Trust and explainability as tools for improving automated writing evaluation

AWE tools: Improvement through trust research

Maria Goldshtein

Human Systems Engineering, Arizona State University, maria.goldshtein@asu.edu

Erin K. Chiou

Human Systems Engineering, Arizona State University, erin.chiou@asu.edu

Amin G. Alhashim

Human Systems Engineering, Arizona State University, amin.alhashim@asu.edu

Rod D. Roscoe

Human Systems Engineering, Arizona State University, rod.roscoe@asu.edu

Automated writing evaluation (AWE) promises faster, more consistent, and potentially less biased judgments than human evaluators. This form of automation has been aiding processes in education in American schools in recent years by providing evaluation and feedback. Despite the rise in use of AWE and its pedagogical importance, it has not been widely examined through a trust lens. The unique system structure involved in AWE use—including three main parties: teachers, students, and AWE tools—gives rise to unique goals and dynamics. This paper fuses concepts from human factors engineering, computer science, and social psychology to highlight the utility of trust-in-automation research in creating actionable insights that could aid with AWE design, implementation, and use.

We briefly review current trust-in-automation frameworks along with specific trust-related issues that may be unique to the parties involved in the use of education technologies such as AWE. Juxtaposing these trust frameworks and educational technologies (e.g., AWE), we outline actionable steps that can aid in AWE design, implementation, and continued use by considering relational trust factors of the human and automation components involved.

CCS CONCEPTS • Human-centered computing → Human computer interaction (HCI) →HCI theory, concepts and models

Additional Keywords and Phrases: Automated writing evaluation, Algorithmic bias, Education, Trust in Automation

1 INTRODUCTION

Modern automated writing evaluation (AWE) systems are computational tools that rely on machine learning processes to evaluate writing based on statistical generalizations derived from prior ratings and corpora (i.e., training data) (McNamara, Crossley, Roscoe, Allen, & Dai, 2015; Wilson & Roscoe, 2020). These algorithms attempt to differentiate between types and standards of writing using variables related to syntactic complexity, sophistication, cohesion, vocabulary, word frequency, and other variables detectable in text. AWE is already implemented in the education system in many states to enable evaluation and feedback for students (Feathers, 2019). The stakes for AWE outcomes are high. Writing assessments can determine students' class placement (e.g., remedial, introductory, or advanced composition courses) (Balinski & Sönmez 1999), high school graduation, university admission (Engler, 2021), and job attainment (Ajunwa et al., 2016; Cocchiara et al., 2016).

Trust is a crucial component in the successful design, implementation, and use of AWE in educational environments. Institutions and teachers need to trust AWE to incorporate them in evaluative and pedagogical processes. Students need to trust AWE tools and the feedback they provide and feel that their writing is being evaluated fairly and constructively. Trusted feedback is more likely to be used by students to improve their writing as intended (Roscoe et al., 2017). Effectively operationalizing and cultivating trust becomes more pressing as the use of AWE becomes more ubiquitous and common.

Understanding trust in human-automation interaction has been a popular framework for guiding the research and design of technology that involves a human in the loop. This research has traditionally been in industrial work settings (Sheridan, 2002). Although this framing has acknowledged the role of social and organizational factors that affect human-automation interactions, traditionally the base level scope of analysis for studying human-automation interactions has focused on a single human operator and the automated system. In the case of AWE tools used in the education system, the base level scope of analysis for a writing evaluation task is normally teacher and student, who have distinct but also interdependent roles and possibly differing goals and expectations with respect to the technology.

