

Figure 5-9: An annotated screenshot of a scenario. In each scenario, players must persuade NPCs like Belle (A) to move into the TOWNSQUARE (B). To do this players must identify which argument from the opposing side appeals to what Belle values. Persuading NPCs costs Energy (C). The bar positioned below each NPC represents their political tribalism. Players must reduce an NPC's tribalism before attempting to persuade them. They can do this by playing Discourse Cards (D) like *Conversation Reset*. Finally, the action menu (E) allows the player to request a hint, end the day/turn, or reference information in their notebook.

interference in our elections. The idea of an alien invasion is also instantly understandable and motivating, and that is in part due to the second (and more important) reason for using an alien invasion: existential threats change our perceptions of what is important. Although the players will be grappling with emotionally evocative political issues that might normally polarize communities into their political tribes, the ever-present existential threat will constantly remind players of the community's shared goals and interests – effectively positioning both sides of the political divide as part of a larger, world-tribe.

5.3.2 Townsquare

In each level of the game, the player's primary objective is to persuade each townsperson to move into the Townsquare. Again, this movement into the Townsquare does not represent a change in belief, but simply a willingness to engage in earnest and meaningful dialogue with those they disagree with. Townspeople with different beliefs are positioned either to the left or right of the townsquare. This is intentionally meant to evoke the prevailing (if not overly-simplistic) real-world dichotomy of America's two-party system. The main purpose of the positioning of townspeople is so that, as players persuade individual townspeople to move into the Townsquare, that productive action is reinforced visually as players see the townspeople physically coming together into a shared space.

5.3.3 Isolation

While it may seem counter-intuitive to design a single-player game for teaching discourse skills (a social activity), there are actually some key benefits of isolated gameplay. First, by taking players out of their normal social context, players can explore their own values and beliefs without pressure from their peers (or family) to conform to the established beliefs of a particular tribe. Second, the simulated social context of our game allows students to take risks and understand the consequences of discourse actions in a way that is impossible in real social contexts, where a poor decision could cause the student to lose face with their peers.

5.3.4 Time Pressure

Players are given a small number of days (generally 3) to solve the problem before aliens invade. This time pressure raises the stakes of the, relatively trivial, issues laid out in the level's scenario – which otherwise might not appear to be time sensitive. We also hope that the urgency introduced by the time pressure helps students to understand politics as not merely a set of mostly inconsequential issues, but rather as a highly consequential method for solving problems as a society.

5.3.5 Audiovisual Language

In a context where players are accustomed to only seeing video games with high production value, educational games can be quickly identified and written-off based solely on their modest audiovisual elements. This is particularly true for games portraying human beings, which can instantly appear dated relative to commercial titles. Rather than invite the comparison, we chose to adopt a retro, 8-bit visual style [44, 12]. We complement this visual style with music [58] and sound effects [78] from the chiptunes genre (i.e., a style that mimics the synthesized music of older video games).

5.4 Game Content

5.4.1 Scenarios

Each level of the game focuses on a scenario, which describes a topic or decision that has divided public opinion (e.g., "Should students be required to wear uniforms?"). Based on insights from our structured interviews with expert instructors, we designed each scenario to be both emotionally evocative and relevant to student lives. While some of the fictional scenarios have real-world analogues (e.g., immigration), we intentionally avoided using well-known, hot-button issues, as students might be inclined to, for example, simply regurgitate the beliefs of their parents about the topic. Using an analogue forces students to form their own beliefs, and hopefully better understand their values in the process.

5.4.2 Discourse Cards

In *Persuasion Invasion*, civil discourse moves are represented by cards that the player can spend energy and/or money to play. Discourse Cards are divided into three categories: *Tribalism Cards* which reduce a townsperson's tribalism, *Intel Cards* which provide clues to a townsperson's values, and *Value Cards* which appeal directly to a certain value.

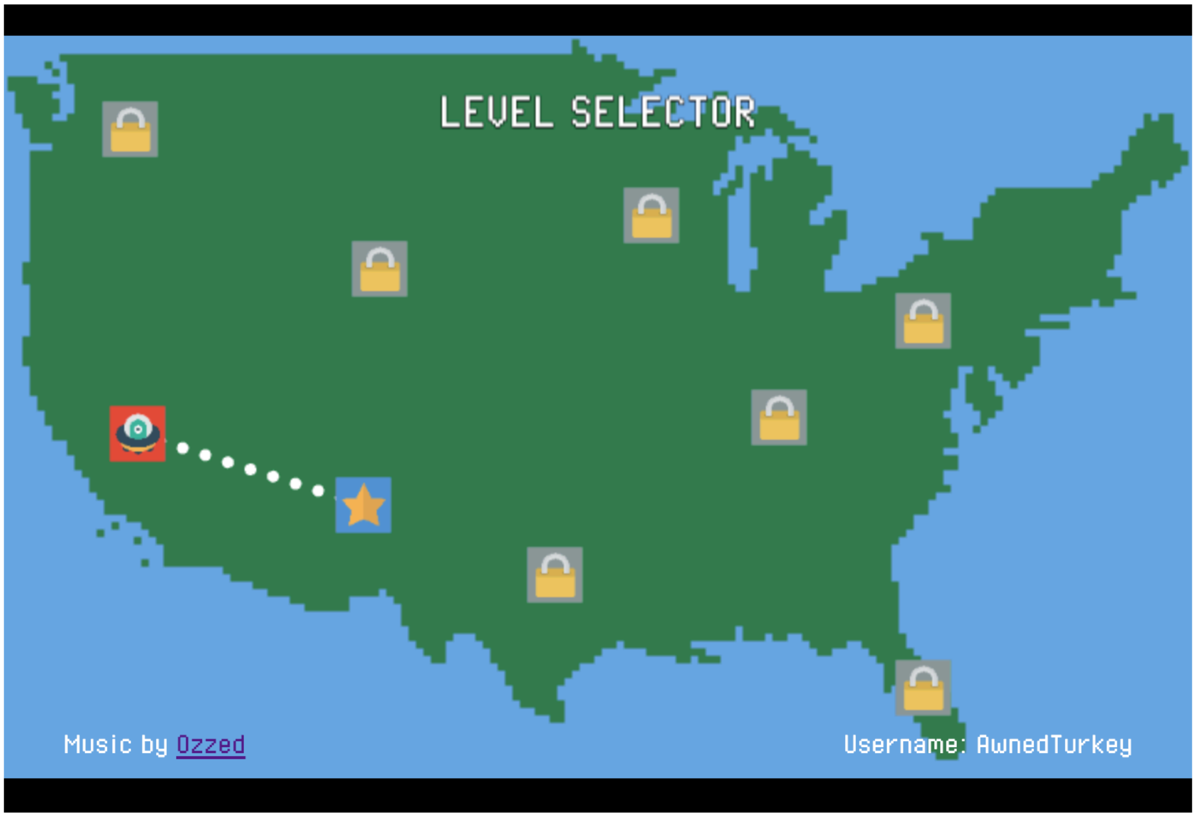


Figure 5-10: A screenshot of the level selector interface. Levels with a star icon have been completed, levels with a UFO icon are unlocked but not completed, and levels with a lock icon are not unlocked. Players complete levels in a set sequence, with the completion of one level unlocking the next.

5.4.3 Unproductive Discourse Moves

In the current iteration, the player only has access to productive discourse cards, but unproductive discourse cards (i.e., cards that increase tribalism) are available to the alien invaders. The inclusion of unproductive discourse moves helps define what make the productive discourse moves good. Additionally, although unproductive moves necessarily have a negative consequence, astute players may be able to derive useful information about a townsperson's values even from this negative interaction. This is similar to how seeing a friend share a debunked news story is undeniably bad, but it can also reveal what kinds of things that friend cares about.

5.4.4 Notebook

Playtesters have described the gameplay as similar to being a detective and slowly uncovering the beliefs and values of a town through these interactions. To help keep track of all this information, players are given a digital notebook that automatically updates when they gather intel about a townsperson. This notebook also allows players to track what values they might think a townsperson cares about.

5.4.5 Currencies

There are three main currencies that players use throughout the game:

Energy

Energy is tied to the idea that, in real-life, engaging in productive civil discourse is exhausting, and our capacity for fighting against our tribalistic instincts is a limited resource. Discourse moves cost more energy in the game when they generally require more mental energy in real life (e.g., starting from a point of agreement or resetting the conversation when it gets heated).

Money

The inclusion of money as a currency is meant to help students understand the real role of money in identifying and appealing to our values. The monetary cost of discourse moves has some basis in their *relative* real-world cost, but is also influenced by how valuable the card is for gameplay. However, these two factors often align.

Tribalism

Tribalism represents a townsperson's unwillingness to set aside the powerful social influence of their group identity (e.g., liberal or conservative) in order to earnestly consider a viewpoint from the opposing side. We include tribalism (i.e., how tribalistic each individual townsperson is) as a currency because although players do not explicitly pay for cards with tribalism, some cards can "cost" (increase) tribalism as a consequence of playing them. For example, the *Gossip* card reveals a piece of *Intel*, but increases tribalism as well.

5.5 Gameplay

Gameplay is turn-based and modeled on the concept of days, with players acting during the day and aliens acting during the night.

5.5.1 Player's Turn

At the start of each day, the player's energy is restored to 5 Energy Points, and they gain some additional money. Additionally, new cards are automatically drawn from the deck until the player has a hand of 5 cards. During their turn, players can spend their energy and money to play discourse cards from their hand, which reduce tribalism and reveal clues about what a townsperson values. When the player believes they have identified the value of a townsperson and have reduced their tribalism, they can try to *Persuade* that townsperson to move into the townsquare (which costs 2 Energy Points). The *Persuade* action asks players to choose an argument from the opposing side that appeals to the target townsperson's value. For example, if a townsperson values the *Care* foundation and is concerned about the health risks involved in letting kids play football, you might present the argument that, "Football teaches skills like leadership and determination. It does more good than harm." This is a belief held by the opposing side, but it still appeals directly to the townsperson's value (Care).

5.5.2 Alien's Turn

When the player exhausts their energy or money, they can choose to end the day. Night falls and the alien's turn begins. During which, the alien can play their own cards to increase tribalism. The effects of these cards are illustrated with the use of a spaceship tractor beam that highlights the townsperson being targeted. Importantly, aliens cannot effect a townsperson who has already moved into the townsquare.

5.5.3 Mayor's Turn

Starting on the morning of the second day, the mayor is given a brief turn as well. The mayor is a reluctant ally of the player, helping the player by reducing the cost of cards, but only if the player can prove themselves by answering questions. These questions act as additional opportunities to assess the student's progress towards mastering the key learning objectives of the game. If students answer the mayor's question correctly, they are rewarded with some free cards that turn.

5.6 Addressing the Challenging Features of Civics Education

Now that we've laid out the basic structure and mechanics of gameplay, we can return to the three challenging features of civics education that we discussed previously: the **collectivist** nature of civic action, the **feelings of powerlessness** or insignificance associated with being one piece of an enormous system, and the fact that progress is almost always made at a **near-indiscernible pace**. Recall that, ideally, the game we design would address these features in a way that preserves their spirit while minimizing the negative impact they may have on motivation. Meeting our transformational goals means striking a balance between authenticity and enjoyment.

We mitigate **feelings of powerlessness** by making the player assume the role of a government agent. This makes the protagonist a citizen, rather than, say, a super hero who is uniquely capable of saving the world. Instead, any additional power the protagonist holds is derived from their position in society (as a government agent), and the unique knowledge and skills that the player learns alongside the character they are playing. Importantly, the protagonist is not named, and only referred to as "Agent." This allows the player to more easily assume the identity of the protagonist, further breaking down the barrier between the two. We expect this to not only result in greater immersion, but also in increased feelings of civic empowerment.

We address **the near-indiscernible pace of progress** by striking a balance between immediate payoffs and slow, incremental intra- and inter-level progress. First, we have players work within small towns, where the impact of their actions is more easily seen. This is perhaps the largest concession to enjoyment at the expense of authenticity, as successful civic action in these communities only requires persuading a few townspeople to move into the *Townsquare*. However, it is difficult to imagine sustaining motivation if players were asked to persuade a realistic number of people. Alternatively, we could have had players try to influence smaller groups of people (e.g., clubs or organizations), but reducing the size of the group also reduces the relative importance of player actions.

The second largest concession to enjoyment is the immediacy of feedback. Players receive both correctness and descriptive feedback after certain correct and incorrect actions. While this is, of course, not representative of real-life civil discourse, the consequences of correct and incorrect actions are designed to align with real-world consequences. For example, engaging in gossip with one NPC increases the political tribalism of a random NPC. Conversely, correctly confirming your understanding of an NPC's value decreases their tribalism.

The game also fails to simulate faulty feedback mechanisms in civil discourse. For example, a witty retort or personal attack may lead to feelings of superiority or righteousness that could be mistaken as personally positive outcomes of civil discourse, despite generally being unproductive discourse moves. This is an area of future work in this domain.

The reduced scale of small towns combined with immediate feedback results in easily discernible progress. We balance these measures designed to sustain motivation by slowing the pace of progress both within and across levels. Each level (or scenario) is played out over a maximum of three days. This timescale is short, but allows us to limit the impact of any one action on an NPC so that the effect is relatively small. Players also encounter many scenarios distributed across a map of the United States. This is meant to remind players of the real scale of the problem, while also effectively dividing the problem into manageable pieces.

Finally, we address the **collectivist** nature of civic action by taking players out of competition with one another and into competition with an external force. By making the antagonists of the game invading aliens, we continually remind players that, while the NPCs may be divided into political tribes along some trivial issue, they are united in their membership to a national or world tribe. In other words, in reducing tribalism, the goal of the player is not to persuade NPCs to adopt the beliefs of a particular tribe, but rather to adopt an expanded definition of their tribal membership. A pending alien invasion offers a simple and intrinsically motivating reminder of at least one shared goal: continued survival.

5.7 Conclusion

In this chapter, we discussed some of the unique challenges and affordances of games in the civic education space, our iterative development process, and finally a description of the iteration that we ultimately deployed in classrooms. In the next chapter, we will discuss the details of that deployment and our methods for evaluating the effectiveness of the game we developed in relation to our transformational goals.

Chapter 6

Experiment Design and Methodology

"Thanks for teaching us this persuasion stuff – it worked on my mom last night."
- A student during the pilot study

A small pilot study and larger, full-scale experiment were designed to evaluate the effectiveness of our game in relation to our transformational goals. In this chapter, we will discuss the structure of these experiments and the materials used to collect student data. Finally, we will detail the methods we used to detect gaming behavior, measure learning, and estimate the impact of bias on student performance.

6.1 Experimental Materials

6.1.1 Pre-Test Questionnaire

Immediately after logging into the game's website for the first time, students were directed to a short pre-test consisting of the following two sub-questionnaires:

Moral Foundations Theory Questionnaire

As in prior work, all students were asked to complete a modified[1] Moral Foundations Theory Questionnaire. Students completed this questionnaire as part of the pre-test given before playing the game. This may not be an ideal sequence (as taking the questionnaire before playing the game may alter how they play the game), but it is unfortunately necessary, as the questionnaire responses drive the value-adaptive intervention. See Section 4.3.1 for a more detailed description of Moral Foundations Theory and see Appendix A for a complete list of the questions used in the modified Moral Foundations Theory Questionnaire.

Skill Assessment Questionnaire

In addition to the Moral Foundations Theory Questionnaire, the pre-test also included six skill assessment questions. These questions were designed to capture a student's prior knowledge or ability with respect to the primary learning objectives of the game: choosing the most

[1] Students in the pilot study completed the standard Moral Foundations Theory Questionnaire. However, several of these 12th grade AP students were unsure of the definitions of certain terms. Therefore, some of the questionnaire's wording was modified to be more reading-level appropriate for the full-scale experiment.

productive discourse move, identifying values, choosing high- over low-quality arguments, and choosing aligned over unaligned arguments. See Appendix B for a complete list of the questions asked.

6.1.2 Post-Test Questionnaire

After the completion of the final scenario or at the end of the final gameplay session (whichever came first), students were directed to a post-test consisting of the following sub-questionnaires:

Player Experience Questionnaire

In addition to designing a game that efficiently teaches key civil discourse skills, another key goal of this work was to design a game that is authentic, meaningful, and enjoyable to play. To capture data about the experience players had while playing the game, we included a series of questions adapted from the Player Experience Inventory (PXI) [84] (see Appendix C for a full list of questions). We added two additional questions to those adapted from the PXI, a question explicitly about *learning* (as the *mastery* question may be ambiguous) and a question about *long-term change*.

Qualitative Feedback

In addition to the above quantitative measures of player experience, we also provide students with three opportunities to provide open-ended feedback. These three open-ended questions were presented to students:

1. Do you have any additional feedback/suggestions about the game?
2. What did you like about this game?
3. What did you dislike about this game?

Class-Specific Questions

We also asked students to directly relate their experience playing the game to their normal class content. Several of these questions were generated by the teacher overseeing the classes used in the pilot experiment, who was given the opportunity to add questions to the post-test questionnaire. These class-specific questions included:

1. What skills from class helped you while playing this persuasive game? Please check all that apply.
    - recognizing or using claims of value, fact, policy, or definition
    - finding common ground with your specific audience
    - acknowledging opposing arguments
    - employing counterarguments
    - understanding which appeal (emotional or logical) is best to choose for your specific audience
    - avoiding logical fallacies

2. Was this an effective way to practice the new knowledge you have learned in class? (yes or no)

3. Should your teacher incorporate this game into the course? (yes or no)

Demographic Questions

Finally, we ask students to answer a small number of demographics questions:

1. What race/ethnicity do you identify as?

2. Gender

3. I play video games (Regularly, Sometimes, Rarely, Almost Never)

Note that we did not ask for student age, as it is expected to be mostly redundant with grade level.

