Towards Value-Adaptive Instruction: A Data-Driven Method for Addressing Bias in Argument Evaluation Tasks

Nicholas Diana
Carnegie Mellon University
Pittsburgh, USA
ndiana@cmu.edu

John Stamper
Carnegie Mellon University
Pittsburgh, USA
john@stamper.org

Kenneth Koedinger
Carnegie Mellon University
Pittsburgh, USA
koedinger@cmu.edu

ABSTRACT

As the media landscape is increasingly populated by less than reputable sources of information, educators have turned to argument evaluation training as a potential solution. Unfortunately, the bias literature suggests that our ability to objectively evaluate an argument is, to a large extent, determined by the relationship between our own beliefs and the beliefs latent in the argument we are evaluating. If the argument supports our worldview, we are much more likely to overlook logical errors. Teachers recognize this need to adapt argument evaluation instruction to the specific beliefs of students. For instance, a teacher might intentionally assign a student an argument that the student disagrees with. Unfortunately, this kind of value-adaptive instruction is infrequent due to its unscalability. We propose a novel method for data-driven value-adaptive instruction in instructional technologies. This method can be used to combat bias in real-world contexts and support human reasoning during media consumption.

Author Keywords

Educational Technology; Civic Education; Civic Technology; Adaptive Instruction; Human-Computer Interaction

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI); HCI design and evaluation methods;** •**Applied computing** → **Education;** *Computer-assisted instruction;*

INTRODUCTION

The ability to objectively evaluate arguments is an essential skill if one hopes to navigate a media landscape littered with misleading or patently false information. Particularly in recent years, a great deal of energy has been devoted to designing instruction with the explicit goal of making people more critical media consumers [13, 1]. And while technology may be commonly viewed as a contributor to many civic engagement challenges (e.g., information bubbles, unproductive discourse [4]), supporting human reasoning when it is most susceptible to bias may be one way that technology can play a positive role in overcoming these challenges.

A citizenry capable of objectively evaluating arguments in the media is a core goal of civic education that is emphasized throughout popular frameworks for Social Studies curricula, such as College, Career, and Civic Life Framework [28] and the 2011 Guardian of Democracy report from the Center for Information and Research on Civic Learning and Engagement (CIRCLE) [8]. For example, the CIRCLE report warns that "media coverage responds to consumer demand and can only be as good as the consumers it serves" (p. 12) [8], and makes the specific prescription 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" (p. 12) [8].

On its surface, the motivation behind this kind of argument evaluation training seems straightforward: If we give people practice thinking critically about arguments, when they encounter dubious arguments in the future, they will think more critically about them. While this may in fact be true in some cases, this straightforward view of real-world argument evaluation fails to explain phenomena like the *motivated numeracy effect*. In a series of experiments, Kahan and colleagues [16, 15] presented participants with problems that required them to critically evaluate quantitative data. They also measured the *numeracy* of each participant, which is a measure of not only the participant's mathematical skills, but also their tendency to "engage quantitative information in a reflective and systematic way and to use it to support valid inferences" [15]. If the "better critical thinkers = better media consumers" hypothesis is correct, then we would expect individuals with high numeracy to be relatively immune to the political valence of the scenario they are being asked to evaluate. In other words, we would expect that their quantitative training will help them evaluate evidence more objectively.

As we might expect, when given a problem scenario that was politically neutral (e.g., a new skin rash treatment), participants with higher numeracy (i.e., more likely to use quantitative data) did better than participants with lower numeracy. However, when given a politically charged issue (e.g., gun control), accuracy decreased, and participants with high numeracy showed evidence of *greater polarization* (i.e., less objectivity) than their peers with less numeracy. The authors hypothesize that, contrary to the "better critical thinkers = better media consumers" hypothesis, participants with high numeracy may be using their quantitative reasoning skills to further justify

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.3376469

their own beliefs rather than to objectively evaluate evidence that might run counter to their established worldview.

A related line of research pursued by Stanovich and colleagues has centered around our tendency to evaluate arguments more favorably when the argument aligns with our own beliefs or worldview. This phenomenon, termed *Myside Bias*, has been shown to have powerful and reliable effects on informal argument evaluation tasks, in which participants indicate that arguments that they agree with are stronger than arguments they disagree with. Moreover, myside bias has been shown to impact argument evaluation performance independent of the participant's intelligence.