Applying a relational trust framework to the design and implementation of AWE tools allows for the consideration of potentially conflicting goals across the different stakeholders involved with the technology. The specific demands involved in designing and implementing an AWE tool first require an examination of the different relationships involved in AWE as a task. Clarifying these relationships can subsequently help with the characterization, cultivation and maintaining of trust in AWE tools. Specific topics of interest in exploring AWE implementation and use through the trust lens pertain to the different goals and metrics teachers and students may have for assessing AWE tools and maintaining trust in them. The linguistic nature of the automated evaluation raises unique trust-related concerns. Initial and continued trust in the system can pertain to teacher's and student's existing trust in the fairness with which the tool might assess their language use. This assessment by teacher and student can be affected by their previous experiences or information on linguistic bias

In the next sections we review existing trust in automation frameworks, outline the goals of the various parties involved in AWE processes (focusing on teachers and students), and discuss ways in which recent frameworks can advance the design, implementation, and use of AWE tools. We show that the systematic application of a "trusting automation" (Chiou & Lee, 2021) lens to AWE tool design, implementation and use can provide for a novel and valuable set of actionable insights that can improve all facets of AWE tools. Finally, we discuss the sources and expressions of mistrust and distrust in automation stances, which should play a role in designing strategies and behaviors within the AWE system.

2 TRUST IN AUTOMATION STATE OF THE ART: A STARTING POINT FOR TRUST AND AWE RESEARCH

Several frameworks have been proposed to explain trust in automation or trust in AI-enabled systems (e.g., "AI teammates") (Hoff & Bashir, 2015; Hou et al., 2021; Huang et al., 2021), including frameworks inspired by interpersonal trust research (Mayer et al., 1995). However, appropriate conceptualizations of human trust in automation need to be based on social psychological theory and must consider the possible differences between interpersonal trust and the concept of trust that guides our understanding of human-automation relationships (Lee & See, 2004; Madhavan, P., & Wiegmann, 2007). A more recent framework of trusting automation has been proposed that accounts for increasingly interactive, AI-enabled technologies that may also be increasingly capable at tasks that were originally thought to be human-only (Chiou & Lee, 2021). This framework specifies that increasingly interactive and capable technology may need to consider how to *sustain* a trusting relationship through the technology's responsivity. This framework which expands on the more traditional conceptualizations of trust in automation that are based on an information processing perspective and "calibrating trust" according to the automation's capabilities.

Trust in automation research has not traditionally focused on educational technology, a field that has its own unique set of goals, stakeholders, and metrics for success (Roscoe, Craig & Douglas., 2017). Educational technologies embedded in a broader educational system involve complex relationships that are intended to change with time, as students make progress in their learning and self-efficacy. Furthermore, having both teachers and students involved in the process of using educational technology like AWE tools means that both those parties need to have their own understanding of how the AWE functions. Teachers and students also separately assess the AWE's success in addressing their goals. For teachers, the goals relate to providing evaluation and feedback that are relevant to the pedagogical principles and materials the teacher employs. For students, the goals relate to providing actionable and clear feedback, or trustworthy evaluation in the case of the student. In the next section we break down relevant sources of trust and mistrust in AWE and ways in which these issues can be characterized and addressed through trust in automation research.

3 THE UTILITY OF TRUST FRAMEWORKS IN AWE DEVELOPMENT, IMPLEMENTATION, AND USE

The successful design, implementation, and use of AWE tools requires the trust of all parties involved in the process. In this section we apply concepts from information processing (Eisenberg et al., 2003; Loewenstein & Lerner, 2003; Wickens & Carswell, 2006) and relational (Chiou & Lee, 2021) trust frameworks to some of the main issues in AWE implementation and continued use.

Examining trusting and trust calibration, along with responsivity issues in AWE contexts, can contribute to the design of context-aware actions, along with the promotion of mutual trusting in educational technology in general.