6.1.3 Anonymous Player Usernames

Each student was given a system-generated username on an index card, and then asked to write their name below the username. These cards were collected at the end of class to be distributed by the instructor at the beginning of the next class. This method ensures 1) that the students do not forget or misspell their usernames, and 2) that the student's data is anonymized at the point of collection, as the researchers never collect the student's real names.

These usernames were generated by simply combining an adjective (selected randomly from a list) to an animal (also selected randomly from a list), and then querying the server to ensure that the new username was indeed unique. It is worth noting that these usernames proved to be a source of joy throughout the experiment, with teachers, for example, joking that, "This is how I'm going to refer to [some student] from now on." An additional unanticipated benefit of this method was the memorableness of the usernames. By the third session, many students did not need to reference their username card to login to the system.

6.1.4 Data Servers

Data was collected at the transaction level and logged to a database hosted on a secure server. Location data was not collected, nor were user IP addresses.

6.2 Pilot Study

A pilot study was conducted to test the functionality and potential effectiveness of the game before a full-scale implementation. Three sections of a local high school's 12th Grade AP English Language and Composition course were recruited by their instructor for participation. The game was administered as part of normal classroom instruction, though students (n=48) were given the option to opt-out of data-analysis (no students chose to opt-out).

After logging in for the first time, students were directed to the pre-test described above. After completing the pretest, students were directed to the game proper, beginning with a short tutorial. After completing the tutorial, the students were tasked with completing a series of eight levels (see Chapter 5 for a complete description). Students were given three 40 minute class periods to play the game. Many students used close to the entire time allotted to finish the game, though some students completed the game in as little as two class periods.

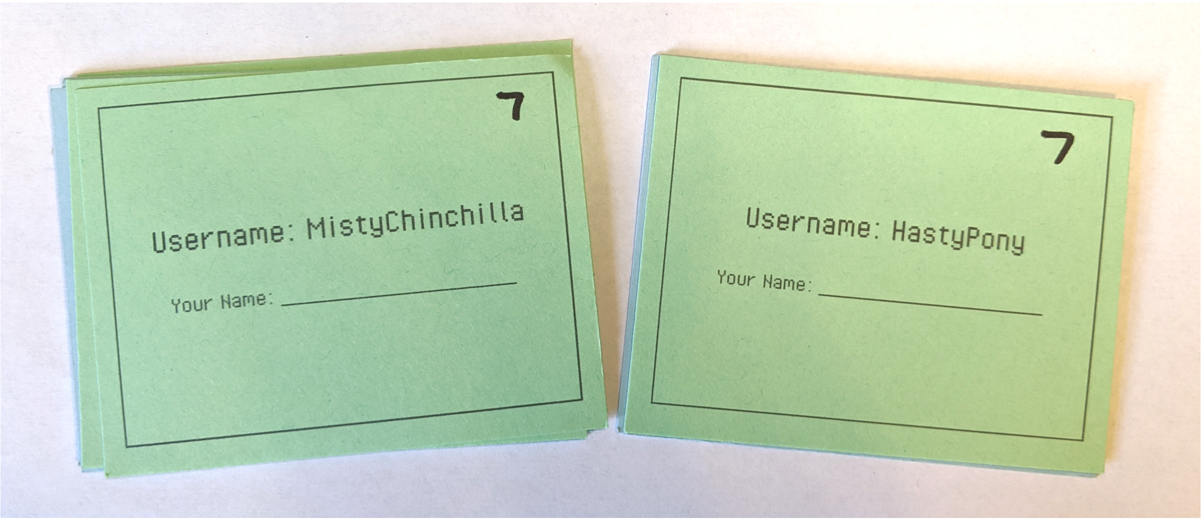


Figure 6-1: Two unused username cards from the experiment. Students were instructed to write their names below their usernames. The students' teacher distributed these cards at the beginning of each gameplay session and collected them at the end of the session. This eliminated any issues with invalid usernames while ensuring that the data was anonymous at the point of collection.

6.2.1 Post-Pilot Changes for the Full-Scale Experiment

Preliminary analysis of the data collected during the pilot prompted several changes to the game. The following changes were implemented prior to beginning the full-scale experiment.

1. The card "Facts and Figures" was removed from the game. Examination of the log data showed that that students were using this card to bypass a *Persuade* event. It may be the case that, in this respect, the card served as a bottom-out hint, but it was nevertheless removed due to the already limited number of practice opportunities and because the card provides no instruction or feedback to the student.

2. The Moral Foundations Questionnaire was updated with reading-level appropriate language. For example, the meaning of the word "chastity" was unclear to several students in the pilot.

3. Several new pieces of instructional content were added to the game in order to clarify mechanics or concepts that students in the pilot frequently asked about. This new content includes instruction pertaining to: the notebook value-selector interface, the bonus energy mechanic, the discard card option.

4. Many more opportunities for value-specific feedback were added to the game. Unlike simple correctness feedback, this feedback includes information about the specific value a student chose. Changes included feedback after a successful *Persuade* action, feedback about both values after an unsuccessful *Empathize* action, and feedback after both a successful and unsuccessful *Confirm Understanding* action.

Perhaps the single largest change between the pilot and full scale experiment was the order and content of scenarios. In the pilot experiment, new, more difficult knowledge components were introduced earlier in the sequence of scenarios. Preliminary analysis of the pilot data suggested that these new knowledge components may have been introduced before students had enough opportunities to master the ability to identify the value latent in text, a foundational knowledge component upon which the later KCs rely. In the full scale experiment, these more complex KCs were introduced after students have had multiple opportunities to practice identifying values (matching the order depicted in Figure 5-8 in Chapter 5).

The content of many scenarios was changed to increase the number of value-specific feedback opportunities. The number of scenarios remained the same (8), so the increase in feedback opportunities was achieved by 1) having the players interact with additional npcs, and/or 2) structuring the scenario in a way that encouraged (or required) players to employ actions that resulted in value-specific feedback (e.g., the *Confirm Understanding* or *Empathize* actions). This new instructional plan is depicted in Figure 5-7 in Chapter 5. This sequence is designed to give students at least 6 opportunities to practice identifying each value. More practice opportunities were given for identifying *Loyalty*, *Authority*, and *Sanctity*, as analysis of the pilot data showed these values to be the most difficult for students to identify[2].

6.2.2 Student Feedback During the Pilot

The primary goals of this pilot were to ensure that 1) the purpose of the game was clear, 2) the in-game instructions were clear, and 3) the game was leveled appropriately with respect to difficulty. Data from the pilot suggest that the game was successful on all three counts, with 84% of students reporting that they quickly grasped the overall goal of the game, 83% of students reporting that they quickly grasped how to perform in-game actions, and 79% of students agreeing that the game was not too easy and not too difficult to play. The overall reception of the game during the pilot was positive. The vast majority of students (84%) reported that they enjoyed playing the game, and that they found the game to be valuable to them (64%).

The classes used in the pilot study provided two additional sources of data, which were both unanticipated and beneficial. First, two of the three classes used in the pilot study were quite comfortable speaking freely during the gameplay sessions. This provided additional evidence for enjoyment ("I love the Social Media Snoop – it's my favorite."), collaborative help ("I think it means...go to your notebook."), and the value of humor (several students commented on the text shown when players lose a scenario).

The second unanticipated source of data were student "affirmations" – a regular end-of-week activity where each student states something that happened during the week that they are appreciative of. Several students mentioned the game specifically. For example, one student was thankful for, "3 days of not having English class." Another student stated that, "Usually computer stuff in English class is boring, but this was actually fun."

Finally, there was also some anecdotal evidence that students continued to think about the game outside of class. For instance, during a gameplay session, one student stated, "My friend just texted me: Don't let the aliens invade again." Additionally, during student affirmations, one student expressed gratitude for the opportunity to learn these civil discourse skills, adding, "It worked on my mom last night!" These out-of-class connections suggest that the game was perceived as meaningful and relevant to their lives by at least some players.

[2] There are a number of explanations for why *Care* and *Fairness* were easier to identify. One potential reason may be that the students in the pilot tended to value *Care* and *Fairness* more than the other values. Alternatively, the difficulty of each value identification task is variable, so it may be the case that the *Care* and *Fairness* tasks are, by chance, easier (irrespective of the influence of bias).

6.3 Full-Scale Experiment

Following the pilot study, a full-scale experiment was designed. In addition to the changes to the game content described above, this full-scale experiment differed from the pilot in two key ways:

1. A larger and more diverse sample of students were recruited for the full-scale experiment. The game was implemented in classes across grade levels (10th, 11th, and 12th), subject areas (Civics and English), and communities (rural and suburban).

2. In addition to evaluating in-game performance, this experiment was designed to evaluate the impact of gameplay on real-world discourse.

The full-scale experiment, originally, was to implement the following switching replications design (see Figure 6-2): First, all classes would engage in a pre-study discussion for which the classroom audio was recorded. Following the class discussion (approximately one week later), half of the classes were to play the game for three class periods, while the other half engaged in typical, business-as-usual classroom activities. Approximately a week after this first set of gameplay sessions, all classes would engage in another recorded mid-study discussion. Following this discussion, the classes that had previously not played the game would then play the game over three class periods (with the other classes resuming normal classroom activity). Finally, following this second set of gameplay sessions, all classes would engage in a final recorded class discussion.

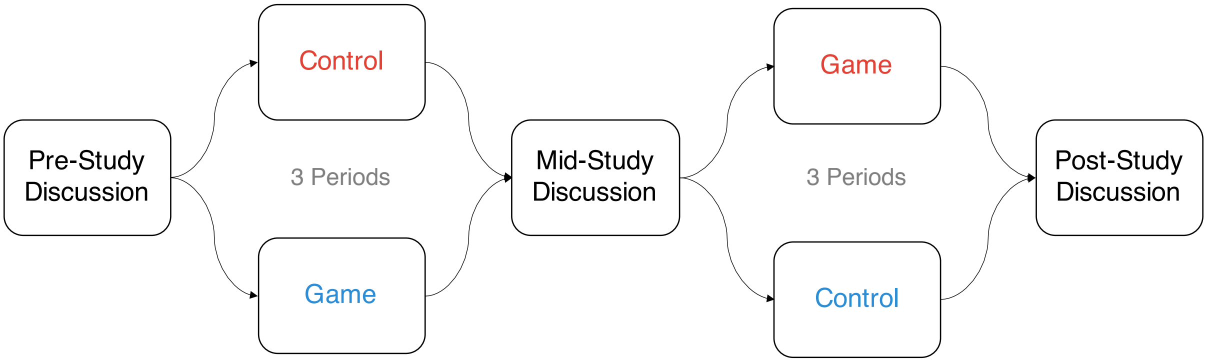


Figure 6-2: The original switching replications design that was to be implemented in the full-scale experiment. The COVID-19 pandemic prevented us from collecting data past the first set of gameplay sessions.

6.3.1 Classroom Discussions

Teachers were intentionally given very little guidance on how to conduct the recorded classroom discussions. The only instruction given was that: 1) the discussion should be no different than the class discussions that they typically engage in as part of their normal curriculum, and 2) the discussion should center on an historical or current issue that is divisive (to prompt discussion). Each discussion was to focus on a different topic, again in an effort to mimic business-as-usual classroom practices. As the game aims to teach general civil discourse skills (rather than topic-specific content), we expect the impact of the game to be present regardless of topic. Each student had the opportunity to opt their entire class out of being recorded during classroom discussions. No students chose to opt out. In addition to collecting audio data, students were asked at the end of the period to complete a brief questionnaire about various aspects of the discussion (see Appendix D).

6.3.2 Impact of the COVID-19 Pandemic

Unfortunately, the data collection phase of the experiment was cut short by the closure of local schools [28] as a result of the COVID-19 pandemic. Fortunately, data from the first, pre-study discussion as well as the first set of gameplay sessions were able to be collected before the closures. The primary consequence of this unanticipated complication is our inability to analyze the impact of gameplay on classroom discussions (as only the pre-study discussions were collected).

As a result, our analyses will focus on the data collected from those students who played the game. We will focus on the nature and extent of their learning, and the quality of their experience. Additionally, while we can no longer conduct a verbal protocol analysis that compares the productivity of discourse across classroom discussions, the data collected from the pre-study discussion may be useful as a descriptive snapshot.

6.3.3 Value-Adaptive Intervention

In addition to using our construct *Alignment* to measure the impact of bias, we also used *Alignment* to guide a value-adaptive intervention. All students were randomly assigned to one of two conditions: a control condition and an adaptive condition. The two conditions were identical in every respect with one exception: When players in the adaptive condition were asked to choose which statement would be most persuasive to an NPC, they saw one of the options presented in orange-colored text with an additional piece of instruction that read:

> **Caution: Orange-colored options might seem more persuasive to you (based on your values). Remember to choose the best response for [NPC NAME]**.

The color orange was chosen because it is attention-grabbing, however it isn't traditionally associated with correctness (as colors like red or green are). Remember that the orange colored option was not any more or less likely to be the correct answer; this intervention is simply designed to elicit a more critical analysis of the options.

6.3.4 Demographic Information

A total of 87 students from high schools located in the Northeastern Region of the United States participated in the study. Note that all demographics questions were optional, and a small number of students chose not to answer some questions. The students where evenly split with respect to sex (41 females, 43 males, 1 other), and reflective of the racial demographics of the area (9 black or African American students, 72 white or Caucasian students, and 4 students identifying as other/more than one race). Table 6.1 lists relevant features of the six classes used in the experiment, including which of the two schools the class took place at, the grade level, the subject, and the number of students.

School	Community	Grade	Subject	N
School A	Rural	10	Remedial English	10
		11	AP Lang. and Composition	36
		12	AP Literature	18
School B	Suburban	11	US History	14
		11	AP US History	9

Table 6.1: Features of the six classes used in the experiment. Note that there were two classes of 11th grade AP Language and Composition.

6.4 Data Analyses

6.4.1 Detecting Gaming Behaviors

While we hope the majority of students will make an honest effort to engage with the educational technology we developed, a subset of students will undoubtedly try to game the system – that is, try to fulfill their participation requirement without engaging with the material in a conscientious way. This lack of engagement is noteworthy in and of itself, perhaps speaking to a failure to engage a particular subpopulation. However, apathetic responses are also a significant source of noise when attempting to measure the impact of the game on those students who made an effort to learn. For this reason, we specify several exclusion criteria designed to capture so-called gaming behavior.

Reading Traps

We used two questions built into the pre- and post-tests to classify gaming behavior. Embedded into the Moral Foundations Theory Questionnaire on the pre-test, students are asked to indicate the degree to which they agree with the statement, “It is better to do good than to do bad.” We would expect that everyone would indicate that they agree with this statement. However, students that are responding to the prompts without reading them may accidentally indicate that they disagree. Additionally, embedded into the post-test is the question:

> While completing the study, it’s important that you read each prompt. To demonstrate that you’re reading each question, please select the third choice.

This question is designed to catch users who may be answering post-test questions without reading. Note that it is possible that some subset of students pass the reading trap on the pre-test, but fail the trap on the post-test. These students may have experienced fatigue by the conclusion of the experiment. To mitigate the potential impact of fatigue, we consider only students who fail both reading traps (i.e., students who demonstrate potential gaming behavior at both the start and end of the experiment).

Duration-Based Classification

In addition to using reading traps to classify potential gaming behavior, we also used task duration-based classification. Specifically, we examined amount of time a student took to read and provide an answer to a persuade task (i.e., “Which statement would be most persuasive?”).

Importantly, we only examined the duration of the student's first attempt at persuading each NPC. This helps us make a distinction between probable gaming behavior (which is uninformative) and guessing behavior (which may be informative). Consider the following scenario:

Imagine a student who is engaged in conscientious gameplay. On the first attempt, the student reads the options carefully and reasons about each. The student quickly eliminates one of the three options, but is unsure which of the two remaining options is correct, as a reasonable argument could be made for either. After some internal deliberation, the student makes a choice, but is immediately notified that it was the wrong one. The feedback offers a clue as to why the chosen answer was incorrect and why the alternative choice is correct. With this information, the next time the student has the opportunity to perform this same persuade action, they do not need to carefully read and consider the options. Instead they quickly choose the correct option.

Because the options given to students during persuade action actions are fixed (though presented in random order), if we label every rapid attempt as a gaming behavior we risk mistakenly classifying interactions like the one described above as gaming. Contrast the above scenario with a student whose first attempt has an extremely short duration. If the student makes their first attempt before they could have finished reading and reasoning about the available options, we can reasonably assume that the student is engaging in gaming behavior.

Still, the parameters of this definition leave room for ambiguity. For example, what threshold of duration distinguishes gaming behavior from conscientious engagement? Typically, one might generate this threshold by multiplying the average number of words we expect the participant to read per opportunity (e.g., 31.5 words) with known average reading speeds (250 words per minute [3]) to compute an estimation of how long we might expect a participant to read every word of the task (7.5 seconds).

However, there are reasonable explanations (other than gaming behavior) that could explain how a student could provide good-faith answer in a duration shorter than 7.5 seconds. For example, a student may read the first option and be confident enough that it is the correct answer to choose it without reading the other two options. Alternatively, the student may quickly skim the options for keywords related to the value they are trying to match (e.g., the word "equal" if they are looking to match to the *Fairness* foundation). While this strategy may not always work, it is certainly ecologically valid.

Because we do not want to misclassify these valid strategies as gaming behaviors, we adopt a more lenient measure of gaming with respect to duration. First, we count the number of times a student's first attempt at a unique task lasts less than 3 seconds (approximately the time it would take to read one of the three options). Then we divide that number by the total number of unique tasks the student attempted (as to not penalize students who completed more scenarios). Using this ratio (the proportion of first attempts made in under 3 seconds) allows us to use a generous duration threshold with the assumption that students engaged in gaming behavior will continue to regularly employ this strategy throughout the game.