Haidt's Social Intuitionist Model [9] provides a theoretical framework for understanding the mechanism of myside bias. He argues that when we see a politically charged argument, judgments of correctness/incorrectness or rightness/wrongness are made instantly and intuitively, using Kahneman's System 1 thinking. The System 2 thinking (which we would need to engage in order to critically evaluate the argument), is almost never activated if the argument aligns with our beliefs. If it feels true, we have little incentive to exert more cognitive effort to do something that risks undermining our worldview. However, when presented with an argument that conflicts with our beliefs or worldview, System 2 is activated, but not in an effort to seek the truth or evaluate evidence objectively. Instead, we activate System 2 in order to justify our own position or attack the rationale of the conflicting opinion.

These studies and frameworks paint a picture of a more nuanced relationship between training and argument evaluation. They suggest that addressing society's apparent inability to accurately evaluate arguments in news media [26] likely requires more than general instruction on argument evaluation. This is because the skill we refer to as *argument evaluation* is likely at least two separate skills:

1. The ability to evaluate an argument that *aligns* with your beliefs

2. The ability to evaluate an argument that *does not align* with your beliefs

A media-literate citizenry must be able to do both, but this requires instructional design that is able to make the distinction between these two very different kinds of argument evaluation tasks. In order to make this distinction however, we must first have 1) a working knowledge or estimate of student beliefs, 2) a working knowledge or estimate of the beliefs latent in the argument the student is evaluating, and 3) a way to estimate the alignment between those two sets of beliefs.

Value-Adaptive Instruction in the Classroom

It is important to recognize that teachers are experts at adapting instruction to the beliefs of a particular student. For example, in a recent series of semi-structured interviews, a high school Social Studies teacher described a particular lesson in which each student is matched to an attendee of the Constitutional Convention and asked to portray that historical figure during a mock convention in class. The teacher recounted one instance in which they matched a student with a specific historical figure, solely because they knew the student's beliefs conflicted with the beliefs of the person they would be asked to portray. Why go through all this trouble? Presumably, the teacher understood that there was some benefit to be gained from having a student defend (and perhaps empathize with) a perspective different than their own. This is value-adaptive instruction.

But value-adaptive instruction as it currently exists in the classroom is unscalable. Effective and efficient value-adaptive instruction would require that a teacher has a working knowledge of each of their students' beliefs, and the ability to individually administer, to each student, the specific type of argument that best supports their mastery of the above two distinct skills. This would be difficult in a class of more than a few students, and certainly is an entirely unreasonable request in a much more typical classroom of 20-30 students.

However, similar kinds of scalable, individualized, and adaptive instruction, have been achieved in other domains through the use of intelligent tutoring systems. For example, a recent large-scale study conducted by the RAND corporation found that adaptive tutoring systems, on average, improved student performance in algebra by about eight percentage points [23]. The implementation of a similar adaptive system in a college statistics course resulted in a 15% increase in learning rate compared to the traditional course [20]. Moreover, students showed these gains despite having spent roughly half the time in the adaptive course than their peers in the traditional classroom condition [20], demonstrating not only an increase in efficacy, but efficiency as well.

Unfortunately, there been far less work on the impact of such adaptive systems in the civic technology space. The work that exists, however, illustrates that specific aspects of this ill-defined domain [21] can be defined and measured in a way that makes them a tractable topic for instructional technologies. For example, philosophers have developed software tools for supporting the diagramming and evaluation of arguments [25, 11]. Additionally, researchers [1] and civics educators [8] have highlighted the usefulness of educational games as a scalable way to have students engage in "civic simulations" (e.g., voting, engaging in civil discourse, etc.), and researchers have developed such civic games to teach specific civic skills like media literacy [13] and perspective taking [5].

While these solutions demonstrate that technology can play a positive role in civic learning, the bias literature suggests that the effectiveness of this instruction will be limited by its failure to address the role of bias in argument evaluation. To effectively address and combat bias in civic learning, our educational technology needs to be capable of doing what good civics educators do intuitively: adapting instruction to the specific values of individual students.