3.1 Goal alignment in educational technology

As previously discussed, the typical system scope of AWE involves three parties: a teacher, a student, and AWE technology. This system structure is common in education technology but is less researched in other contexts of trust in automation where the interaction is between one human and an automated system. Students differ from workers, and teachers differ from employers or managers. The primary goal of technological intervention in education seems to relate to student empowerment and self-efficiency. In the workplace, on the other hand, the primary goal of technological intervention relates to performance in a specific task, or to focus on complex problems in team processes. This perspective of student empowerment might be a useful one to consider in workplace situations as well, when considering a human-centered approach to worker empowerment through AI-enabled systems. The automated feedback provided by AWE is meant to aid the student in improving their writing by providing an objective evaluation of their current writing abilities and (whether and) how it differs from the desired writing abilities (Blake, 2013; Chun et a., 2016; Shermis & Burstein, 2013; Stab & Gurevych, 2014). The evaluation standards are determined by educators, through the mediation of the programmers developing the AWE tool. The goals of the three parties involved in AWE systems are outlined in Table 1 below.

Table 1: An illustration of some goals different stakeholders have within an AWE system. Including the independent and shared goals of the three parties

Role	Independent Goals	Shared Goals
AWE tool	Cultivate and maintain trust of other parties Empower students via clear and self-efficient learning Respond to dynamic needs of human parties	Objective, consistent feedback Student progress in writing AWE tool responsivity
Student	Improve writing Minimize efforts	
Teacher	Ease workload Information on student progress Evaluation commensurable with their own	

AWE tools clearly do not have their desires and goals. We view the goals of the tool as a representation of the goals of the tool creators (Educational Technology companies, develoeprs), and the goals of institutions who acquire the tools. AWE tools have defined utilities, and their success in achieving them can be quantified. The goals of AWE tools include: empowering students through self-efficient learning processes. In addition, AWE tools can contribute to easing teachers' workload by serving as a proxy for the teacher in providing evaluation on student writing in a quick and consistent manner. AWE is meant to remove tasks from teachers' plates while still allowing them to provide students with consistent, objective, and reliable feedback, helping the students gauge and work on their own progress. The continued effectiveness of the AWE tool requires it to adapt to the needs of students and teachers. Those needs may vary and change.

The human parties in AWE contexts share the global goals of providing students with consistent, unbiased, pedagogically useful, and efficient feedback or evaluation of student writing. All parties also share the goal of feedback provided by AWE being comparable to the feedback provided by teachers. The feedback AWE provides needs to rely on an evaluation rubric comparable to ones used by teachers. An additional global goal of AWE is for feedback and evaluation to be less biased than teachers might be. Human bias, whether implicit or explicit (Greenwald & Krieger, 2006), can contribute to inconsistency in writing evaluation.

While the parties share several global goals, their goals may differ on a more local level. Students who have other responsibilities and school activities might want to minimize their efforts activities dedicated to writing. The students' desire to spend less time learning may interfere with the goals of teachers or technologies to maximize learning, goals that requires *more* effort from the students.

The goal of receiving unbiased feedback and evaluation is a global goal. However, this goal might be more salient and more local for students who have experienced bias within the education system, and within their everyday experiences outside school. This goal might also be salient for teachers who have experienced or witnessed bias. While AWE tools are not aware of student backgrounds, bias can still arise through linguistic cues (e.g., lexicon, syntactic constructions, etc.). People who are the targets of that bias can aware of being judged on their language use.

There are additional stakeholders involved in AWE systems in less direct ways. One such stakeholder involved in the acquisition and use of AWE tools is the educational institution acquiring the tool. Institutions have some unique goals in AWE (e.g., cost effectiveness, long-term solutions) and some goals that overlap with the other parties (e.g., educational benefits, efficiency). The companies and developers of AWE technologies are additional stakeholders who have their own sets of goals (e.g., profits, positive feedback, long-term contracts). Future research can include the goals of the latter stakeholders and the ways in which those shape the design and use of AWE. However, this paper focuses on the teacher, student and AWE tool.