A histogram of the frequencies of this ratio (hereafter referred to as the gaming ratio) suggested an appropriate ratio threshold between .2 and .5. We decided to adopt a ratio threshold of .4 for two reasons. First, theoretically, it is reasonable to assume that students who intend to game-the-system will engage in gaming behavior at least 40% of the time. Second, this threshold allowed us to generate an improved model without sacrificing a significant portion of the data (as was the case with more conservative thresholds).

[3] generally researchers use 300 words per minute, but because we are estimating the reading speed of high school students, we use the more conservative estimate of 250 words per minute.

6.4.2 Measuring Learning Outcomes

We used a student's performance on a series of assessment events to estimate learning. Some of these assessment events (e.g., *Persuade* actions and *Empathize* actions) mimic traditional assessment methods (e.g., multiple choice questions). However, other events (e.g., identifying the value in a new piece of intel) happen in unseen or so-called "stealth" [71] assessments.

In both cases, we estimated learning primarily by modeling the relationship between practice opportunities and performance. If students perform better with practice, that is evidence that learning has occurred. Specifically, we use the following hierarchical mixed-effects model to measure the impact of prior opportunities on performance:

$$Outcome \sim priorOpportunities + (1|AP/Student) \tag{6.1}$$

With *Outcome* representing a binary correctness score (0 = incorrect, 1 = correct); *priorOpportunities* representing the number of times a student has already attempted a problem of this kind[4]; and $(1|AP/Student)$ representing a nested random effect with the student's unique identifier nested under AP, a binary variable indicating whether or not the class the student played the game in was an AP course[5]. This model served as the base, onto which more complex models are built to answer more nuanced questions. What follows is a description of our four primary hypotheses, with respect to our primary learning objectives, and a description of the model used to test each hypothesis.

Hypothesis 1 *Performance on value-identification tasks (in general) will improve with practice.*

To model the impact of performance on value-identification tasks in general, we select from the log data the outcomes for each *Persuade*, *Confirm Understanding*, and *Empathize* action, the three player actions that both require value-identification and provide correctness feedback. Each one of these opportunities is counted incrementally, grouped by the value being identified (e.g., attempt 1 at identifying the *Fairness* foundation). Note that because empathize actions require the player to identify two values, these actions are counted twice, once for each value identified. Tutorial scenarios are excluded from analysis.

Because some values may be more easily identified than others, we add an additional random effect, *valueType*, which fits a random intercept to each of the five values the student is being asked to identify. The resulting model is of the following form:

$$Outcome \sim priorOpportunities + (1|AP/Student) + (1|valueType) \tag{6.2}$$

Hypothesis 2 *Performance on value-identification tasks that require students to identify shared values across ideological divides will improve with practice.*

In addition to value-identification in general, we investigated whether or not students' improved with respect to identifying values across ideological divides. In the game, students

[4] Because student gameplay was limited to three class periods, many students began problems that did not get a chance to complete. These incomplete problems artificially inflate the error rates of later scenarios. To mitigate this source of noise we excluded problems that were started but never completed. This effectively steps a student's progress back one scenario (to the last scenario they completed).

[5] Note that there are several other factors that could have served as grouping variables (e.g., teacher, grade), but our analysis showed that these factors explained very little variance. In contrast, whether or not the class was an AP class explained some variance, though not nearly as much as the student's unique identifier.

practice this skill when performing the *Persuade* action. Modeling the impact of practice on these tasks was modeled using a similar method as described above, with one exception: only outcomes for *Persuade* actions were selected (other outcomes were excluded from analysis).

Hypothesis 3 *Students will use tribalism reducing discourse moves more efficiently with practice.*

In addition to identifying the values latent in text, students also have the opportunity to practice tribalism reduction. Throughout the game, students are faced with NPCs that are resistant to persuasion due to their high tribalism. Recall that a successful persuade action depends on 1) the student choosing the correct action, and 2) that a random number drawn between 1 and 100 is higher than $(100 - NPCTribalism)$. In order to maximize the chances of a successful persuade action, students need to lower the tribalism of the NPC they are trying to persuade.

To measure student performance on tribalism reduction over time, we first constructed a state model that stepped through the log data, recreating the game state for each student during each *Persuade* event. This state model provides one key piece of information: the current tribalism score of the NPC being persuaded. We can compare the NPC's current tribalism score with their default score (i.e., the score the NPC is initialized with at the beginning of the scenario) to compute a tribalism reduction ratio. It is important to compute a ratio in this case, as later scenarios are associated with higher tribalism scores. We place this ratio into the following model:

$$\mathit{tribalismRatio} \sim \mathit{priorOpportunities} + (1|AP/Student) \tag{6.3}$$

Where *tribalismRatio* is the outcome variable, predicted by the fixed effect *priorOpportunities* and the nested random effects of *AP status* and the *Student Identifier*. If students become more efficient at tribalism reduction, we expect that priorOpportunities will have a negative coefficient in this model. That is, as practice opportunities increase, the tribalism ratio decreases.

Hypothesis 4 *Performance on tasks that require students to "Choose the best discourse move" will improve with practice.*

Finally, there are a number of other civil discourse skills and concepts that play an important, but secondary role in gameplay. These skills are reinforced throughout the game, but explicitly assessed at the beginning of turns 2 and 3 of each scenario, when players are presented with a random, multiple choice question. In the game, a correct answer to this question serves to prove the worthiness of the player to the mayor of the town, who in turn makes a certain type of discourse move free-of-cost for a turn. As resources are limited, it is highly advantageous for players to try to win these free cards.

To measure performance on these cards, we first constructed a series of simple knowledge component (KC) models: 1) a single skill KC model, in which we presume all questions capture performance on a single skill; 2) a unique-step KC in which we presume each question captures performance on its own, unique skill; 3) and a theory-driven model, in which questions that were designed to capture the same skills are grouped together. We then use a model, similar to our base model, to estimate the performance on these KC models:

$$\mathit{Outcome} \sim \mathit{priorOpportunities} + (1|AP/Student) \tag{6.4}$$

Where *Outcome* represents the binary correctness score (0 = incorrect, 1 = correct) for the multiple choice problem, and *priorOpportunities* represents the number of times, prior to the current opportunity, that the player has seen a question that targets the same KC. As always, the nested random effects of *AP* status and *Student* identifier are included as well.

6.4.3 Measuring the Impact of Bias

Like most educational technologies, our game was primarily instrumented to capture changes in conceptual understanding and skill acquisition. That is, it was designed to measure performance along the dimension of learning. However, our game was also instrumented to capture performance along a secondary dimension: bias. We expect that these two dimensions are not independent of one another. This subsection details our various methods for testing whether or not our game can be used to measure the impact of bias on performance. The next subsection details how we tested our debiasing intervention.

We constructed one general model to test the impact of our bias estimates on in-game performance. We generate the model twice, in order to compare a baseline estimate of potential bias with our data-driven estimate of potential bias (which we refer to as *Alignment*). The baseline and data-driven estimates are described in detail below.

Questionnaire Scores as a Predictor of Bias

As a baseline measure, we used a student's raw scores on the Moral Foundations Theory Questionnaire as an estimate of their potential bias. That is, each value-identification task requires the student to identify a specific value. We use the student's questionnaire score for that specific value as a fixed effect in the model. If this score is a good estimate of potential bias, we expect a relationship between the student's questionnaire score (for the value in question) and the student's performance when asked to identify that value.

Alignment as a Predictor of Bias

We compared the above baseline measure to our data-driven measure, *Alignment* (see Section 4.3 for additional details). In this case, the values latent in the text for the correct answer in a value-identification task are compared to the student's values (as measured by the Moral Foundations Theory Questionnaire) to compute *Alignment*. Because *Alignment* captures the relationship between user and content across five dimensions (as opposed to just one in the method described above), we expect this estimate of potential bias to be more nuanced, and thus, a better predictor of performance.

Differing Relationships Between Bias and Performance

In our previous work, the impact of *Alignment* on outcomes was always negative. But recall that in these tasks we were asking participants to evaluate the strength of arguments, we were not evaluating their ability to identify the values latent in them. In our game, the relationship between *Alignment* and performance on value-identification tasks is less clear. We hypothesize that students will have an easier time identifying values that they share. However, it is possible that the process of identifying a value that you don't share prompts System 2 processes that improve student accuracy.

Limitations of Bias Estimation

Before proceeding, it is worth noting that we are careful to specify that these are estimates of *potential* bias. A more complete model would include some estimate of one's propensity to overcome one's biases as well (a feature not captured in these models). Unfortunately, such a measure is difficult to estimate, and traditional measures (e.g., analytic thinking tasks [26]) are likely not informative with our population of interest.

Bias Model

We use these two estimates of potential bias (baseline and *Alignment*) in two models (problem-level and decision-level). The first, problem-level, model follows our base model:

$$Outcome \sim Estimate + (1|AP/Student) \tag{6.5}$$

Where *Outcome* represents a binary measure of correctness for each value-identification opportunity, and estimate represents either the baseline estimate of potential bias or *Alignment*. We expect that both the baseline estimate and *Alignment* will result in a significant, positive coefficient. That is, students will have an easier time identifying values they share. Due to the additional nuance captured by *Alignment*, we expect that *Alignment* will outperform the baseline measure in its predictive power. These expectations can be restated as the following two hypotheses:

Hypothesis 5 *Both the baseline estimate of potential bias and Alignment will be predictive of performance on value-identification tasks.*

Hypothesis 6 *The predictive power of Alignment will outperform the baseline measure for predicting performance on value-identification tasks.*

6.4.4 Intervention

The final major area of analysis we explored was the impact of a value-adaptive debiasing intervention. It is worth reiterating that merely measuring bias in a scalable and data-driven way is an important and novel innovation, and the primary contribution of this work. Still, not designing and testing a debiasing intervention that leverages this innovation would be a missed opportunity.

We evaluated the impact of our intervention on numerous measures of student performance and engagement (e.g., performance on value-identification tasks, enjoyment, problem duration). Because students were randomly assigned to one of two conditions (Control or Adaptive), the majority of these analyses use independent samples t-tests to test for differences between conditions. However, in cases that require additional model complexity, we implemented a variant of our base hierarchical mixed effects model, with an additional fixed effect of *Condition*.

6.5 Conclusion

In this chapter, we described the experimental materials used to evaluate the effectiveness of our game. We also described the successes and shortcomings of a small-scale pilot, and the changes made before deploying the game in a larger, full-scale experiment. Finally, we described the methods we used to classify gaming behavior, measure learning, and estimate bias. In the next chapter, we will detail the results of that experiment, both with respect to learning and player experience.

Chapter 7

Results

"The concept is pretty neat and I liked the graphics.
It was a useful tool for class and an engaging way to practice"
- One student

"I liked that I didn't have to write an essay."
- Another student

In this chapter we will detail the results of an experiment designed to test the impact of our educational game. These results are organized into four sections:

1. **Student Experience:** Was the game meaningful and enjoyable to students?
2. **Learning:** Did the game improve performance on key skills?
3. **Impact of Bias:** Was our measure of *Alignment* predictive of bias?
4. **Debiasing Invervention:** Was our value-adaptive intervention effective at reducing the impact of bias?

A total of 87 students were included in analyses (after 3 students were excluded for exhibiting behavior that met our criteria for gaming).

7.1 Student Experience

At the conclusion of the study, all students were asked to complete a post-test questionnaire that included a series of questions inspired by the Player Experience Inventory [84]. Students responded to each question using a Likert scale ranging from 0 ("Strongly Disagree") to 5 ("Strongly Agree"). Figure 7-1 shows student responses to each of the 13 player experience questions. Responses are positioned around zero by positioning "disagree" responses to the left of 0.

7.1.1 Enjoyment

While many of these player experience measures have implications for future iterations of this work, the following analyses focus on only one measure: enjoyment (i.e., "I enjoyed playing this game"). We use enjoyment as a proxy for player experience in general for the purposes of investigating relationships between player experience and the following variables of interest: grade level, video game usage, in-game progress, and in-game performance.

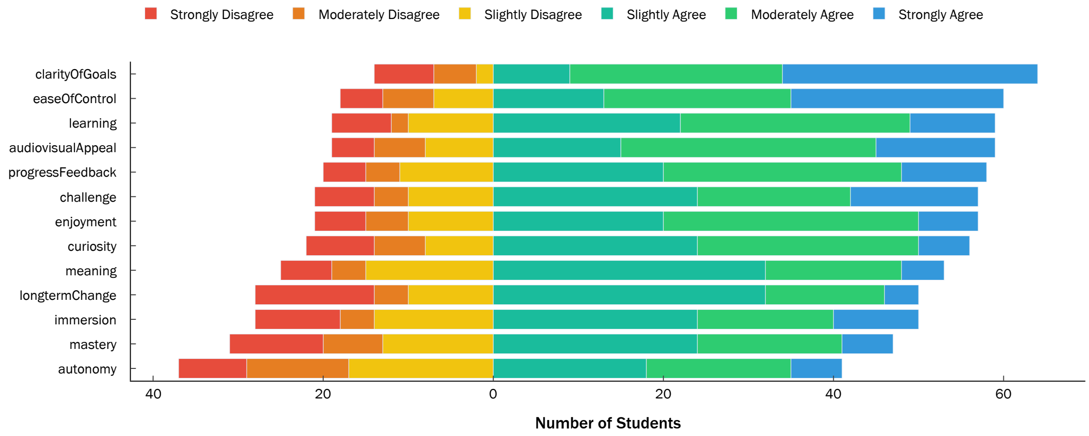


Figure 7-1: Student responses to player experience questions positioned around zero such that negative responses are positioned to the left of 0. A majority of students responded positively on all measures, suggesting that most students had a positive overall experience with the game.

First, a one-way ANOVA showed significant differences in enjoyment between grade levels (F=19.39, p<.001). We performed a multiple pairwise comparison between groups using Tukey's HSD test, and found that all three grades differed significantly with respect to enjoyment, with 12th grade students (M=4.11, SD=0.76) enjoying the game significantly more than 11th grade students (M=3.08, SD=1.26), who enjoyed the game significantly more than 10th grade students[1] (M=1.20, SD=1.32) (see Figure 7-2).

Video game usage (i.e., "I play video games: [regularly, sometimes, rarely, or almost never]") was also significantly correlated with enjoyment (r=.27, p<.05), however this was only true when the 10th grade Remedial English class was excluded from analysis (due to the class' unusual combination of high video game use and low enjoyment).

Our data also suggest that high ratings of enjoyment were associated with student-level motivation and performance metrics. For example, there is a moderate positive correlation between the number of scenarios completed and ratings of enjoyment (r=.48, p<.001) with students who had completed more scenarios reporting higher levels of enjoyment (though we are careful to clarify that the directionality of this relationship cannot be determined). The number of scenarios a student completes is likely a measure of both motivation as well as performance. This is because completing a high number of scenarios is almost certainly attributable to a mastery of the material rather than chance.

To obtain a more direct measure of the relationship between performance and enjoyment, we compared enjoyment ratings to in-game performance on one of the key assessment tasks: identifying which value is latent in some text. We found a moderate positive correlation between average rate of success and enjoyment (r=.47, p<.001), such that higher enjoyment ratings were associated with higher average rates of success. Figure 7-3 shows the relationship between enjoyment and average rate of success.

[1] This study was designed to collect data from an additional, non-Remedial English class. However, the COVID-19 pandemic resulted in school closures before the data could be collected. While the data gathered from this remedial 10th grade class reflects the experience of a valid and important cohort of students, we recognize that any effects related to grade may be skewed by what may be an unrepresentative sample of 10th graders as a whole.

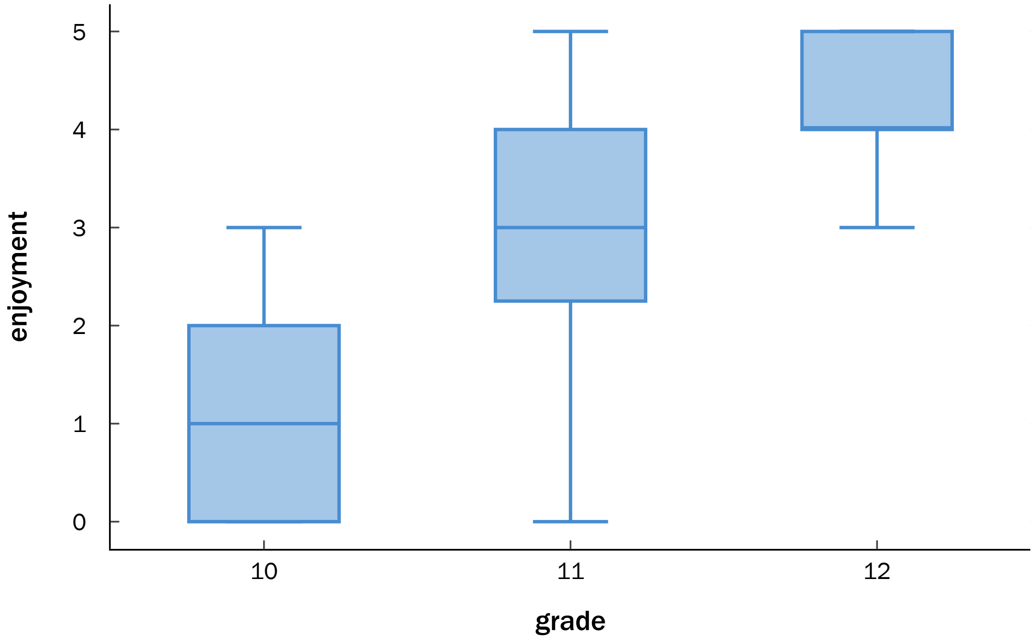


Figure 7-2: Levels of enjoyment across grades. Each of the three grades' scores differed significantly from the others, with each grade level enjoying the game significantly more than their peers in the grade below them.