The main contribution of this paper is a novel approach to implementing value-adaptive instruction in instructional technologies. Our approach is built on a novel method for capturing the relationship between a user's values and the values latent in text content. Capturing this relationship allows us to provide individualized instruction on specific argument

evaluation skills, including those required to combat bias. Furthermore, we can, for the first time, measure bias and the effects of debiasing interventions without any hand-coding of instructional content (i.e., in an entirely data-driven way). In sum, this paper presents a novel, scalable approach to value-adaptive instruction in educational technology.

The broad goals of this project are to target the civic education of both younger adults (in schools and online) and older adults (online). The current project focuses on online use. Online education is important in this domain because we ultimately want to support life-long learning of civics and media literacy skills. This is particularly important in this domain because research [17] (and the results of this study) indicate the older adults show stronger myside bias effects and thus are most in need of educational interventions designed to reduce such bias.

What follows is a discussion of the potential benefits of value-adaptive instruction in education, and a theory- and data-driven method for adapting instruction to user values. Finally, we will briefly discuss some recent success implementing this methodology, as well as future directions and applications that might benefit from value-adaptive instruction. We believe that value-adaptive instruction is not only a useful method in academic contexts, but can also support life-long learning systems where technology is used to support human-reasoning.

AFFORDANCES OF VALUE-ADAPTIVE INSTRUCTION

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

This example illustrates the first benefit of value-sensitive instruction: **targeted justification requests**. Asking a student to justify *a belief they agree with* requires that they engage System 2 thinking. Moreover, Haidt [9] 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 of civil discourse like Constructive Conflict Theory [14].

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

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

As we've discussed above, when our *unintelligent* tutoring system asks students to justify arguments, it cannot distinguish between arguments that align with the student's beliefs and arguments that do not. We've noted that being able to know when a student is justifying their own beliefs is essential for measuring their ability to use evidence to support their arguments. However, it is equally, if not more important, to know when student is justifying beliefs that are not their own. This requires students to engage in the second key benefit of value-sensitive instruction: **targeted perspective taking**. Perspective taking another core skill in civic learning that is present in both the C3 and CIRCLE civic education standards we've referenced above [8, 28].

Myside Bias as a Civil Discourse Difficulty Factor

The third, and primary benefit of value-sensitive instruction is its potential ability to **measure myside bias**. Recall that myside bias is the formal name for our tendency to evaluate arguments more favorably when they align with your own views or beliefs (and conversely, more critically when they do not) [27]. Information 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 illustrating the potential impact of myside bias on civil discourse, and how value-sensitive systems might help mitigate myside bias in reasoning about news media specifically, and in civil discourse more broadly.

A METHOD FOR ADAPTING INSTRUCTION TO VALUES

In its simplest form, value-adaptive instruction can be thought of as a way to measure and adapt to the relationship between

the values of the user and the values latent in the content they are reading or listening to. We call the degree to which a user's values align with the values latent in the content *alignment*. *Alignment* is not a new construct in psychology, as most myside bias experiments require either that an identity or belief be measured (or assumed) and then related to experimental content designed to align with or oppose that belief. What we propose here is a theory-driven method for computing *alignment* between a user's values and *any unlabeled text*. Unlike most myside bias studies, our resulting metric is continuous (as opposed to binary) and multi-dimensional (as opposed to one-dimensional). The method described below is, of course, only one potential method for computing alignment, and while it boasts explainability and theoretical-grounding, there are doubtless other superior methods for computing *alignment* that have yet to be discovered.

Computing the degree of *alignment* between user- and content-values first requires that we estimate those each of those sets of values. We estimate user-values using a theory-driven approach that draws on Haidt's Moral Foundations Theory [9]. Then, we estimate the values latent in text content using Garten's Distributed Disctionary Representations [7]. Finally, we relate these two vectors of values to one another to compute *alignment*. What follows are justifications and detailed descriptions of each of these steps.

Estimating User-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.