3.2 Reasons for trusting

Trust can be defined as a willingness to be vulnerable to another. For automated systems to cultivate and maintain the trust of human parties, they need to display *trustworthiness:* reasons and signs for trust. A related concept to trustworthiness in education technologies is *credibility* (Schroeder, Chiou, & Craig, 2021). The evaluations and feedback produced by AWE, along with the tools themselves, must be perceived as credible by students and teachers for trust to occur and be maintained. Determining the credibility of the information we receive is relevant to all frameworks of communication, which includes communication between people along with human-machine frameworks.

The notion of credibility (Fogg & Tseng, 1999) has been described in literature on decision-making, and on technology as beliefs relating to the trustworthiness of an information source, and relates to public legitimacy (Renn & Levine, 1991). Credibility has also been described as a combining competence, trustworthiness, and goodwill (McCroskey & Teven, 1999), with the concept of trustworthiness consisting of character, sagacity, safety, and honesty. This conceptualization of credibility seems to overlap with the conceptualization of factors affecting perceived trustworthiness, which have been described as: ability, benevolence, and integrity (Mayer, Davis, & Schoorman, 1995). Decades of research on message credibility in the realm of communication (see Pornpitakpan, 2004 for review) have shown that the variables most relevant to the credibility of a message include source, message, channel, receiver, and destination. These credibility properties mean that determining the credibility of a message is a multi-factor calculation which includes the previous experiences of the message receiver, and the intended audience. The different factors involved in credibility tell us that attitudes that human parties hold towards AI, technology and AWE are relevant in designing trustworthy AWE tools, and in communication about their function and outputs. The mode of communication directed at the human parties is also important in determining the trustworthiness of the system (Burgoon et al., 2000; Ha et al., 2020). Thus, the credibility of the output produced by the AWE in the eyes of the human stakeholders is an important component in aiding their trust in the AWE.

Trustworthiness may have overlapping or separate requirements for different parties involved in AWE systems. Meaning that different parties may have different parameters used to determine credibility and trustworthiness. These parameters will depend on the expectations from them within the AWE system, and their own goals.

3.3 Maintaining trusting relationships

Maintaining trusting relationships is a process that requires systems to be *responsive* while also considering parties' goals. Without the possibility of maintaining trusting relationships with AWE tools and its outputs, teachers and students may not use the tool as intended, detracting from its utility. Thus, for AWE tools to achieve their goal within the AWE system, all human parties must be considered in the design and evaluation of such technologies. *Responsivity* is another concept we will utilize to discuss ways in which trust is cultivated and maintained within an AWE system. Responsivity is the system's ability to adjust and adapt to altered conditions and to continue functioning. In the case of AWE, responsivity can and should be operationalized in different ways to cater to different goals of the three parties.

We consider how processes of trust can be supported through automation responsivity using Chiou & Lee's (2021) framing of trust based on four dimensions: the decision situation, semiotics, interaction sequence, and strategy.

To establish the relevant context in which trusting behaviors manifest, it is helpful to assess the decision *situation* to assess whether and to what extent trust is involved during an interaction, and the perceived tradeoffs of that interaction. Additionally, we need to determine the relevant *semiotics*, that is that signals, signs, and symbols that affect trusting decisions and interaction outcomes. The *sequence* of interactions pertains to an assessment of how trust might evolve across repeating interactions and multiple different situations – and how patterns of interaction shape trusting. Lastly, *strategy* is operationalized as an understanding of how people and automation navigate decision situations; understanding strategy can help empower human parties and contribute to shaping patterns of interactions and outcomes. Below we outline the ways in which *situation, semiotics, sequence* and *strategy* might play out in contexts involving AWE systems.