To control for known grade-level effects, we generated the following logistic hierarchical mixed effects model:

$$Outcome \sim Enjoyment + (1|Teacher/Grade/Student) \tag{7.1}$$

The data for this model represent every opportunity each student had to perform the value-identification task. *Outcome* is a binary variable representing whether the student correctly identified the value (1) or incorrectly identified the value (0). *Enjoyment* is the student's enjoyment score. A nested random effect of teacher, grade, and student was also included, again, to control for known grade-level effects. We found that *Enjoyment* was a significant predictor of *Outcome* ($\beta = .19$, p<.001), even when controlling for grade-level.

7.1.2 Out-of-Class Gameplay

During data collection, researchers began noticing unusual "jumps" in student progress between in-class gameplay sessions. Further analysis of user-session timestamps revealed that some students were indeed playing the game either during other free periods throughout the school day or outside of school hours. Approximately 11% of logged transactions occurred outside of a student's class period (including a 10 minute buffer on each side of the interval), with approximately 36% of those transactions occurring outside of school hours entirely. Approximately 28%

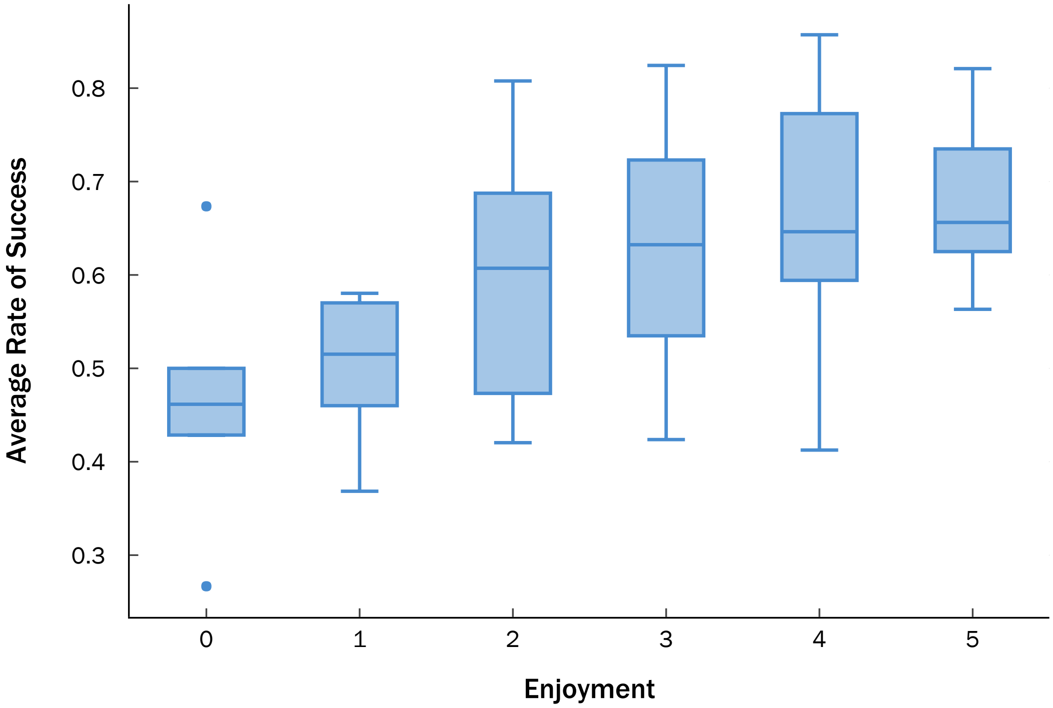


Figure 7-3: Average rate of success for students grouped by enjoyment rating. There was a moderate positive correlation between average rate of success and enjoyment (r=.47, p<.001), such that higher enjoyment ratings were associated with higher average rates of success.

of students had at least one out-of-class transaction, with 15% of students logging at least one action outside of school hours.

While it is possible that out-of-class gameplay was motivated by a desire to be done with the game, there are several problems with this explanation. First, after one student in the classroom had completed the game (generally at the end of the second or beginning of the third gameplay session), students were informed that students who complete the game were to continue working on an assignment from their typical class curriculum. Second, while students were encouraged by their instructors to finish the game, it was made clear to students (generally during the second or third session) that completing the game within the three sessions was not required. Given this information, students that engaged in outside-of-class activity would have to enjoy playing the game less than their typical class activities. This is not borne out the data, which show that the students who played the game outside of class (M=3.54, SD=1.00) enjoyed the game significantly more than students who did not (M=2.86, SD=1.54) (t(76)=2.10, p<.05).

7.1.3 Qualitative Feedback

Students were also provided with an opportunity to provide open-ended feedback about the game. During the post-test, students were asked to describe what they liked and disliked about the game. Eighty-five students generated 170 comments, which were printed and arranged around common themes via affinity diagramming. Note that entries that contained more than one sentiment were split into two or more entries. The results of the diagramming process revealed that student feedback generally related to three common themes: 1) player experience and attitudes, 2) the level of difficulty, and 3) the impact of the game on student knowledge and dispositions. Below we explore student comments related to each of these themes in detail. Figure 7-4 shows the ultimate affinity diagram used to organize student comments, and Table 7.1 lists some of the major categories from the process alongside examples of each.

Player Experience and Attitudes

Many students used this opportunity to provide more descriptive feedback about their experience playing the game or their attitudes about game-based learning. For example, several students said that they enjoyed the artistic style (n=6), game content (n=4), and general idea (n=5). While some students (n=5) spoke about how they enjoyed the autonomy given to players, a large number of students (n=13) expressed unhappiness with the repetitive nature of the game. A somewhat surprising result was the number of students (n=7) that specifically mentioned that they enjoyed the process of persuading townspeople, the core mechanic of the game.

Five students simply enjoyed the fact that the instruction was delivered via a game and "not just paper and pencils," with three students explicitly stating that they preferred our game to their normal classroom activities. In perhaps the game's most glowing review, one student said that they "lost track of time and played it in multiple other classes because [they] wanted to beat the game."

Level of Difficulty

Much of the feedback focused on the game's level of difficulty, with the responses suggesting a varied experience in this regard. Some students (n=8) found the game to be too difficult, while others (n=7) expressed the opposite sentiment, saying that the game was too easy or simple. Four students admitted that they had some difficulty getting a grasp on the actions at the beginning of the game. Similarly, four students said that the difficulty level of the game was balanced, increasing as they progressed through the game. Three students said that they explicitly enjoyed that the game was challenging. Several (n=7) students made comments related to the difficulty of identifying values, suggesting that this is an area that may require more or different instruction.

Impact on Knowledge and Dispositions

Finally, many students took this opportunity to reflect on what they had learned (n=5), how those lessons relate to their lives (n=4), and the game's value as an educational tool (n=9). Students highlighted the authenticity of the in-game interactions ("I liked that it kind of showed how to act in real-life situations") and the meaningfulness of the gameplay ("It had me thinking about my actions/conversations with others"). Perhaps most importantly, this qualitative feedback provided additional evidence that the game successfully conveyed a complex notion fundamental to productive civil discourse. As one student perfectly articulated, "This game showed the importance of understanding other people's values when discussing a controversial matter."

Affinity Category	**Example Entry**
How did students feel about the design?	“I liked the style”
Did students like the content?	“I liked seeing all the different people’s values”
Did students like the concept?	“it was a good concept for a game”
Did students have autonomy?	“I enjoyed how interactive the game was, allowing the player to think for themselves and determine the values of the characters”
Was there enough variety?	“it was too repetitive”
Did students enjoy persuading NPCs?	“I liked how we had to match a person’s value to get them to look at our side of an argument”
How did students feel about game-based learning?	“It being a game appeals to me more”
Did students prefer this to normal classroom activities?	“better than English class amirite?”
Was the game too difficult?	“it was hard and confusing”
Was the game too easy?	“it was very easy to figure out”
Did the game get easier with practice?	“it was confusing at first but you get the hang of it”
Were students continually challenged throughout the game?	“I liked that it got a little harder as it went on...”
Was value-identification instruction clear?	“some ideas I chose that made sense weren’t correct and I wasn’t sure how”
Did students learn what we wanted them to learn?	“helped develop skills regarding understanding opposing arguments.”
Did students think this was a valuable educational tool?	“I like how the game made you think.”
Did students relate the game to real life?	“I liked that it kind of showed how to act in real-life situations.”

Table 7.1: Some of the themes that emerged during the affinity diagramming of student’s open-ended feedback, as well as an example of student feedback exemplifying each theme.

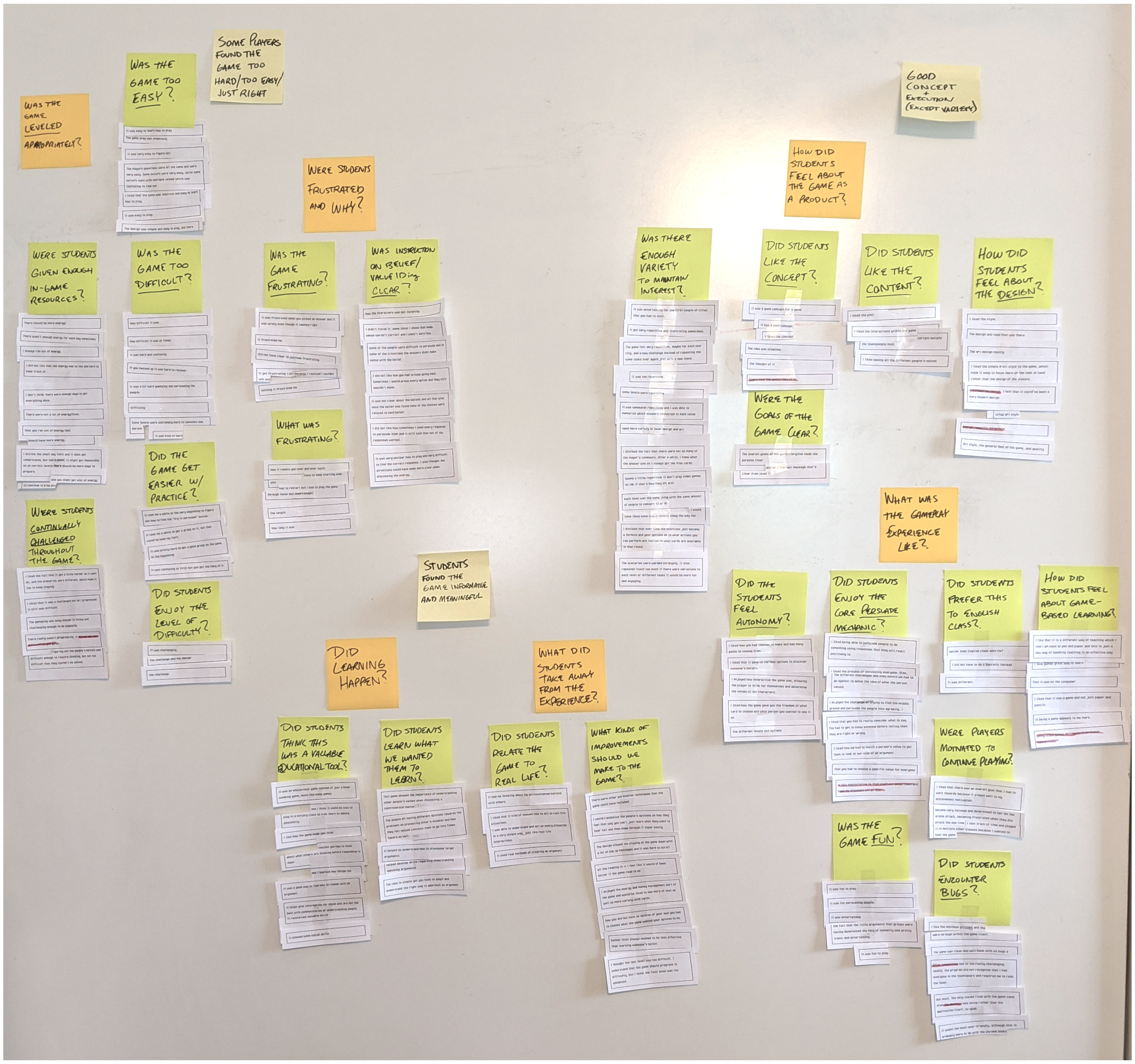

Was the game leveled appropriately?
Was the game too easy?
Some players found the game too hard/too easy/just right
Were students frustrated and why?
Good concept + execution (except variety)
How did students feel about the game as a product?
Were students given enough in-game resources?
Was the game too difficult?
Was the game frustrating?
Was instruction on belief/value IDing clear?
Was there enough variety to maintain interest?
Did students like the concept?
Did students like the content?
How did students feel about the design?
What was frustrating?
Were the goals of the game clear?
Did the game get easier w/ practice?
Were students continually challenged throughout the game?
Did students enjoy the level of difficulty?
Students found the game informative and meaningful
What was the gameplay experience like?
Did the students feel autonomy?
Did students enjoy the core persuade mechanic?
Did students prefer this to English class?
How did students feel about game-based learning?
Did learning happen?
What did students take away from the experience?
Did students think this was a valuable educational tool?
Did students learn what we wanted them to learn?
Did students relate the game to real life?
What kinds of improvements should we make to the game?
Were players motivated to continue playing?
Was the game fun?
Did students encounter bugs?

Figure 7-4: An affinity diagram of student open-ended responses to two questions on the post-test ("What did you like about the game?" and "What did you dislike about the game?"). Three major themes emerged: 1) more detailed comments about player experiences and attitudes, 2) comments on the level of difficulty, and 3) reflections on the practical and educational value of the game.

Coefficients	**Estimate**	**SE**	**t**	**p**
Pre Test Score	0.216	0.109	1.972	< 0.1
Num Scenarios Completed	0.293	0.109	2.677	< 0.01
(intercept)	0.000	0.102	0.000	1.000

Table 7.2: Summary of the linear regression model used to examine the impact of in-game performance on post-test scores, while controlling for the student's pre-test score. All variables are standardized to allow direct comparison. The number of scenarios a student completed was a significant predictor of post-test score.

7.2 Learning

7.2.1 Pre-Post Gains

Students were asked to answer a small set of identical questions before and after gameplay. These questions were designed to measure their mastery of some of the key concepts and skills that they practice throughout the game. A t-test revealed no significant difference between pre- and post-test scores for all students, though students in grade 12 saw a marginally significant gain ($t = -1.94, p < 0.1$) from pre-test ($M = 2.22, SD = 0.81$) to post-test ($M = 2.89, SD = 1.18$).

We used the number of scenarios a student completed as a rough proxy for in-game performance for the purposes of examining the impact of in-game performance on post-test scores. We generated the following simple linear regression model, controlling for the impact of the student's pre-test score:

$$postTestScore \sim preTestScore + numScenariosCompleted \tag{7.2}$$

Table 7.2 shows the results of this model. We found that the number of scenarios a student completed was a significant positive predictor ($\beta = 0.29, p < 0.01$) of post-test scores, even when controlling for a student's pre-test scores (which were also a marginally significant predictor of post-test scores ($\beta = 0.21, p < 0.1$)).

7.2.2 Value-Identification

We modeled the impact of practice on a student's ability to correctly identify the value latent in text using the following hierarchical mixed effects model:

$$Outcome \sim priorOpportunities + (1|AP/Student) + (1|valueType) \tag{7.3}$$

Recall that we include a random effect for value type, as we expect that students learn to identify each value individually (as opposed to a general value-identification skills that is employed regardless of the value being identified). An alternative to including this additional random effect would be to generate a separate model for each of the five foundations. However, opting for the single model provides us with more data, and allows the model to leverage student-level variance (which we expect to be present across foundations).

Table 7.3 shows the results of this model. We found that, as predicted, the number of prior practice opportunities was a significant positive predictor of performance on value-identification tasks. That is, more practice was associated with lower error rates. We also found that both the

Random Effects	**Variance**	**SD**
Student:AP	0.275	0.524
Foundation	0.219	0.468
AP	0.006	0.082

Fixed Effects	**Estimate**	**SE**	**df**	**z**	**p**
Prior Opportunities	0.153	0.034	3278	4.434	< 0.001
(intercept)	0.703	0.236	78	2.974	< 0.05

Table 7.3: Summary of the hierarchical mixed effects model used to measure the impact of practice on value-identification performance. *Prior Opportunities* (i.e., practice) was a positive predictor of *Performance*, the outcome variable. That is, more practice was associated with better performance on value-identification tasks, as predicted.

nested random effect capturing *AP* status and *Student* ID, as well as the random effect capturing the type of *Foundation* being identified accounted for some variance. Mixed effects models generated without these random effects resulted in higher AIC scores (AIC=6780 and AIC=6774 respectively) compared to that of this model (AIC=6626), suggesting that the inclusion of these random effects improved the model.

7.2.3 Shared Value-Identification

We hypothesized that one's ability to identify values in text across ideological lines may be a distinct skill. That is, we that hypothesize that each *Persuade* action relies on two related, but distinct skills:

1. Identifying which one of the five foundations the NPC in question values
2. Selecting an argument from the opposing side that matches that value

To accomplish the first step, players can play cards to gather intel (clues) about what the NPC values. We call the second step *shared value-identification*, as it requires players to understand how that value might be represented in a way that supports the opposing side. For example, if a player discovers that an NPC believes that, "Football is too dangerous" and guesses that the NPC values the *Care* foundation, they might attempt to persuade the NPC with the argument, "Football builds character and does more good than harm" – an argument supporting the opposing side that also appeals to the *Care* foundation.