Moral Foundations Theory

Moral Foundations Theory [10] 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 [2], which are short scenarios designed to evoke a specific moral foundation. See Table 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 when making moral decisions. 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 [9]. 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 across these groups are just manifestations of more fundamental differences 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 perhaps 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 rational are most commonly challenged by another person (who usually disagrees with you). Perspective taking, like the kind supported by value-adaptive instruction, allows us to more accurately simulate the

[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.

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 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 [6] as well as more specific political beliefs (e.g., "Climate change is real") [18, 24]. 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.

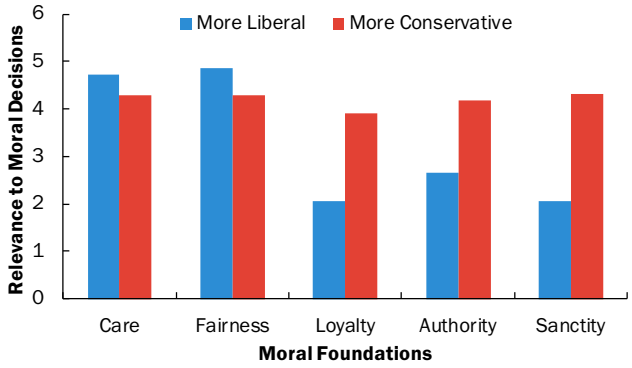

Figure 1. Relevance to Moral Decisions by Moral Foundation for more conservative and more liberal participants in one of our studies. These values closely match previously observed values for liberals and conservatives [10], suggesting that our recruitment pool is politically diverse.

Estimating Content-Values

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

First, this approach is only effective for analyzing large bodies of text. This is because smaller bodies of text (e.g., news headlines, tweets, etc.) may be highly relevant to the target 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) [3]. 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 [22] 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 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 [12], recent implementations (such as the Word2Vec [22] methodology employed in the current study) have demonstrated the effectiveness and efficiency of the method (in terms of computational cost). Mikolov et al. [22] 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.

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 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 [2]).

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 kinds of semantic 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 [7] extended this work in distributed representations to incorporate concept dictionaries. A distributed dictionary representation is computed by simply averaging the distributed representations of all the words in the dictionary. The result is a point in the semantic space that amplifies the shared, core features of each of the component dictionary terms. Because we are ultimately using an abstract representation of a concept, our dictionaries can be highly focused, including only the most relevant terms. Distributed dictionary representations mitigate the two major drawbacks of word-frequency methods. First, because the method calculates the semantic distance between the body of text and the target concept, it does not require that any of the dictionary terms actually be present in the text. This allows for the effective analysis of small bodies of text. Second, because the distributed representations are built using a text-corpus, the estimated meaning of words will be true to the words' contemporary meaning, so long as the text-corpus is contemporary. In our analysis, we use the pre-trained Google News corpus (approximately 100 billion words) Word2Vec model[2], and a Python implementation of Word2Vec [22] called *gensim*. We also use the same concept dictionaries that Garten and colleagues had used in their original paper [7].

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, represented below in Equation 1 as vector *x*.

$$x = [Care_u, Fairness_u, Loyalty_u, Authority_u, Sanctity_u] \quad (1)$$

After we have this estimation of student values, we compute the similarity ($1 - CosineDistance$) 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. This vector is represented below in Equation 2 as vector *y*.

$$y = [Care_t, Fairness_t, Loyalty_t, Authority_t, Sanctity_t] \quad (2)$$

To generate an *Alignment* score, we first compute the similarity ($1 - CosineDistance$) between the student's vector of values and the text's vector of values. The resulting score is a number from 0 to 2 that reflects the extent to which the student and text values align. Finally, we use a normalized log-transformation to correct for skew (see Equation 3).

[2]The pre-trained Google News model can be found here: https://code.google.com/p/word2vec/

$$alignment = \log(1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||_2 \cdot ||\boldsymbol{y}||_2}) \quad (3)$$

Alignment should be computed for each student/content combination. Computing alignment can be done in real-time, but is computationally demanding enough to warrant a dedicated server. Alternatively, because the bulk of the computational load is due to the distributed dictionary analysis, in cases of static, pre-determined text content (i.e., tutoring systems), the DDR analysis can be done beforehand. Scoring the Moral Foundations Theory Quesitonnaire and calculating the cosine distance between those scores and the pre-computed text-content scores is computationally negligible, and can easily be handled on the client-side of any modern internet browser.