Situation. The goal of AWE tools is to evaluate student writing in a way that leads to improvement in student writing. AWE tools are meant to work in tandem with teachers, providing teachers with an accurate assessment of student progress, an assessment equivalent to what the teacher would produce themselves. Identifying the various situations possible in AWE contexts first requires an assessment of the goal environment. For example, knowing when the AWE tool will be used, how the tool will be used, and what will be the weight and role of the tool's output in the pedagogical process might affect interactions with or surrounding the tool. This coordination occurs between teacher, student, and may be mediated or moderated by the AWE tool. For example, if a teacher were to implement an online AWE tool as part of a student's grade, versus as an optional tool for students to self-assess their own writing. In those situations, at-home students attending school remotely would be facing the option of using the tool versus not using the tool, the resulting decision matrix may look like figure 1 below (values are estimates for this thought experiment).

	Instructor uses AWE feedback to assign grade	AWE is used only for student self-assessment
Student use	[-1:+1], +1	+1, 0
Student does not use	-1, 0	0, -1

Figure 1: A decision matrix for two hypothesized AWE decision situations with student and teacher actions, costs, and benefits

With the values in the cells representing costs (-) and benefits (+) to the student and teacher (student value, instructor value) we can see that the first row, first column cell shows that the students' value will range from a cost to a benefit ([-1:+1]), depending on their writing performance as assessed by the AWE, whereas the teacher gains (+1) in this immediate turn given the benefits of relying on the AWE to assist with the task of grading. However, this value may transition to a cost for the instructor (-1) if the AWE is not functioning as intended and the instructor continues to rely on the AWE tool to complete the grading task. To give one more explanation, in the second row, second column cell the student has chosen not the submit their writing for evaluation by the AWE, but the AWE is only used for student self-assessment. This does not cost the student but also does not have any benefits (0) whereas for the instructor this could be seen as a cost (-1) of implementing the tool with no benefit to improvement in student writing because of the tool.

The situation structure may also clarify decisions to use the technology itself, for either the teacher or the student alone. In the example above, the AWE feedback being a component of the student grade will likely motivate more students to utilize the tool. Knowing that including AWE use in the grade will motivate students to utilize them, teachers will be more likely to include the AWE feedback as a part of the grade. In turn, the AWE feedback and its perceived benefit to student and teacher, along with its credibility, will contribute to continued use of the tool. Receiving feedback that is perceived as useful by student and teacher will contribute to a decision matrix in which all parties gain benefits from the tool use. Responsivity is an important feature in AWE decision situations. For example, a student can grasp the situation and determine their subsequent strategy through information exchanged during interactions (Engström et al., 2018; Picard & Friston, 2014). Actions which can perturb the environment and provoke a response from the student can generate new information (Flach et al., 2013; Sadigh et al., 2018). Specifically, after receiving feedback, there can be an option to provide the student further resources on the topic the student struggled with the most.

The resources can be more detailed information, connecting the student to a teacher who can explain the topic further, or providing relevant examples or exercises to strengthen specific skills.

The coordination of teacher and student goals can be used as a framework to assess the situation and to aid in the development of AWE tools (Wagner & Arkin, 2011). Aligning the trust of the teacher and student with the capabilities of the AWE is a required condition for a system's superior performance (Lee & See, 2004), but not a sufficient one (McGuirl & Sarter, 2006; Merritt et al., 2015; Zhang et al., 2020). Another way to operationalize responsivity in AWE decision situations is to provide for the ability of feedback on the tool's performance to flow from the bottom to the top. In this case, responsivity would mean that students and teachers who notice problematic evaluation behaviors (e.g., inconsistency, bias, unclear outputs) could pass this information to developer. Those developer can then address these concerns. Responsivity can serve as another index of trustworthiness and help maintain the trusting relationship between the parties involved in the system, while supporting the system's resilience.