To measure performance on only the second step, we first isolated *Persuade* opportunities where the student has already correctly guessed what foundation the NPC values at the time of the persuade event. Guesses are logged using a value-selector interface, which is a reference tool that allows players to toggle (on/off) each of the five foundations with respect to each NPC. Players can toggle an NPC's potential values at any time by accessing the NPC's entry in the player's notebook, but players are also given the opportunity to toggle potential values after collecting intel about an NPC. It is important to note that the player may or may not be aware of the fact that they have identified the correct value. For example, toggling a value after gathering intel does not provide correctness feedback. However, using discourse moves like *Confirm Understanding* and *Empathize* can confirm or disconfirm a player's guess.

Because we selected only persuade opportunities in which the player has already correctly guessed what the NPC values, the performance on the persuade action directly reflects their ability to perform the second step. We tested the impact of practice on shared value-identification using the same model described above (Equation 7.4), but using this subset of persuade opportunities. Contrary to our hypothesis, we found that the the number of prior opportunities was not a significant predictor of performance on shared value-identification.

Correct Guess x Opportunity Interaction

While this result was surprising, one possibility is that the opportunities that fall into this subset are impacted by a ceiling effect. That is, the students who are correctly identifying the NPC's foundation also excel at identifying shared values. To examine the relationship between correctly guessed NPC foundations and shared value-identification, we selected all *Persuade* action opportunities and constructed the following model:

$$Outcome \sim priorOpportunities * guessedCorrectly + (1|AP/Student) + (1|valueType) \quad (7.4)$$

This model is similar to the previous model described above, with a few key differences. First, the data for this model are not limited to *Persuade* opportunities in which the student has correctly guessed the NPC's foundation; all *Persuade* opportunities are included. Additionally, we include *priorOpportunities* in an interaction term with *guessedCorrectly*, a fixed effect represented whether or not the student guessed the NPC's foundation correctly at the time of the *Persuade* event. Results from this model are reported in Table 7.4. We found evidence for an interaction between *guessedCorrectly* and *priorOpportunities* such that the impact of practice is significantly different for opportunities in which the student has correctly guessed the NPC's foundation. The nature of this interaction is visualized in Figure 7-5. Graphing the data suggests that, while there is no relationship between practice and performance for opportunities with correct guesses, there is a positive relationship between performance and practice for opportunities without correct guesses.

A third model was generated to examine the impact of practice on opportunities without correct guesses. Results from this model support our interpretation, as practice is indeed a marginally significant positive predictor of performance on *Persuade* opportunities without correct guesses ($\beta = 0.109, p < 0.1$).

7.2.4 Tribalism Reduction

Throughout the game, students are faced with NPCs that are resistant to persuasion due to high amounts of political tribalism. Reducing an NPC's tribalism increases the chances of a persuade action being successful. We hypothesized that students will use these tribalism-reducing discourse moves more efficiently with practice.

To measure the impact of practice on tribalism reduction, we first constructed a state model to capture the tribalism score of each NPC at each persuasion event for each student. We compared the NPC's current tribalism score with their default tribalism score to compute a ratio representing the percentage of the NPC's default tribalism that has yet to be reduced. We refer to this simply as the NPC's *tribalism ratio*. It is important to use a ratio rather than a

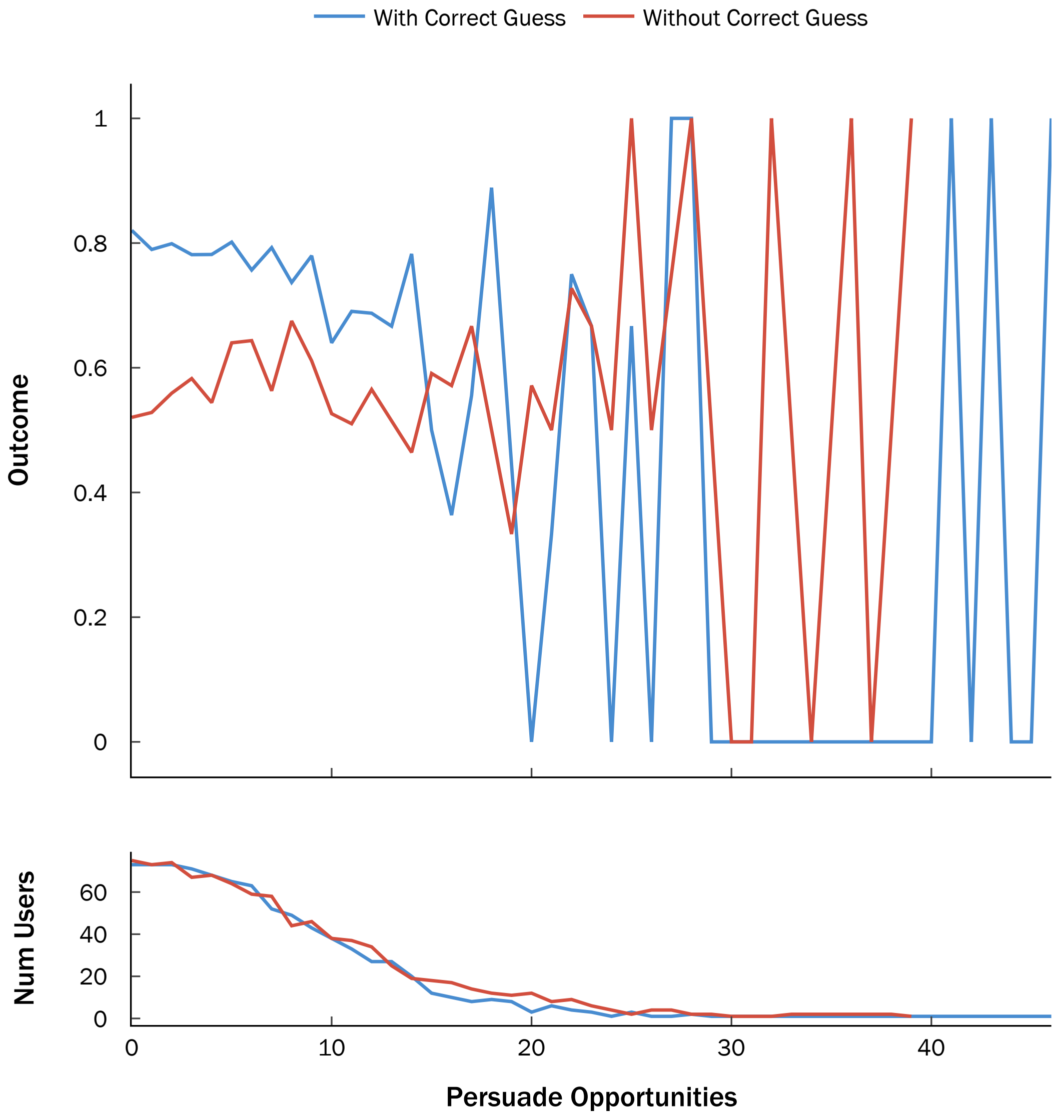


Figure 7-5: The top figure shows the interaction between mean performance at each opportunity and *guessedCorrectly*. There appears to be no relationship between opportunity and performance for attempts made with a correct guess. However, there may be a relationship between opportunity and performance for attempts made without a correct guess.

Random Effects	Variance	SD
Student:AP	0.197	0.444
Foundation	0.185	0.430
AP	0.013	0.115

Fixed Effects	Estimate	SE	z	p
Prior Opportunities*Guessed Correctly	-0.278	0.079	-3.497	< 0.001
Prior Opportunities	0.177	0.055	3.219	< 0.01
Guessed Correctly	0.813	0.085	9.516	< 0.001
(intercept)	0.490	0.231	2.117	< 0.05

Table 7.4: Summary of the hierarchical mixed effects model used to explore the relationship between correct guesses and shared value-identification across practice opportunities. We found evidence for an interaction between *guessedCorrectly* and *priorOpportunities*, suggesting that the impact of practice on *Persuade* action performance is different for opportunities in which the student has correctly guessed the NPC's foundation.

raw score, because later scenarios generally contain NPCs with higher tribalism scores. Note that NPCs with a default tribalism of 0 were excluded from this analysis.

These tribalism ratio scores were integrated into a hierarchical mixed effects model as the outcome variable, with *number of prior persuade opportunities* as a fixed effect and a nested random effect capturing *AP* status and each player's unique *student* identifier. If students used tribalism reduction discourse moves more efficiently with practice, we would expect that the *number of prior persuade opportunities* will yield a negative coefficient (i.e., as practice increases, tribalism ratios decrease).

Table 7.5 shows the results of the mixed effect model. We found that the *number of prior persuade opportunities* was a significant, negative predictor of *tribalism ratio* ($\beta = -0.10, p < .001$). Figure 7-9 shows mean tribalism ratio scores across *Persuade* opportunities in conjunction with the number of users at each opportunity count.

7.2.5 Choosing the Best Civil Discourse Move

In addition to the above skills, which constitute the primary learning objectives of the game, students also practice a number of skills that serve as secondary learning objectives. These skills and concepts are reinforced throughout gameplay, and assessed explicitly at the start of turns 2 and 3. Generally speaking, student performance on these assessments represents their ability to choose the most appropriate civil discourse move from a list of options. Individually, each question assesses the student's understanding of a specific concept.

This data represents each opportunity each student has to answer one of a random set of multiple-choice questions. The answers to each question correspond to discourse move cards that the students interact with throughout the game. As some cards serve the same general function, some questions may have more than one correct answer. Therefore, the choices presented to the student at each opportunity are generated as follows: First, a correct answer is randomly chosen

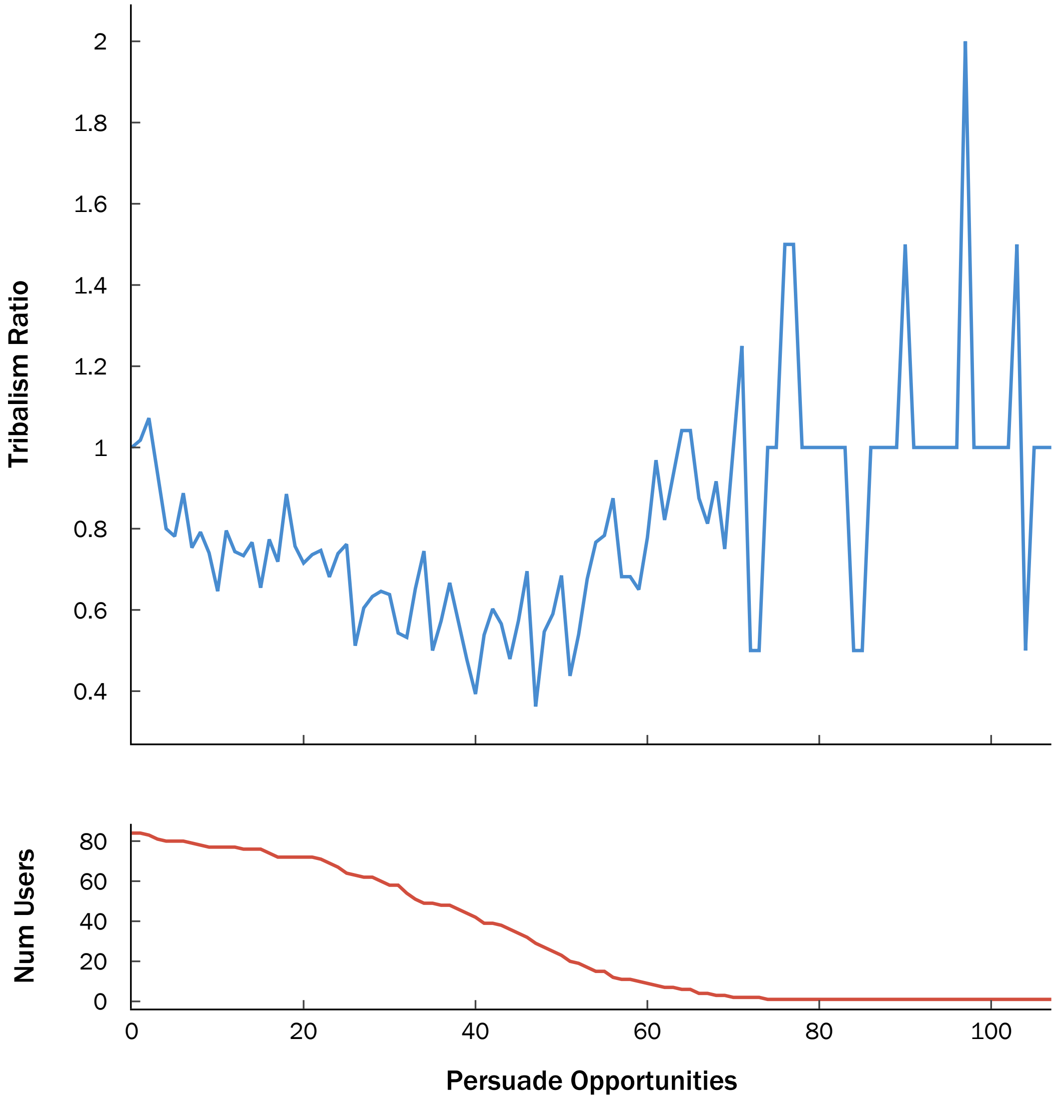


Figure 7-6: The top figure shows mean tribalism ratio scores by *Persuade* opportunities. The bottom figure shows the number of users by *Persuade* opportunities. We see that mean tribalism ratio decreases steadily with practice, while a majority of users were in the system. Note that ratio scores of greater than one are due to tribalism increases as a result of either player actions (e.g., playing the *Gossip* card) or alien actions.

Random Effects	Variance	SD
Student:AP	0.065	0.258
AP	0.000	0.000
Residual	0.405	0.637

Fixed Effects	Estimate	SE	df	t	p
Prior Opportunities	-0.101	0.012	3278	-7.957	< 0.001
(intercept)	0.715	0.030	78	23.078	< 0.001

Table 7.5: Summary of the hierarchical mixed effects model used to measure the impact of practice on the use of tribalism reduction discourse moves. *Prior Opportunities* (i.e., practice) was a negative predictor of *Tribalism Ratio*, the outcome variable. That is, more practice was associated with greater reductions in tribalism ratio, as predicted.

from a set of possible correct answers. Second, two incorrect answers are randomly chosen from a set of possible incorrect answers. Finally, the order in which the three options are presented is randomized.

General Performance Across Questions

First, we tested whether or not, with practice, student acquired a general skill representing their ability to chose the most appropriate civil discourse move from a list of options. The following hierarchical mixed effect model was generated to test the impact of practice on question performance:

$$Outcome \sim priorOpportunities + (1|AP/Student) \tag{7.5}$$

Where *Outcome* is a binary variable representing correctness and *priorOpportunities* represents the number of times a student has previously attempted any of these multiple choice questions (regardless of the concept the question assesses). We found that the number of prior opportunities was a significant positive predictor of performance ($\beta = 0.469, p < .001$). This suggests that students' ability to correctly choose the most appropriate civil discourse move did improve with practice. Figure 7-7 shows the relationship between opportunities and performance on these multiple-choice questions.

Performance on Individual KCs

The previous analysis represents a general or single-skill KC model. We also examined performance using a more specific four-skill KC model. The four skills in question are:

1. **Listen:** If you're not sure what someone values you can either listen carefully to what they're saying, or simply ask them what they think the most important issue is. People won't listen to you if you don't listen to them first.

2. **Confirm Understanding:** Even if you're pretty sure you know what someone values, it's important to confirm your guess with them. This also builds respect because it shows

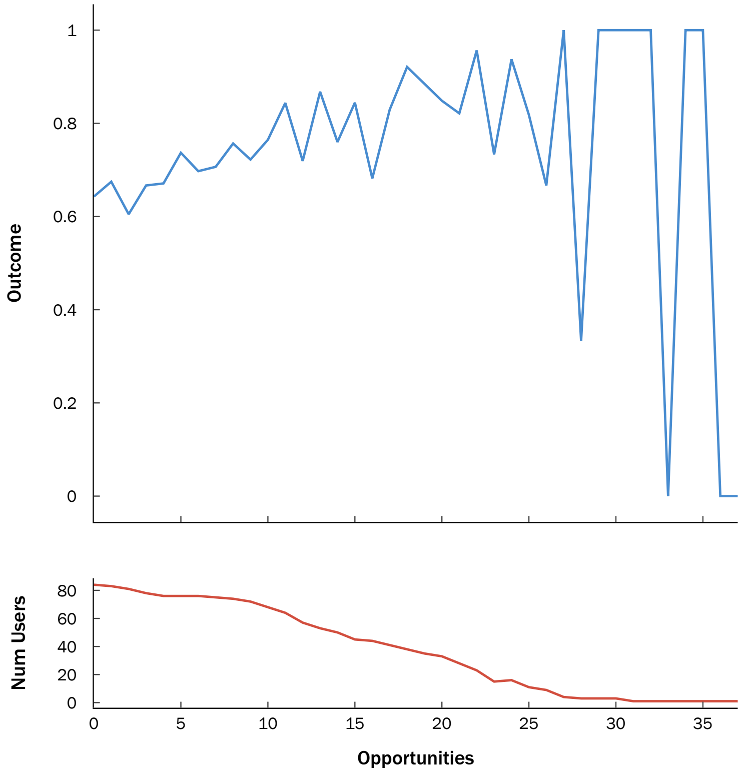


Figure 7-7: The relationship between opportunities and performance on multiple-choice questions generally. Below is the number of users at each opportunity. Performance improves with practice, while a majority of users were in the system.

you were listening.

3. **Conversation Reset:** When a discussion gets heated, it's best to take a step back and remind everyone of your shared goals.

4. **Point of Agreement:** Starting from a point of agreement helps remind everyone of your shared goals.

Performance on each of these skills was assessed using the following hierarchical mixed effects model:

$$Outcome \sim priorOpportunities + (1|AP/Student) + (1|KC) \tag{7.6}$$

Where *Outcome* is a binary variable representing correctness and *priorOpportunities* represents the number of times a student has previously attempted a question targeting the same KC as the current question. The additional random effect *KC* is a categorical variable representing which KC the student is practicing during the given opportunity.