It is worth reiterating the method we outline here is only one potential way to measure the construct of *alignment*. Any piece of this equation (i.e., the estimation of user values, the estimation of values latent in text, or the method for relating those two estimations) can be replaced or iterated upon to improve the construct's validity. What follows is a case study of an experiment we conducted that used the above method to compute *alignment*. We offer this case study in both an effort to make the particulars of the method more concrete, but also in an effort to demonstrate that, while our estimation of *alignment* is surely not perfect, it is accurate enough to be useful in some contexts.

CASE STUDY: 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*. Of the 60 participants, 38 identified as male, 20 as female, and 2 as other. Participants ranged in age from 18-62 years old (M=31.10). With respect to race and ethnicity, 50 participants identified as Caucasian, 6 as Hispanic, 3 as Black, and 1 as Asian. The majority (59%) of participants reported having completed a college level education or higher, and a high number of participants reported completing a master's degree (n=19). While Moral Foundations Theory claims that moral intuitions are innate and thus presumably present (in some form) throughout childhood, engaging with those intuitions in a meaningful way requires a level of intellectual maturity that begins to develop in young adulthood. As such, interventions that incorporate this methodology would likely be most effective if they target learners at the high school level and beyond.

Given the wide age range of our target learners (>= 17 years old), we believe the sample used in the case study presented above is appropriate. These participants were recruited from the general population with the restriction that they reside in the United States (this condition presents its own limitations, which are discussed above). In the assessment of value-adaptive systems, recruiting from an online participant pool allowed us to access a more politically diverse sample of participants than recruiting from our local community.

Participants were 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 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. The value-aligned arguments used in this study were based on concepts used in the empirically validated Moral Foundations Vignettes [2]. 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):

$$rating \sim quality + alignment + (1|participantID) + (1|argumentID) \quad (4)$$

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* ($\beta = 1.33$, $p < 0.01$).

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 [17]. 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 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).

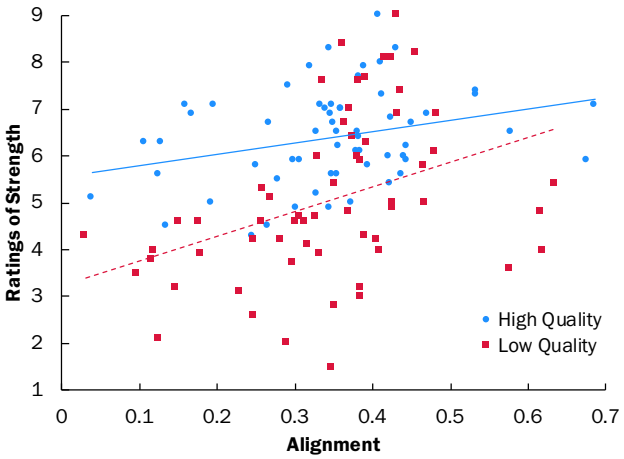

Figure 2. Relative impact of alignment on the ratings of high and low quality arguments. Each data point represents the average rating and alignment for all arguments within a category (high or low quality) for one participant. On average, participants rated high quality arguments as stronger than low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

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 3).

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

LIMITATIONS

It is important to note that while the model specified in the case study is only one instance of how this method might be implemented, the specified model has several important limitations that future work should aim to address. First, the reliance on an English news corpus would likely limit the

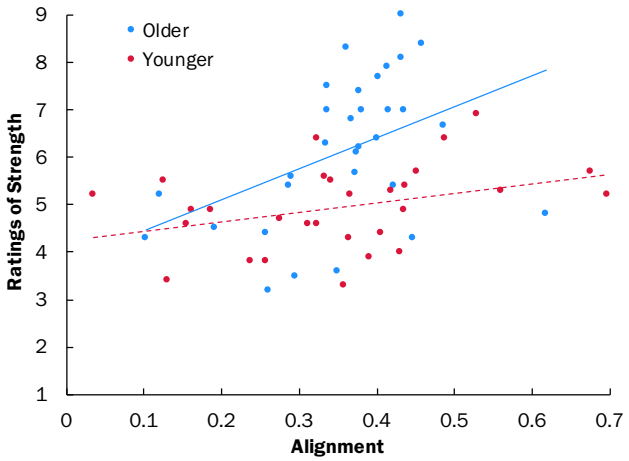

Figure 3. The interaction between age and alignment. Each data point represents one participant's average rating and alignment scores. Alignment had a much larger impact on ratings of strength for older participants (participants above the median age) than younger participants. This conforms with previous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