Semiotics pertains to signs and how those are interpreted. The De Saussurian (2011, initially 1916) notion of a dyadic sign includes the signifier (the perceived form) and signified (the concept it denotes), Peirce (1974) turned the sign relationship to a triadic one, adding the *interpretant,* the person perceiving the communicated signifier and interpretating its meaning. In the context of a relational trust framework, signs of trust can be symbols (e.g., emblems, badges, certifications) and names, icons and indexes, or symptoms and signals – incidental indicators of trust (Riegelsberger et al., 2005). Within the context of trusting AWE systems, signs for trust can occur between the AWE tool and the human parties, combined or separately – in ways that pertain to their personal goals. AWE tools can have certifications relating to their performance, which may have a positive effect on trust from teachers. AWE tools can display symptoms cultivating trust in students, like outputs that are clear, relevant, and actionable in ways that minimize the student's efforts. The system's responsivity can be manifested through signs by being more interactive. An example of responsivity through signs can be providing students with personalized feedback pertaining to improvement on issues with which they struggled or reminding them of those issues are still are source of struggle. A personalized approach can signal to a student that the system is catering to their personal needs, and that the feedback received is specific to their educational goals. Thus, following the AWE system's advice can be perceived as a less effortful way to achieve progress and good grades.

Signs of trust can also occur between the human parties in an AWE system. The mere fact of teacher trust in the AWE tool can signal to the student that it is a useful tool. Further work on AWE situations should give rise to additional relevant signs within the semiotics of the AWE system. Another way in which semiotics can help maintain trust is through transparency, AWE tools providing detailed information on the reasons for the evaluations given (e.g., parts of the essay related with a certain bit of feedback).

The *sequence* of interactions pertains to how trust evolves over time, depending on how situations change, and how patterns of interaction shape trusting. Sequence relates how interactions between two or more parties in a system are timed and ordered. In the case of AWE systems decisions are made asynchronously, students write essays, their writing gets assessed by AWE tools, output with evaluation and feedback is disseminated to students and teachers (separately, or not), and then teachers and students make decisions on their next steps – based on the received output. The next steps can be coordinated between teacher and student but can also occur in each party separately. The decisions made by teachers and students may rely on past actions and signs, as those decisions are made asynchronously. The first interaction with the AWE tool also might have a heavier weight on teacher and student decisions, compared to later interactions (Chiou & Lee, 2021). The initial interaction will provide student and teacher with actual information about how the system works and the costs and benefits of using it. This will be the anchor for the perception of the system and trust in it, which later interactions will update.

Strategy describes a series of goal-oriented actions, norms, or policies that guide interactions. In AWE focused situations, strategies might include changing the outcome values of a situation to test an agent's trustworthiness or leveraging factors in the goal environment to create a new situation. For example, the AWE tool can attempt giving students different advice that targets the same writing related behaviors and gauge whether the student incorporates that feedback in future writing samples. This allows the AWE tool to gauge which feedback approaches are more useful for each student. As illustrated in table 1, teachers and students require overlapping but non-identical outputs from the AWE, thus the strategies for maintaining trust in providing those outputs would need to cater to some separate teacher and student needs. The AWE can interact with the teacher about student needs and progress through different strategies. One example could be providing the teacher with information on areas that require further classroom instruction from the teacher. This would allow the teacher to tailor the teaching materials to trends in student needs arising from the AWE output. Alternatively, similar goals can be achieved by providing the teachers with specific advice catered to the students' needs. In that case, the teachers can just guide the students to further interact with the AWE tool and follow the tool's personalized advice to further their writing goals. For the student, the AWE outputs should include feedback that can lead to actionable processes of learning, and feedback on progress over time. The explainability (Miller, 2019; Philips et al., 2020) and credibility of the outputs of an AWE tool in the eyes of teachers should play a role in AWE strategies, to help in cultivating and maintaining trust in the AWE system. Teachers' trust in the AWE tool may serve as an index (Chiou & Lee, 2021) of trust that students can use for credibility. Students need to view the outputs of AWE tools as credible to have trust in the feedback they receive from that tool. Importantly, strategies that focus on student or teacher goals separately are ultimately mutually beneficial for all parties involved in the triadic AWE system.