We found that the number of prior opportunities was a significant positive predictor of performance ($\beta = 0.472, p < .001$). Importantly, the model specified using the four-skill KC model (AIC=1484) outperforms the model specified using the single-skill KC model (AIC=1522), suggesting that, while students' performance on these multiple choice questions generally improves with practice, the relationship between practice and performance is more accurately captured if we consider these skills individually. Figure 7-8 shows the relationship between opportunities and performance across the four KCs modeled.

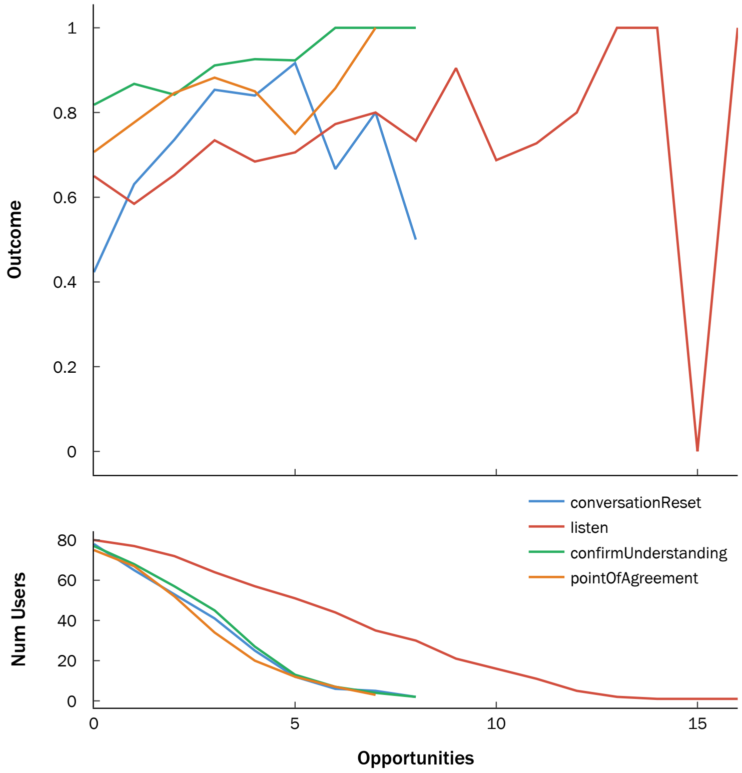


Figure 7-8: The relationship between opportunities and performance on multiple-choice questions targeting specific knowledge components. Below is the number of users at each practice opportunity for each KC. Across all KCs, performance improves with practice, while a majority of users were in the system.

7.3 Measuring the Impact of Bias

7.3.1 Problem-Level Estimates

We first examined our ability to estimate bias at the problem-level. The data for this model are each opportunity at a *Persuade* action for each student. We used the following model to test two estimates of myside bias:

$$Outcome \sim biasEstimate + (1|AP/Student) \tag{7.7}$$

Where *biasEstimate* is one of two metrics used to estimate the potential impact of myside bias. Recall that *Persuade* actions require students to identify which foundation an NPC values. The first metric used to estimate bias is simply the student's score on the Moral Foundations Questionnaire for the value they are asked to identify (i.e., the NPC's value). This metric served as our baseline measure. We found that this baseline measure of potential bias was not a significant predictor of performance.

The second metric used to estimate potential bias is our data-driven measure, *alignment*. Recall that *alignment* represents the cosine similarity between the student's values (as measured by the Moral Foundations Theory Questionnaire) and the values latent in some small piece of text (as estimated using distributed dictionary representations). In this case, the small piece of text that we use to compute *alignment* is the text of the correct response to the *Persuade* action. Put simply, *alignment*, in this case, is the degree to which a user's values align with the values latent in correct response to the *Persuade* action in question. We found that, like our baseline measure, *alignment* was not a significant predictor of performance.

In Section 7.2.3: *Shared Value Identification* we discussed that the *Persuade* action is likely actually composed of two separate but related knowledge components: 1) identifying which foundation the NPC values, and 2) identifying an argument from the opposing side that appeals to that foundation. Recall that we previously isolated the second knowledge component by selected a subset of *Persuade* opportunities, in which the player has already correctly identified the NPC's value. Performance on this subset of opportunities represents performance on this second KC (i.e., how well students are able to identify values across ideological lines). To examine the potential impact of bias on this process, we implemented the following hierarchical mixed effects model using this subset of *Persuade* opportunities as the data:

$$Outcome \sim priorOpportunities + biasEstimate + (1|AP/Student) \tag{7.8}$$

Interestingly, we found that our baseline metric of potential bias we not a significant predictor of performance in this model, though *alignment* was a significant positive predictor of performance ($\beta = 1.12, p < .01$). This suggests that students had an easier time identifying value across ideological lines when their values aligned with the values they were asked to identify.

7.4 Debiasing Intervention

Finally, we examined the impact of an intervention (designed to reduce bias) on both in-game performance and pre-post measures. Recall that students in the experimental condition had an in-game experience identical to those in the control condition with one exception: during *Persuade* actions, students in the experimental condition saw an additional piece of instruction that highlighted the option that most aligned with their values (i.e., had the highest computed *alignment*) alongside a messaged warning the player that they may be biased to select the highlighted option.

Inclusion of *Experimental Condition* as fixed effect to the above reported models of value-identification task performance showed that condition was not a significant predictor of performance. Similarly, a series of t-tests showed no significant differences across conditions on measures of enjoyment, motivation, and pre-post test gains. However, the presence of the intervention appears to have had an impact on the relationship between bias regulation and performance.

7.4.1 A Composite Measure of Bias Regulation

Up to this point, our *alignment*-based estimates of potential bias have represented the extent to which the user's values align with the values of the correct persuade option's text. This made a direct comparison between *alignment* and our baseline measure possible. However, this *alignment*-based estimate is limited in that it fails to account for the alignment between the user and the other potential options. We used these other alignment scores to generate a more nuanced estimate of the amount of potential bias a student may be overcoming at each opportunity. This new composite metric, which we call the *Bias Regulation Index* (BRI), is computed as follows:

$$BRI = (Alignment_{highest} - Alignment_{chosen}) + (Alignment_{correct} - Alignment_{chosen}) \tag{7.9}$$

In this model, $Alignment_{highest}$ represents the alignment score of the option with the highest alignment score (i.e., the option we would expect a completely biased player to pick).

Similarly, $Alignment_{chosen}$ represents the alignment score of the option the player chose, and $Alignment_{correct}$ represents the alignment score of the correct option. The first set of parentheses in this equation essentially gives the player credit for choosing an option that isn't the option with the highest alignment, and gives them more credit the farther away their choice's score is from that highest score. This first set of parentheses cannot penalize players, as they cannot chose an option with a score higher than the highest score.

The second set of parentheses penalizes the player if they chose an option with a higher alignment score than the correct option. If $Alignment_{correct} < Alignment_{chosen}$, then the result of this second set of parenthesis is set equal to 0 to keep the metric from crediting the players for choosing an incorrect option with lower alignment than the correct option. Importantly, players are neither penalized or credited in this metric for choosing the correct option. The resulting sum of these two sets of parentheses represents a student's ability to overcome bias to choose the correct answer. Positive scores on this metric capture those instances in which a student chooses the low-aligned correct score over the high-aligned incorrect one. Negative scores capture those instances in which the player chooses a high-aligned incorrect option over a lower-aligned correct one.

This metric is more nuanced than simply including the *alignment* score of the correct option, as it mitigates the impact of the option's correctness on choice. That is, did the player choose this because it is the correct option, or because it aligned with their values. When the correct option is also the option with the highest alignment score, the choice is easy and uninteresting. This metric focuses on instances in which the choice is difficult.

7.4.2 Interaction Between Condition, BRI, and Number of Opportunities

We expect that the relationship between bias regulation and performance may be impacted by experimental condition (i.e., the presence or absence of the intervention) and practice. To test for interactions between experimental condition, practice, and students' ability to regulate bias with respect to performance, we incorporated this new *Bias Regulation Index* (BRI) into the following hierarchical mixed effects model:

$$Outcome \sim BRI * priorOpportunities * condition + (1|AP/Student) \qquad (7.10)$$

As expected, we found a significant three-way interaction between *Bias Regulation Index*, the number of prior practice opportunities, and experimental condition. We used the R library *interactions* to explore and visualize this interaction. Figure 7-9 shows the relationship between BRI and Performance at three different opportunity counts. We see that, while the relationship between performance and BRI remains relatively stable across practice opportunities in the control condition, the relationship between these variables changes with practice in the adaptive condition. Recall that BRI scores below zero indicate opportunities in which the student chose a high-aligned incorrect option over a low-aligned correct one, and positive scores indicate opportunities in which the student chose a low-aligned correct option over a high-aligned incorrect one. This graph suggests that the intervention may have caused students with low bias-regulation to perform worse (potentially choosing the salient orange-colored option more). However, students with high bias-regulation seemed to benefit from seeing the intervention, outperforming their peers in the control condition.

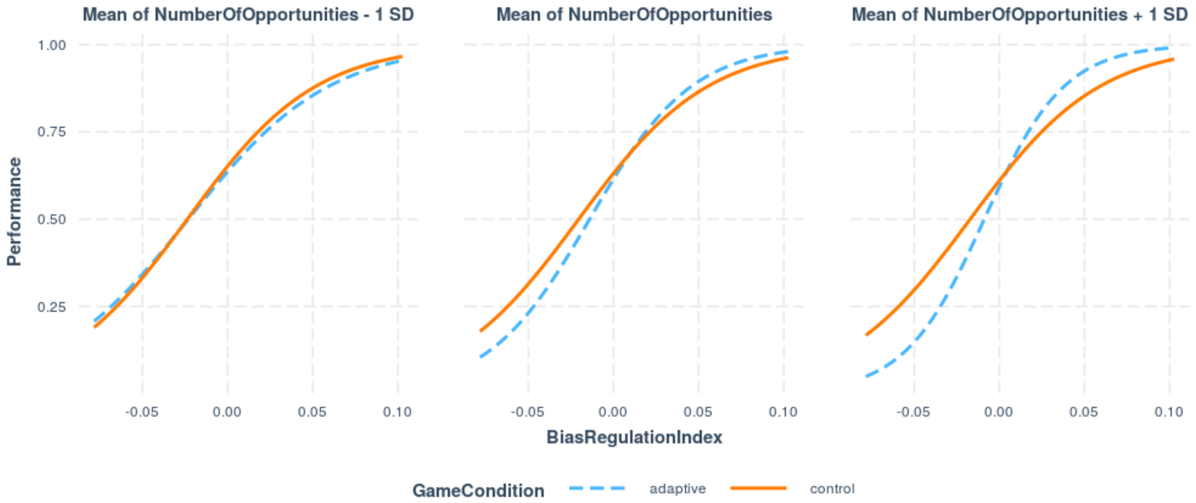


Figure 7-9: The three-way interaction between condition, opportunities and *Bias Regulation Index* (BRI). We see that, while the relationship between BRI and performance remains relatively constant with additional practice in the control condition, the relationship seems to change in the adaptive condition. As the number of practice opportunities increase, students in the adaptive condition with low bias-regulation appear to do worse than their peers in the control condition, whereas students in the adaptive condition with high bias-regulation appear to benefit from the intervention compared to their peers in the control condition.

7.5 Conclusion

In this chapter we reported the results of our full-scale experiment with respect to four key areas: player experience, learning, bias estimation, and the impact of our intervention. We found that, in general, students enjoyed the game, their performance on key skills improved with practice, and our data-driven measure of *alignment* allowed us to estimate the impact of myside bias on student performance. In the next chapter, we will discuss the implications of these results with respect to the broader contexts of civic technology, learning science and human-computer interaction.

Chapter 8

Discussion

"The fact that the little arguments that groups were having determined the fate of humanity was pretty ironic and entertaining."
- a student that gets it

While the results presented in the previous chapter are promising, it is difficult to assess their worth in isolation. In this chapter, we discuss how the results of our experiment relate to three broader contexts: civic technology, learning science, and human-computer interaction. This chapter also includes a discussion of the limitations of this work, including improvements to future iterations of the game.

8.1 Games for Civic Education

Our results provided a great deal of evidence supporting the use of games to teach civic skills. In general, students found our game to be meaningful to their real-world lives and enjoyable to play. Our results also showed that the game was an effective way to practice fundamental civil discourse skills, perhaps due, in part, to the unique features of game environments.

8.1.1 Meaningful Connections to Students' Lives

During the expert interviews conducted at the beginning of this project, several instructors emphasized the importance of relevance in maintaining student interest and motivation. One teacher went so far as to say that, "If it's not relevant to their lives, they don't care." We took this advice very seriously during the design process, altering the content of scenarios to reflect issues students might have a stake in (e.g., school uniforms, parks, corporal punishment). There is, unfortunately, no evidence of the impact of these design choices, as the content of scenarios was only mentioned once in student feedback (to say they were ironically petty). We did, however, find evidence suggesting that students found the material relevant on a deeper, more meaningful level.

The post-test survey found that a majority of students agreed that playing the game was valuable to them, and the open-ended responses provided some clues as to what aspects of the game may have been valuable to students. Again, it is worth emphasizing that students did not mention any of the topics covered in the game in their open-ended responses. That is, no students said that playing helped them form a belief or got them to thinking about some specific topic. Instead, students reflected on how the game helped them practice empathizing with opposing positions and de-escalating unproductive arguments. This feedback is important on two fronts.

First, it suggests that we achieved our goal of designing a civics game that focuses on civic skills rather than surface-level conceptual knowledge. Despite the fact that students encountered a variety of interesting topics throughout the game, student feedback suggests that the deeper, skill-oriented goals of the game were clear. Second, several students related these skills to their lives directly (e.g., "I liked how it kind of showed how to to act in real life situations"). This may be, in some cases, what students found to be meaningful and valuable, and highlights the shortsightedness of our attempts to increase relevance. While we focused a great deal of effort on crafting relevant content, we ignored the potential relevance the skills the game was designed to teach. In truth, sustaining student interest likely requires both relevant content and relevant skills, but future work should make an intentional effort to capitalize on this untapped reservoir of relevance. One simple solution is to make the implicit connections between in-game skills and real-life explicit. For example, the game could ask players to imagine a scenario playing out in their own life, and to choose the most appropriate civil discourse move. However, this approach also comes with drawbacks: Asking players to jump from the game world to the real world and back will likely break immersion, and removing the emotional distance between player and character may make players more resistant to behavioral change.

Understanding how best to strike a balance between making connections to the student's real life and maintaining immersion requires more research in this area. However, it is worth reiterating that the game made no explicit connections to students' real lives, and yet prompted students to make these connections regardless. This may be due to our choice to design a game that is a simplistic model of realistic interactions about real-world topics. Even the main fantasy element (the alien invasion) was a thinly-veiled metaphor for real-world threats (e.g., foreign election interference). Elements of realism may be one method for creating bridges to students' real lives without breaking immersion.

One somewhat inconsistent result that seems to undercut the meaningfulness of the game is the percentage of students that agreed with the statement, "Playing this game will change how I discuss politics in the future." While a majority of students still agreed with this sentiment, the percentage was curiously smaller than students who, for example, enjoyed the game or admitted to learning something. One possible explanation for this discrepancy is some potential ambiguity in the question's wording. Rather than interpreting the phrase "how I discuss politics" as meaning one's approach to engaging in political discussions, some students may have interpreted this to mean, "this game has changed my opinion on some topic" (something the game is intentionally not designed to do). A second, related, potential explanation relates to the honesty of students' self-reflections. That is, this question requires that students implicitly admit that the way they engaged in political discussions before was, in some way, flawed or in need of improvement. Because our political beliefs are so intricately tied to our sense of self, a kind of self-preservation instinct might inhibit complete honesty in this response. On first glance, it may seem reasonable to rule out this second possibility considering the fact that students seem willing to honestly admit shortcomings in their knowledge or ability (as per the *learning* question). However, it is possible that one holds their political beliefs closer to their sense of self than their knowledge or ability.

Key Takeaway 1 *Student feedback suggested that players found the game to be meaningful and valuable to their real-world lives. This is likely related to our choice to focus on practical civic skills.*

8.1.2 Improving Performance on Civic Skills

As we discussed above, civics games can allow us to provide an experience that is authentic and meaningful. Our results also suggest that games in this space can also provide practice that effectively progresses students towards the mastery of key civic skills. Below is a discussion of the key findings of this work with respect to civic education.

First, we found that students' ability to identify the values latent in text improved with practice. This is the first time, to our knowledge, that a value-identification task drawing directly from Moral Foundations Theory has been successfully implemented in an educational game. Importantly, our results suggest that what we characterize as *value-identification* is likely a general class of five specific skills (one per foundation), as student performance varied across foundations. There are many potential reasons for this variation. For example, it may be due to population variations in foundational distribution (i.e., student populations may tend to prioritize certain foundations over others). It is also possible that this variation is due to conceptual complexity. The *Sanctity* foundation, for example, is easy to feel but difficult to verbalize (consider Justice Stewart's difficulty formally defining the threshold for obscenity, instead saying simply, "I know it when I see it"). Contrast that with the *Care* or *Fairness* foundations which are easy to verbalize and whose texts may even contain linguistic cues that hint at the foundation's presence.

In addition to the general skill of identifying the values latent in text, we also examined a more specific skill: identifying values latent in text across ideological lines. That is, how well can a student identify a value from a belief on one side of the issue, and then identify a belief that matches that value from the other side of the issue. Our results in this respect were inconclusive. This is potentially due to the fact that our method for distinguishing between the general value-identification task and this more specific shared value-identification task was limited by a ceiling effect. While these differences in the impact of practice do not provide direct evidence of two separate value-identification skills, they justify further work in this area. And further work in this area is critical, as it is perhaps this skill more than any other that opens the door to productive discourse between two ideologically opposed discussants.