applicability of this work to learners for whom English is their primary language. A central claim of Moral Foundations Theory is that the foundations are universal, but expanding this work to non-English speakers would require the construction and validation of new concept dictionaries for each foundation.

It is also important to recognize that any technology built by and dependent on data generated by biased human beings will inherently contain bias. The consideration and acknowledgment of algorithm bias is particularly important in the development of technological interventions designed to reduce human bias. These systems likely require a perception of relative neutrality in order to be effective, and as technology can never truly be value-neutral, honesty about the bias inherent in such systems is critical for building user trust.

While we demonstrate that this particular measure of alignment can be used to successfully predict bias in the general population, how well this method generalizes to a traditional classroom remains an open question. We would expect young adults (like their older counterparts) to similarly draw on their intuitive, System 1 thinking to make moral decisions, and thus be susceptible to myside bias. However, the known interaction between age and myside bias may alter the efficacy of debiasing interventions. For example, it may be the case that debiasing interventions are more effective in classroom contexts due to young adults potentially being more receptive to new ideas. In contrast, it may be the case that debiasing interventions are less effective in classroom contexts due to myside bias in young adults being less developed or pronounced. Examining the impact of age-related effects on the efficacy of this methodology is the subject of future work.

FUTURE DIRECTIONS AND APPLICATIONS

As we've discussed above, value-adaptive instruction is useful for distinguishing between two key civil discourse skills: 1) justifying an argument you agree with and 2) justifying an argument that you disagree with (i.e., perspective taking). But we've also discussed that there is a more fundamental benefit afforded by value-adaptive instruction: the acknowledgement and measurement of the role of bias in our evaluation of politically charged arguments. Measuring the impact of bias on critical reasoning requires that we have some estimation of user beliefs, as well as the beliefs latent in the argument the user is reasoning about. Adapting instruction in this way is too computationally expensive for a human instructor in class sizes larger than a few students. However, by using theory-driven estimates of user values and data-driven estimates of the values latent in text content, we can equip educational technology with the ability to adapt instruction based on user values. Moreover, our estimates of user-beliefs are topic-general, which allows us to apply them to any context. This, along with the data-driven nature of our content-value estimations, makes our method scalable to classrooms of any size, and theoretically applicable to even future, novel political contexts.

We propose that this value-adaptive methodology can be used to address the following gaps in the learning science, civic education, and media literacy spaces:

Long-Lasting Debiasing Effects

Developing successful debiasing interventions is notoriously difficult [19], and even when promising results are reported, whether the benefits of the intervention have long-lasting effects is rarely examined. Lilenfield and colleagues [19] use the metaphor of bias as a chronic disease, and suggest that, like a disease, bias requires continual treatment over time. We believe that this data-driven approach to measuring the impact of bias in a context-independent way can support the development of the kinds of debiasing ecosystems necessary to continually combat the biases that afflict our ability reason objectively.

Destabilizing Information Bubbles

From our theoretical perspective (based on the Social Intuitionist Model), the goal of an information bubble is to maximize the "this feels true," System 1 thinking, and conversely minimize the more rational and critical, System 2 thinking (which risks "popping" the bubble). Information bubbles benefit from three unfortunate facts: 1) engaging in critical (System 2) thinking is generally more effortful (cognitively demanding), 2) if we agree with the belief in question, we have an incentive to confirm our own perception of reality, and 3) even if we disagree with the belief in question, there is still often a social risk associated with standing in opposition to the group and being seen as not-a-team-player [16]. These powerful cognitive and social forces do not simply make truth-seeking difficult, they inhibit the activation of the System 2 thinking necessary for truth-seeking. A key strength of technology in this space is that technology is (largely) unaffected by these forces, and could continue to probe the user to think critically precisely in the moments when these social or self-preservation forces might ordinarily prevent them from doing so. In this way, technology is a partner (or perhaps sidekick) in the reasoning process – providing us another set of eyes when we might otherwise be blinded by bias.