4 UNIQUE TRUST CONCERNS FOR AWE

In this section we delve further into the unique structure of AWE systems, mistrust and distrust in AWE, and language attitudes that lead to different trust stances. We define *distrust* as an informed lack of trust in the system based on relevant information or negative prior interactions within it. *Mistrust* refers to false trust or false distrust (Itoh & Tanaka, 2000) that emerges from inaccurate or irrelevant information. An additional relevant state of trust is low trust, which can be understood as an agnostic attitude.

4.1 Reasons for negative mistrust: language attitudes

AWE tools rely on written language data from training and operation; this raises distinct trust-related concerns. AWE tools use training data that rely on human raters' evaluations of writing. This means that raters' biases, or biases in the rating criteria, will be imported into the AWE tool and replicated by it if these biases are sufficiently consistent. Rater biases includes language attitudes (Eckert, 2008; Garrett et al., 2003; Konovalova & Le Mens, 2020; Lev-Ari & Keysar, 2010; Woolard & Schieffelin, 1994), specifically the perception of Standard American English as the preferred dialect, with certain varieties being judged more harshly (e.g., African American English, Southern dialects of English) (Lippi-Green, 2012). Language attitudes can be positive, negative, or neutral. When evaluating how people speak or write, audiences may infer or imagine the writer's age, gender, race, ethnicity, origin, socioeconomic status, or even personality traits (see Johnson et al., 2017). These language-based judgements can then influence the evaluation of writing. Speakers of commonly disparaged dialects might justifiably worry that AWE technologies will recreate these everyday linguistic biases.

Given the inherent potential for bias in AWE based in machine-learning, an additional benefit of responsivity relates to its ability to contribute to the *dismantling* of unjust perceptions and their perpetuation within automated systems. Specifically, feedback from teachers and students can contribute to a continued process of de-basing in automated tools, assisted by those most affected by bias. Ways to incorporate this feedback into the design and daily operation of AWE would support the ability of such tools to improve their responsivity to its human counterparts, and to evolve continuously with society.

4.2 Pre-existing trust and mistrust in automation and specifically AWE

People are not blank slates and may harbor negative attitudes towards AI and automation prior to their interaction. Understanding existing issues of distrust and mistrust can aid in developing the system's resilience. In the case of AWE systems, this would be the system's ability to maintain sustained engagement and adaptability in various environments.

Distrust and mistrust in technology are as common as trust (Dietvorst et al., 2015). Although many people are relatively technologically savvy, there is still a gap in the public's ability to understand the inner workings of the automated systems they interact with daily. One reason for this lack of understanding is the black box nature of many automated systems (Dubnick, 1998). In the case of AWE tools, users (e.g., institutions, teachers, and students) do not typically have access to information about the training datasets used AWE algorithms nor the generalizations underlying that development process (e.g., the driving assumptions of linear regression or machine learning analyses). Users are typically unaware of the internal "rubrics" that AWE technologies employ to evaluate writing. Other reasons for negative mistrust and distrust can relate to existing negative stances towards governments and institutions, or to negative or discriminatory personal experiences with such algorithms and the institutions that employ them.

It is worthwhile to disentangle two causes for mistrust in automation. First, negative mistrust in automated systems (including AWE) can stem from the perception of those technologies as flawed or prone to errors and biases. Importantly, this may be a stance supported by current events and contemporary evidence (Eubanks, 2018; Noble, 2018; O'Neil, 2016). Second, negative mistrust, and in some cases distrust, may reflect attitudes toward the developers and users of automated tools. To be clear, this position does not critique the technology itself as "untrustworthy," but instead questions the origins or intentions behind the tool (e.g., designed to cause harm or to benefit only people in power). Although both types of negative mistrust might result in similar responses to new technology, their roots are different and need to be identified and addressed in different ways.