Students also learned to more efficiently use tribalism-reducing discourse moves with practice. The strategic use of some of these discourse moves (e.g., *Point of Agreement*, *Conversation Reset*) requires that students understand when best to use them. Thus, our metric (tribalism ratio) provides a measure of the student's prerequisite conceptual understanding as well as their performance on the skill of choosing the most appropriate discourse move for reducing tribalism. Importantly, the strategic effectiveness of these discourse moves in-game mirrors their strategic effectiveness in real-life, a connection that some students commented on in their open-ended responses. These tribalism-reducing strategies are particularly important for populations (like students) in which a large proportion of social interactions are mediated via technologies that intentionally or unintentionally increase political tribalism.

Finally, students seemed to develop a conceptual understanding of several concepts that served as secondary learning goals. Students' ability to identify when to use discourse moves like *Confirm Understanding*, *Listen*, *Point of Agreement*, and *Conversation Reset* improved with practice. Again, this suggests a conceptual understanding as well as the development of the skill required to discern the most appropriate move among other options.

Key Takeaway 2 *Students' performance on many of the civic skills of interest improved with practice. The impact of practice on students' ability to identify values across ideological lines may have been obscured by a ceiling effect.*

8.1.3 Using Simulations to Assess Civic Skills

Assessing student performance on civic skills can be challenging, as there is often not one correct action that one ought to take. And while it is possible to assess student responses to tasks like, "Describe how you would respond to this argument," such assessments can be difficult, prone to bias, and unscalable. A simple solution to this problem is to ask students to choose the most appropriate response from a list of options. This transposes the task from the correct/incorrect dichotomy to a better or worse spectrum, which minimizes (but does not eliminate) many of the above drawbacks.

Games are uniquely suited to implementing these kinds of "choose the best option" comparison tasks because games can immerse players in environments rich with relevant contextual information. Consider the pieces of information one might need or want when navigating a tense political discussion. At a minimum, the discourse moves one takes are dependent on who they are talking with, what their relationship to that person is, and what they are talking about. A written test would require all of these relevant details be written out and read by students. However in games, as in real-life, this information can be inferred through past experience. Practically speaking, this reduces the cognitive load of students and mitigates issues associated with asking students to read long, detail-rich passages to answer a question (e.g., time, fatigue, reading comprehension).

The richness of contextual information in game environments is especially powerful when: 1) in-game contexts are designed to mirror real-life contexts, and 2) in-game actions are designed to mirror real-life skills. These two features likely create additional bridges from the game world to the player's life, which may prompt near-transfer.

Key Takeaway 3 *Civic skills may be more easily assessed using comparison tasks. Game environments can provide the important contextual information required to answer these comparison tasks more efficiently than standard, paper-and-pencil testing methods.*

8.1.4 Enjoyment, Motivation, and Learning

Games, especially educational games, are not always fun, but we found that a majority of students found the game to be an enjoyable experience. And while some students were quick to remind us that the bar, in this case, is relatively low (e.g., "I liked that I didn't have to write an essay"), we also found evidence of genuine enjoyment in some students. For example, we found that a sizable proportion of students continued to play the game outside of their normal classroom hours, with some choosing to play the game at home (where we might assume the bar for what qualifies as an enjoyable experience is much higher). It is also worth reiterating that these positive outcomes on measures of enjoyment and motivation are likely not due to an easy or straightforward gameplay experience. On the contrary, many students' open-ended responses related to how challenging the game was, with a small number of students admitting that they enjoyed the challenging nature of the game.

Students' open-ended responses provide some hints as to what the students found enjoyable. Positive feedback relating to the design and central plot of the game validated some of our design choices (e.g., the use of an alien invasion to increase motivation). Similarly, several students mentioned enjoying the core persuasion mechanic of the game, suggesting that our simulated civic activity was sufficiently scaffolded as to not be overwhelming or boring. Finally, several students said that they, "liked seeing all the different people's values." It is unclear why this was compelling. For example, it may be the case that the most valuable feature of the game to some students was the chance to be exposed to different beliefs surrounding a topic. Alternatively, the act of "playing detective" to discover NPC beliefs may be intrinsically rewarding. Narrowing in on why students found these features to be enjoyable is a subject of future work.

Key Takeaway 4 *Student feedback suggests that they found the game to be enjoyable and motivating. Some students pointed specifically to the style, concept, and core mechanic as contributing positively to their experience.*

8.1.5 Balancing Scaffolding with Player Autonomy

While a majority of students agreed that they "felt like [they] had choices regarding how [they] wanted to play the game" (autonomy), a large number of students (46%) disagreed with that statement. This feeling of restriction expressed by some students may be due to the game following a relatively inflexible instructional plan. This plan is designed to scaffold student learning, introducing new concepts gradually in an effort to minimize cognitive load. Unfortunately, this approach requires student actions be limited (particularly during early scenarios), to ensure that students are learning the newly introduced skill or concept. Later scenarios provide the students with more freedom, but many students did not progress to those scenarios during the allotted gameplay time. The data support this possible explanation, as students who completed more than half of the scenarios reported significantly higher (t(28), $p<.01$) ratings of autonomy ($M=2.72$, $SD=1.41$) than those who completed half or fewer than half of the scenarios ($M=2.19$, $SD=1.51$). Future iterations of the game might increase feelings of autonomy by alternating between scenarios designed to progress the student through the instructional plan, and scenarios designed to allow the student to freely practice the skills they have learned.

Key Takeaway 5 *Though scaffolding instruction likely increased the efficiency of learning, it may have come at the cost of player autonomy.*

8.2 Expanding the Boundaries of Personalized User Experiences

Unlike other domains, the nature of tasks in the civics domain can change depending upon the beliefs of the person reading the content or doing the task. Perhaps the most novel contribution of this work is the development of a system capable of measuring this dynamic relationship between user and content, and creating individualized instruction based on that relationship.

8.2.1 Differing Approaches to Value-Adaptive Instruction

We tested our data-driven estimate of potential bias, *Alignment*, against a reasonable baseline estimate: the student's score for the foundation of interest on the Moral Foundations Theory Questionnaire. It is worth reiterating what is meant by baseline in this case. First, recall the business-as-usual approach to value-adaptive instruction in the civics domain, in which teachers are required to maintain a working knowledge of student beliefs and then adapt instruction based on how those beliefs relate to class content. While this approach certainly has the potential to be the most nuanced (as content is adapted on an individual basis by an expert instructor), it is also unscalable.

Consider next our baseline measure. Recall that value-adaptive instruction requires us to relate user-values to the values-latent in content. Using a student's scores from the Moral Foundations Theory questionnaire would allow us to develop a value-adaptive system that is scalable on the user side of that equation, but the system would still require that humans

Estimate	**Scalable**	**Requires Labeled Data**	**Multidimensional**
Student Interactions			X
Moral Foundations Score	X	X	
Alignment	X		X

Table 8.1: A comparison of different approaches to estimating bias for value-adaptive instruction. *Alignment* provides a scalable and nuanced way to estimate bias.

label the values latent in the content, limiting scalability. This method is also limited by its uni-dimensionality, as the system would only be able to consider each foundation in isolation.

Finally, consider our new, data-driven estimate of bias, *Alignment*. This metric not only eliminates the need for human-labeled content, but also provides a more nuanced, multidimensional estimate of potential bias. This mutlidimensionality is important, as our beliefs are often informed by more than one moral foundation. Table 8.1 summarizes the key differences between these three approaches to value-adaptive instruction

8.2.2 Alignment as an Estimate of Potential Bias

While neither *alignment* nor our baseline measure was predictive of student performance on a general value-identification task, *alignment* (but not the baseline measure) was predictive of performance on shared value-identification tasks. Recall that these are tasks in which the user has correctly completed the first step of the Persuasion action (identifying which foundation the NPC values), but not the second step (choosing which argument from the opposing side appeals to that foundation). This is the same subset of opportunities for which we suspect that a ceiling effect may have obscured the impact of practice on performance.

This suggests that students had an easier time identifying values across ideological lines when their values aligned with those they were asked to identify. That is, a student that values the *Care* and *Fairness* foundations will find it easier to identify these values in opposing arguments. This result has an important implication for engaging in productive civil discourse: People will likely find it easier to empathize across ideological lines when the beliefs are rooted in shared values. This is in contrast to, for example, attempting to empathize with a value you do not hold, which is likely more difficult.

Key Takeaway 6 *Alignment (but not the baseline measure) was predictive of performance on tasks that required player to identify values across ideological lines. This suggests that perhaps the easiest way to find common ground is to empathize with arguments from the opposing side that appeal to the values that you hold.*

8.2.3 Value-Adaptive Debiasing Interventions

In addition to exploring the use of *alignment* as an estimate of bias, we tested an *alignment*-driven value-adaptive debiasing intervention. The impact of this intervention on performance proved to be subtle. For example, we found no significant differences in performance between the adaptive condition (which included the intervention) and the control condition (which did not). However, we did find that the intervention seemed to impact the relationship between a student's ability to regulate their own bias and performance. Our data suggest that students in the adaptive condition with low bias-regulation may have been negatively impacted by the intervention relative to their peers in the control condition. However, students in the adaptive condition with high bias-regulation may have benefited from the presence of the intervention relative to their peers in the control condition.

In one sense, these results are promising. They show that for some students, a debiasing intervention may have effectively reduced their bias, allowing them to more accurately identify and appeal to another's values. However, the apparent negative impact of the intervention on some portion of students in the adaptive condition suggests that there is still a number of challenges to be overcome in this area (e.g., system trust, instruction about the role of bias). One potential explanation for this negative impact is the salience of the intervention. Instruction on what bias is and how it impacts decision-making was limited. Despite the presence of an explanation, students may have misunderstood what the orange-color represented, and took it to be an indication of correctness or incorrectness, when, in actuality, the color has no relation to an option's correctness. An alternative explanation (based on our previous work testing debiasing interventions) is that some students did not trust the accuracy of the intervention's prediction. However, in this case we would expect that these students would ignore the intervention and produce performance roughly equivalent to their peers in the control condition – a result not supported by the data. Future work will include opportunities for qualitative feedback that address the nature and source of this negative consequence more directly.

In the context of education research more broadly, this result demonstrates the practical usefulness of value-adaptive instruction as a tool for addressing complex and important challenges, like debiasing, in a nuanced and scalable way.

Key Takeaway 7 *For some students, the value-adaptive debiasing intervention may have effectively reduced their bias, allowing them to more accurately identify and appeal to another's values. However, for students exhibiting low bias-regulation, the intervention may have negatively impacted performance, potentially due misunderstandings about the system and the role of bias in general.*

8.3 Limitations

Any complex social domain is fraught with opportunities for well-intentioned choices to have unintended consequences. For example, the powerful visual metaphor of townspeople coming together into the townsquare might also inadvertently endorse an *Argument to Moderation*, a logical fallacy that states that the best solution always lies between the two opposing sides[1]. More generally, we expect the most likely source of unintended consequences to be a result of the recharactization of the ill-defined domain of civil discourse into relatively well-defined game mechanics. Making this domain tractable necessarily requires the over-simplification of complex social processes and open-textured concepts. For example, in the game, a certain discourse move might predictably lower tribalism. However in real life, the same action, if employed without tact or in the wrong context, may result in increased tribalism.

Similarly, the simulation of faulty feedback mechanisms was outside of the scope of the game. Helping students understand how unproductive discourse moves can result in false feelings of superiority or righteousness in the short term, but be damaging in the long-term is an important direction for future work. Our results (particularly those related to the value-adaptive intervention) may support this work in productive ways. For example, one could imagine a tool that

[1] One student's open-ended response suggests that this may have happened.

monitors a user's self-generated content (i.e., responses to social media posts), using alignment to the user's known values as an indicator of self-righteous speech.

Other limitations of this work relate to the sample of students used. First, as we noted above, the onset of the COVID-19 pandemic left us unable to collect data from an additional, non-remedial English class. Thus, any grade level effects may be skewed by what may be an unrepresentative sample of 10th graders. That said, it is likely the case that the game is more appropriately leveled for older students. Future work will expand participation to adults of all ages.

Second, while Haidt claims his Moral Foundations Theory can be applied across cultures [35], the scenario content in this game was specifically designed for the American political context and may not transfer well to other political contexts. That said, provided the content can be adjusted appropriately, this game may be a tool for studying differences in moral foundation weights across cultures – an exciting direction for future research.

Finally, this game is limited by its single-player nature. As one student pointed out, "Doing video games distracts from human connections." And while this game was intended to be just one part of a larger experiment that included real-world class discussions, future work will also make efforts to better capture the social dimensions of this work. For example, teachers may have students engage in a hybrid classroom discussion in which they are given, as a reference, physical copies of the discourse move cards that they use in the game.

8.3.1 Improvements to the Game

There are a number of improvements to the game that will be implemented in future iterations, what follows are some of the most important: Future iterations will make connections between the game and the real-world more explicit. For example, an explanation of the use of alien invaders as a metaphor for foreign election interference will be incorporated into the post-test. Future iterations will also include a new discourse move designed to assess students' mastery of shared value-identification. Finally, future iterations will include a more robust student feedback system. The current system allowed us to determine, for example, that the system was enjoyable and meaningful, but it would be much more useful to know what aspects are enjoyable or meaningful.

8.4 Conclusion

In this chapter, we discussed in detail the most important implications and contributions of this work with respect to three areas of interest: civic technology, learning science, and human-computer interaction. In the final chapter, we will summarize these contributions and expand on the future work supported by our findings.

Chapter 9

Conclusions and Future Work

In this document, we have discussed three major 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
3. the absence of any acknowledgement of the role of values in informing our beliefs and biasing our reasoning

In an effort to address these gaps, we developed an educational game, Persuasion Invasion, that offers students:

1. opportunities to practice civil discourse skills in a scaffolded, low-consequence environment
2. instruction designed to help students understand their own values and the values of others (outside of the social pressure to conform to the established beliefs of the tribe)
3. adaptive instruction that uses a model of the interaction between student and content values to estimate and reduce myside bias

We found that students found the game to be both enjoyable to play and meaningful to their real-world lives. Log data collected during gameplay also shows that student performance on key civil discourse skills generally improved with practice. Additionally, we found that a data-driven estimate of the relationship between a player's values and the values latent in the content they were reading was predictive of performance on a value-identification task. This suggests that our game, for the first time to our knowledge, was able to estimate potential myside bias in a scalable way. Finally, our debiasing intervention demonstrated, for the first time, that value-adaptive systems can be used to combat biased reasoning in a nuanced and scalable way.

9.1 Directions for Future Work

9.1.1 A Language of Common Values

In the 2011 CIRCLE report on the state of civic education in America, the authors emphasize the importance of public discourse driven by core ideals, citing it as a "means of addressing the root causes of vacuous and sometimes vicious dialogue." To that end, our game emphasizes the skill of identifying when these common values might be motivating a belief. However, there may also be some utility in simply having a shared language of values. Future work will examine the role of Moral Foundations Theory not only as a tool for understanding beliefs, but as a sort of linguistic scaffold, used to structure or support dialogue between students familiar with the concepts.

9.1.2 Augmenting Reasoning in Bias-Prone Interactions

The validation of *Alignment* as a measure of bias in high school students has direct applications in civics classrooms, as it allows us to create instruction that is not only personalized, but also scalable. However, the notion of a value-adaptive system may have applications outside of the classroom as well.

In any interaction that requires human reasoning, the incorporation of measures like *Alignment* into systems represents a shift from a narrow, cognitive-first approach to a more holistic and realistic approach. We envision new systems that form a reasoning partnership with their users, leveraging the strengths of each. The machine can tirelessly search (with relative objectivity) for content that may leave the user prone to biased reasoning, but offload the act of reasoning about that content to the expert human reasoner. These just-in-time interventions may be able to counteract the effects of content designed to short-circuit our critical reasoning faculties by appealing directly to our System 1 moral intuitions. In highly connected networks (such as social media platforms), even a small number of users armed with a second, more critical look may stop emotionally-appealing misinformation from propagating with impunity. In this sense, augmented reasoning systems may be one method for destabilizing information bubbles.

Recall that bias is akin to a chronic disease, requiring continual treatment rather than a one-time cure. As such, any augmented reasoning system would likely need to be integrated into users' lives, at least for some extended period of time. For example, the continual treatment of our bias might require that users install a browser extension that acts as a second set of (relatively) unbiased eyes and periodically alerts the user when they are reading some content that may leave them susceptible to bias. While the technical challenges of implementing such a system are clear, how users will feel about interacting with this kind of solution is less clear. Building trust in the system, an understanding of bias, and an appreciation for its consequences are all tasks not well suited to the brief interactions of a just-in-time intervention. It is likely that this is reason why our previous attempts to create effective debiasing interventions have failed. Instead, these transformations likely require a more immersive and focused educational experience. If the browser extension is akin to spell-check, users still need a foundational course in grammar. Our game was designed, to a large extent, to provide that foundational, immersive learning experience. Future work will focus on pairing this foundational experience with the brief, just-in-time feedback intended to address bias over time and in the real-world.

9.1.3 Democratizing Editorial Discretion

The rise of social media has been accompanied by a rise in smaller, decentralized media sources. Until now, social media platforms themselves have had editorial discretion over which headlines get seen, a responsibility that they have (to a large extent) given to algorithms designed to maximize screen time (rather than a diversity of perspectives or journalistic integrity). While readers may have a desire for more-trustworthy news sources, there is a recognition that our biases limit the reliability of self-regulated media-consumption. Media consumers are seemingly faced with an impossible task. Without guidance, consumers are likely to choose (subconsciously or otherwise) stories that appeal to their values and beliefs. With the guidance of algorithm-curated news feeds, the problem is exacerbated, as consumers are only shown stories that appeal to their values and beliefs.