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 given to algorithms designed to maximize screen time (rather than a diversity of perspectives or journalistic integrity). While there may be a desire for more-trustworthy news sources, there must also be a recognition that our biases limit the reliability of our media-consumption habits. We believe that this value-adaptive methodology can support new, user-centered approaches to augmenting media consumption. By making the relationship between a user's values and the values latent in their media more explicit, we create an media consumption environment that is more transparent than the one curated by black-box algorithms. Moreover, this transparency would make it easier for users to identify (and hopefully empathize with) the values that motivate an opposing perspective. At the very least, this approach empowers users to make more informed judgments about the kind of media they are consuming.

Each of these future directions center around life-long learning in real-contexts and rely on the computationally expensive process of estimating user beliefs and then relating those beliefs to the content they are consuming. Until now, this kind of individualized, value-adaptive instruction has been impossible in contexts of more than a few students. The main contribution of this paper is a novel, scalable approach for value-adaptive instruction that can be integrated into instructional tools in a variety of contexts. We believe that capturing this relationship between user- and content-values not only helps us discriminate skills in domains like argument evaluation, but it also helps us measure and combat the biases that interfere with our objectivity and make it difficult to engage in productive civil discourse.

ACKNOWLEDGEMENTS

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

REFERENCES

[1] Elizabeth AS Bagley and David Williamson Shaffer. 2011. Promoting civic thinking through epistemic game play. In *Discoveries in gaming and computer-mediated simulations: New interdisciplinary applications*. IGI Global, 111–127.

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

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

[4] Maeve Duggan and Aaron Smith. 2016. The Political Environment on Social Media. *Pew Research Center* (25 10 2016). https://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

[5] Matthew W Easterday, Yanna Krupnikov, Colin Fitzpatrick, Salwa Barhumi, and Alexis Hope. 2019. Political Agenda: Designing a Cognitive Game for Political Perspective Taking. In *Civic Engagement and Politics: Concepts, Methodologies, Tools, and Applications*. IGI Global, 361–390.

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

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

[8] Jonathan Gould, Kathleen Hall Jamieson, Peter Levine, Ted McConnell, and David B Smith. 2011. 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).

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

[10] Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20, 1 (2007), 98–116. DOI: http://dx.doi.org/10.1007/s11211-007-0034-z

[11] Maralee Harrell. 2007. Using argument diagramming software to teach critical thinking skills. In *Proceedings of the 5th international conference on education and information systems, technologies and applications*. Citeseer.

[12] Geoffrey E Hinton. 1984. Distributed representations. (1984).

[13] Bob Hone, Joyce Rice, Chas Brown, and Maggie Farley. 2018. Factitious. (2018). factitious.augamestudio.com

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

[15] Dan M Kahan, Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. Motivated numeracy and enlightened self-government. *Behavioural Public Policy* 1, 1 (2017), 54–86.

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

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

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

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

[20] Marsha Lovett, Oded Meyer, and Candace Thille. 2008. The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. *Journal of Interactive Media in Education* (2008).

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

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

[23] John F Pane, Beth Ann Griffin, Daniel F McCaffrey, and Rita Karam. 2014. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis* 36, 2 (2014), 127–144.

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

[25] Glenn Rowe, Fabrizio Macagno, Chris Reed, and Douglas Walton. 2006. Araucaria as a tool for diagramming arguments in teaching and studying philosophy. *Teaching Philosophy* 29, 2 (2006), 111–124.

[26] Craig Silverman and Jeremy Singer-Vine. 2016. Most Americans who see fake news believe it, new survey says. (Dec 2016). `http://www.buzzfeed.com/craigsilverman/fake-news-survey`

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

[28] Kathy Swan, Keith C Barton, Stephen Buckles, Flannery Burke, Jim Charkins, SG Grant, Susan W Hardwick, John Lee, Peter Levine, Meira Levinson, and others. 2013. 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).