In this respect, this future work represents just one instance of a new, user-centered approach to augmenting media consumption in a way that is more transparent and user-empowering than algorithm-driven news feeds. Making potential bias visible to media consumers can 1) help users to better regulate their own bias and 2) make apparent the information bubbles created by algorithm-curated news feeds.

9.2 Conclusion

Civil discourse is how we, as citizens, build the consensus needed to reach shared goals. It is a foundational civic activity for any functioning democracy. As such, unproductive discourse does not simply result in uncomfortable conversations, it also threatens to halt progress towards any goal that requires thoughtful consideration from all members of a community. In spite of this importance, students rarely have an opportunity to practice the skills that support productive civil discourse. Instead, as the national civics curriculum continues to dwindle, what remains is content focused on the structure of government, content largely devoid of the practical skills students need to master in order to be able to effectively engage with one another as citizens in a democracy.

This shift in the focus of the civics curriculum has important implications for civic engagement. In the expert interviews conducted in preparation of this work (see Section 4.7), teachers described passionate (and sometimes combative) students deeply invested in classroom discussions, even as the same teachers admit that it is often a struggle to keep students engaged when teaching the curriculum content. We believe games can help close this “passion-gap” by immersing students in simulated environments that are both rich with contextual information and carefully scaffolded to maintain motivation and foster efficient learning. Provided the in-game interactions are authentic, these experiences can also be meaningful and valuable to students in their real-world civic lives.

The skills students can learn throughout meaningful and enjoyable civic simulations provide a roadmap for interacting with likeminded peers, those they disagree with, and questionable news media. Conversely, by failing to provide students with practice and guidance on these skills, we leave them vulnerable to the allure of political tribalism and unequipped to engage in civil discourse that is productive. Value-adaptive instruction addresses only a small number of the challenges present in the civic education space. However, adapting instruction to user values also represents a fundamental shift in civic education priorities. It requires us to recognize the critical importance of practical civil discourse skills and the role that bias plays in how and when we exercise those skills. It requires that we impart an appreciation both for the immense power that civil discourse has in shaping the course of our democracy, and for the serious consequences of not exercising that power judiciously.

Bibliography

[1] Monica Anderson and Jingjing Jiang. Teens, social media & technology 2018. *Pew Research Center*, 2018.

[2] Mahmood Reza Atai and Fatemeh Chahkandi. Democracy in computer-mediated communication: Gender, communicative style, and amount of participation in professional listservs. *Computers in Human Behavior*, 28(3):881–888, 2012.

[3] Marc David Baer. An enemy old and new: The dönme, anti-semitism, and conspiracy theories in the ottoman empire and turkish republic. *Jewish Quarterly Review*, 103(4):523–555, 2013.

[4] Bence Bago and Wim De Neys. Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, 158:90–109, 2017.

[5] Benjamin R Barber. Three scenarios for the future of technology and strong democracy. *Political Science Quarterly*, 113(4):573–589, 1998.

[6] Thomas W Benson. Rhetoric, civility, and community: Political debate on computer bulletin boards. *Communication Quarterly*, 44(3):359–378, 1996.

[7] Bill Bishop. *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt, 2009.

[8] Michael X Delli Carpini and Scott Keeter. *What Americans know about politics and why it matters.* Yale University Press, 1996.

[9] James Carse. *Finite and infinite games.* Simon and Schuster, 2011.

[10] Civic Education Classrooms. Paths to 21st century competencies through civic education classrooms. 2009.

[11] Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198, 2015.

[12] CodeMoose. Rpg pixel sprites - mini folk, 2020.

[13] Joshua Cohen. Deliberation and democratic legitimacy. *1997*, pages 67–92, 1989.

[14] Stephen Coleman and John Gotze. *Bowling together: Online public engagement in policy deliberation.* Hansard Society London, 2001.

[15] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82. American Society for Information Science, 2015.

[16] Terry Crowley and Claire Bowern. *An introduction to historical linguistics.* Oxford University Press, 2010.

[17] Sabrina Culyba. *The Transformational Framework: A Process Tool for the Development of Transformational Games.* 2018.

[18] Lincoln Dahlberg. Computer-mediated communication and the public sphere: A critical analysis. *Journal of Computer-mediated communication*, 7(1):JCMC714, 2001.

[19] Patricia G Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5, 1989.

[20] Maeve Duggan and Aaron Smith. The political environment on social media. *Pew Research Center*, 2016.

[21] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 153–162. ACM, 2015.

[22] J St BT Evans, Julie L Barston, and Paul Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306, 1983.

[23] J St BT Evans, SE Newstead, JL Allen, and P Pollard. Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, 6(3):263–285, 1994.

[24] Jonathan St BT Evans. Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468, 1984.

[25] Andrew S Franks and Kyle C Scherr. Using moral foundations to predict voting behavior: Regression models from the 2012 us presidential election. *Analyses of Social Issues and Public Policy*, 15(1):213–232, 2015.

[26] Shane Frederick. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42, 2005.

[27] Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, 50(1):344–361, 2018.

[28] Andrew Goldstein. Gov. wolf closes all k-12 schools in pa. for two weeks. *Pittsburgh Post-Gazette*, Mar 2020.

[29] Al Gore. Information superhighways speech, 3 1994. Remarks by Vice President Al Gore to the International Telecommunications Union, [Accessed: 2019 04 21].

[30] Jonathan Gould, Kathleen Hall Jamieson, Peter Levine, Ted McConnell, and David B Smith. Guardian of democracy: The civic mission of schools. *Report for the Campaign for the Civic Mission of Schools. Philadelphia, PA: University of Pennsylvania Leonore Annenberg Institute for Civics*, 2011.

[31] Lisa Guilfoile and Brady Delander. Six proven practices for effective civic learning. *Education Commission of the States and National Center for Learning and Civic Engagement*, 2014.

[32] Jurgen Habermas and Jürgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society.* MIT press, 1991.

[33] Jonathan Haidt. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814, 2001.

[34] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

[35] Jonathan Haidt and Jesse Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116, 2007.

[36] Geoffrey E Hinton. Distributed representations. 1984.

[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.

[39] David W Johnson and Roger T Johnson. Energizing learning: The instructional power of conflict. *Educational Researcher*, 38(1):37–51, 2009.

[40] David W Johnson, Roger T Johnson, and Dean Tjosvold. Constructive controversy: The value of intellectual opposition. 2000.

[41] Dan M Kahan, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature climate change*, 2(10):732, 2012.

[42] Daniel Kahneman. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697, 2003.

[43] Jonas T Kaplan, Sarah I Gimbel, and Sam Harris. Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific reports*, 6:39589, 2016.

[44] Kenney. Roguelike modern city, 2020.

[45] Sara Kiesler, Jane Siegel, and Timothy W McGuire. Social psychological aspects of computer-mediated communication. *American psychologist*, 39(10):1123, 1984.

[46] Paul A Klaczynski and Billi Robinson. Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology and Aging*, 15(3):400, 2000.

[47] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[48] Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality*, 46(2):184–194, 2012.

[49] Scott O Lilienfeld, Rachel Ammirati, and Kristin Landfield. Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on psychological science*, 4(4):390–398, 2009.

[50] Collin Lynch, Kevin D Ashley, Niels Pinkwart, and Vincent Aleven. Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, 19(3):253–266, 2009.

[51] Collin F Lynch, Kevin D Ashley, Vincent Aleven, and Niels Pinkwart. Defining ill-defined domains; a literature survey. In *Intelligent Tutoring Systems (ITS 2006): Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, 2006.

[52] Henry Markovits and Guilaine Nantel. The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17(1):11–17, 1989.

[53] Michael McDevitt and Spiro Kiousis. Experiments in political socialization: Kids voting usa as a model for civic education reform. circle working paper 49. *Center for Information and Research on Civic Learning and Engagement (CIRCLE), University of Maryland*, 2006.

[54] Sarah McGrew, Teresa Ortega, Joel Breakstone, and Sam Wineburg. The challenge that's bigger than fake news: Civic reasoning in a social media environment. *American Educator*, 41(3):4, 2017.

[55] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[56] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[57] Pennsylvania Department of Education. Pa state standards: Civics and government, 2019.

[58] Ozzed. Ozzed.net - 8-bit and chiptune music for everyone, 2020.

[59] Zizi Papacharissi. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283, 2004.

[60] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[61] Richard E Petty and John T Cacioppo. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer, 1986.

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

[63] Pew Research Center. More now say it's 'stressful' to discuss politics with people they disagree with. 2018.

[64] Pew Research Center. Social media fact sheet. 2018.

[65] Emily Pronin, Thomas Gilovich, and Lee Ross. Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychological review*, 111(3):781, 2004.

[66] Hunter Railey and Jan Brennan. 50 state comparison: Civic education. *Education Commission of the States*, 2016.

[67] R Revlin, V Leirer, H Yopp, and R Yopp. The belief-bias effect in formal reasoning: the influence of knowledge on logic. *Memory & cognition*, 8(6):584–592, 1980.

[68] Manoel Horta Ribeiro, Pedro H Calais, Virgílio AF Almeida, and Wagner Meira Jr. " everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation. *arXiv preprint arXiv:1706.05924*, 2017.

[69] Joshua Rottman, Deborah Kelemen, and Liane Young. Tainting the soul: Purity concerns predict moral judgments of suicide. *Cognition*, 130(2):217–226, 2014.

[70] Seth Schiesel. Former justice promotes web-based civics lessons, Jun 2008.

[71] Valerie J Shute. Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55(2):503–524, 2011.

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

[73] Craig Silverman and Jeremy Singer-Vine. Most americans who see fake news believe it, new survey says, Dec 2016.

[74] James H Snider. Democracy on-line. *The Futurist*, 28(5):15, 1994.

[75] Keith E Stanovich and Richard F West. Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89(2):342, 1997.

[76] Keith E Stanovich, Richard F West, and Maggie E Toplak. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264, 2013.

[77] Jennifer Stromer-Galley and Alexis Wichowski. Political discussion online. *The handbook of internet studies*, 11:168, 2011.

[78] SubspaceAudio. 512 sound effects (8-bit style), 2016.

[79] Viren Swami. Social psychological origins of conspiracy theories: the case of the jewish conspiracy theory in malaysia. *Frontiers in Psychology*, 3:280, 2012.

[80] Viren Swami, Ingo W Nader, Jakob Pietschnig, Stefan Stieger, Ulrich S Tran, and Martin Voracek. Personality and individual difference correlates of attitudes toward human rights and civil liberties. *Personality and Individual Differences*, 53(4):443–447, 2012.

[81] Viren Swami, Martin Voracek, Stefan Stieger, Ulrich S Tran, and Adrian Furnham. Analytic thinking reduces belief in conspiracy theories. *Cognition*, 133(3):572–585, 2014.

[82] Kathy Swan, Keith C Barton, Stephen Buckles, Flannery Burke, Jim Charkins, SG Grant, Susan W Hardwick, John Lee, Peter Levine, Meira Levinson, et al. The college, career, and civic life (c3) framework for social studies state standards: Guidance for enhancing the rigor of k-12 civics, economics, geography, and history. 2013.

[83] Valerie Thompson and Jonathan St BT Evans. Belief bias in informal reasoning. *Thinking & Reasoning*, 18(3):278–310, 2012.

[84] Vero Vanden Abeele, Lennart E Nacke, Elisa D Mekler, and Daniel Johnson. Design and preliminary validation of the player experience inventory. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pages 335–341, 2016.

[85] Friedrich Waismann. Verifiability. In *How I See Philosophy*, pages 39–66. Springer, 1968.

[86] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857. ACM, 2018.

[87] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository.*, 2016.

Appendix A

Modified Moral Foundations Theory Questionnaire

A.1 Part 1

When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale:

0 = not at all relevant (This consideration has nothing to do with my judgments of right and wrong)

1 = not very relevant

2 = slightly relevant

3 = somewhat relevant

4 = very relevant

5 = extremely relevant (This is one of the most important factors when I judge right and wrong)

1. Whether or not someone suffered emotionally
2. Whether or not some people were treated differently than others
3. Whether or not someone's action showed love for his or her country
4. Whether or not someone showed a lack of respect for authority
5. Whether or not someone violated standards of purity and decency
6. Whether or not someone was good at math
7. Whether or not someone cared for someone weak or vulnerable
8. Whether or not someone acted unfairly
9. Whether or not someone did something to betray his or her group
10. Whether or not someone followed to the traditions and unwritten rules of society

11. Whether or not someone did something disgusting
12. Whether or not someone was cruel
13. Whether or not someone was denied his or her rights
14. Whether or not someone showed a lack of loyalty
15. Whether or not an action caused chaos or disorder
16. Whether or not someone acted in a way that God would approve of

A.2 Part 2

Please read the following sentences and indicate your agreement or disagreement:

1. Compassion for those who are suffering is the most crucial virtue.
2. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
3. I am proud of my country's history.
4. Respect for authority is something all children need to learn.
5. People should not do things that are disgusting, even if no one is harmed.
6. It is better to do good than to do bad.
7. One of the worst things a person could do is hurt a defenseless animal.
8. Justice is the most important requirement for a society.
9. People should be loyal to their family members, even when they have done something wrong.
10. Men and women each have different roles to play in society.
11. I would call some acts wrong on the grounds that they are unnatural.
12. It can never be right to kill a human being.
13. I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.
14. It is more important to be a team player than to express oneself.
15. If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.
16. Waiting until you're married to have sex is an important and valuable virtue.

Appendix B

Skill Assessment Questions

Students were asked to answer the following six skill assessment questions during both the pre-test and the post-test. Correct answers are in bold.

B.1 Choosing the Most Productive Discourse Move

1. Your neighbor thinks that the school day should be extended from 8 hours to 10 hours. You disagree. Choose the best action:

    - Tell him about the studies that have shown that extending the school day reduces productivity because students are so tired.
    - Don't get into an argument. You'll never change their mind anyway.
    - **Start by asking why he feels that way to get an idea of what they value.**
    - Explain that longer schooldays mean less time for extracurriculars.

2. Your neighbor thinks: "The Annual School Raffle should be banned. It feels like gambling, which is against school rules." Which response would be most persuasive?

    - The raffle raises lots of money every year for field trips.
    - Don't get into an argument. You'll never change their mind anyway.
    - **The raffle is a school tradition. We've been holding an annual raffle for over 100 years.**
    - The students learn a lot about budgeting and marketing from doing the raffle.

B.2 Identifying Values

3. Your neighbor thinks: "The Young Adventurer's Club has always been for boys. Allowing girls to join would be breaking a 100 years of tradition." Which response would be most persuasive?

    - Being involved in the Young Adventurer Club builds leadership skills and helps kids grow.
    - Don't get into an argument. You'll never change their mind anyway.

- **In the original Club Handbook, the club founder wrote, "All children deserve adventures."**
- That's just plain not fair. Girls can be adventurers just as much as boys can.

4. A group of kids were caught having a secret dance party in the church cemetery. Your neighbor thinks: "Nobody got hurt. No harm done." Which response would be most persuasive?

   - The cemetary is technically private property.
   - Don't get into an argument. You'll never change their mind anyway.
   - **The pastor was upset when he saw they had stepped on some of the tulips in his garden.**
   - The cemetary is a sacred place. There should be some kind of punishment for dancing on the graves of our ancestors.

B.3 Choosing High- over Low-Quality Arguments

5. Your neighbor thinks: "Students should be required to say the Pledge of Allegiance every day." Which response would be most persuasive?:

   - There's nothing in the school rules requiring students to say the Pledge of Allegiance every day.
   - Don't get into an argument. You'll never change his mind anyway.
   - Giving students the freedom to choose is more American.
   - **Studies show that having students exercise their first amendment rights increases their patriotism than reciting the pledge.**

B.4 Choosing Aligned over Unaligned Arguments

6. Your neighbor thinks: "I think students who fail tests should get free tutoring. We need to care about our weakest students more." Which response would be most persuasive?:

   - There has never been a tutoring program in the history of the school.
   - Don't get into an argument. You'll never change his mind anyway.
   - Research shows that giving unequal treatment to a small set of students can increase feelings of unfairness.
   - **Caring for our students also means letting them learn to deal with the consequences of their actions.**

Appendix C

Player Experience Questions

Students were asked to indicate the degree to which they agreed with the following statements:

Enjoyment: I enjoyed playing the game.

Meaning: Playing this game was valuable to me.

Immersion: I lost track of time while playing the game.

Curiosity: I wanted to find out how the game progressed.

Autonomy: I felt like I had choices regarding how I wanted to play the game.

Mastery: I felt a sense of mastery playing this game.

Progress Feedback: The game gave clear feedback on my progress towards the goals.

Clarity Of Goals: I grasped the overall goal of the game.

Ease Of Control: I quickly grasped how to perform in-game actions.

Challenge: The game was not too easy and not too hard to play.

Audiovisual Appeal: I liked the artistic design of the game.

Learning: I learned something new.

Longterm Change: Playing this game will change how I discuss politics in the future.

Appendix D

Discussion Questionnaire

When people disagreed, did they generally attack the person they disagreed with or the idea they disagreed with?

The person				The idea
1	2	3	4	5

When someone disagreed with me, I felt

Attacked personally				Not attacked personally
1	2	3	4	5

Did people focus more on finding the best solution to the problem, or winning the argument?

Winning the argument				Finding the best solution
1	2	3	4	5

Did most people participate, or did only a few people participate?

Most people				A few people
1	2	3	4	5

If someone was talking that I disagreed with, I

Listened to their perspective				Tuned them out
1	2	3	4	5

As I listened, I

Tried to understand both sides of the issue				Thought about how best to defend my position
1	2	3	4	5

When I hear good evidence that doesn't support my belief, I

Stick to what I believe				Change my mind
1	2	3	4	5