Negative mistrust in technology (or the institutions that use the technology) and low trust can also emerge from our familiarity with cases of biased technology. For example, Buolamwini & Gebru (2018) discussed several different facial recognition algorithms that displayed large discrepancies in accuracy. Specifically, they reported that "darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%". These differences are impactful when such algorithms are applied by police departments and have indeed contributed to false arrests of misidentified people (e.g., Hill, 2020). Importantly, teachers and students who are aware of bias in automated systems via the news might be skeptical of automated systems using artificial intelligence. In addition, teachers and students belonging to minoritzed groups will have their own various experiences of bias to draw from, which can lead to mistrust in the system. Cases of familiarity with bias leading to lack of trust in AWE tools can be categorized as *distrust* if they are based on experiences with similar systems (whereas evidence that is less relevant or is anecdotal, would be categorized as *negative mistrust*).

On the other hand, some people have positive mistrust in technology. For this group of people automated tools that rely on technology are expected to produce reliable and trustworthy results (Sheridan, 2019). This attitude relies on perceiving technology as more objective and less likely to be biased, inconsistent or unfair than a human being performing the same task.

Both types of negative mistrust rely on insufficient information about the technology at hand, the social contexts within which the technology is operating, and potentially a lack of technological literacy. Initial attitudes towards AWE tools are informed by existing attitudes towards technology and societal structures, including the attitude holder's place within those structures. The difference in people's attitudes and base-level trust should be considered when designing strategies for cultivating and maintaining trust within the AWE system. Acknowledging and outlining these attitudes can contribute to the system's resilience through informed explanation (Miller, 2019), and other system features that address the potential trust situations.

5 CONCLUSION

This paper reviews the unique structure of AWE systems (i.e., a triad of teachers, students, and AWE technologies) and some unique societal and personal concerns can lead to teacher and student mistrust in the system. Addressing this mistrust necessitates carefully laying out the potential goals of different parties, and understanding which goals are shared and which goals might differ or even be contradictory. Understanding the different parties' goals and concerns becomes further complicated when we consider linguistic and social factors involved in AWE development and implementation.

In this brief paper we have reviewed the unique structure of AWE systems, the importance of examining trust in AWE contexts, and the specific ways in which trusting and trustworthiness manifest and should be maintained in those contexts. Through the application of information structure (Eisenberg et al., 2003; Loewenstein & Lerner, 2003; Wickens & Carswell, 2006) and relational (Chiou & Lee, 2021) frameworks, we might address unique issues that AWE entails.

ACKNOWLEDGMENTS

This research was supported by the Gates Foundation (INV-006213)

REFERENCES

Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 1–17. https://doi.org/10.1177/1461444816676645

Balinski, M., & Sönmez, T. (1999). A tale of two mechanisms: student placement. Journal of Economic theory, 84(1), 73-94.

Blake, R. J. (2013). Brave new digital classroom: Technology and foreign language learning. Georgetown University Press.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human–computer interaction: A study of credibility, understanding, and influence. Computers in human behavior, 16(6), 553-574.

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. Human Factors: The Journal of the Human Factors and Ergonomics Society

Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. The Modern Language Journal, 100(S1), 64-80.

Cocchiara, F. K., Bell, M. P., & Casper, W. J. (2016). Sounding "different": The role of sociolinguistic cues in evaluating job candidates. Human Resource Management, 55(3), 463-477.

De Saussure, F. (2011). Course in general linguistics. Columbia University Press.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114.

Dubnick, M. J. (1998). Clarifying accountability: An ethical theory framework. In N. Preston & C.-A. Bois (Eds.), Public Sector Ethics: Finding and Implementing Values (pp. 98–81). Sydney, Australia: The Federation Press.

Eckert, P. (2008). Variation and the indexical field 1. Journal of sociolinguistics, 12(4), 453-476.

Eisenberg, N., Losoya, S., & Spinrad, T. (2003). Affect and prosocial responding. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), Handbook of affective sciences (pp. 787–803). Oxford University Press.

Engler, 2021. Enrollment algorithms are contributing to the crises of higher education. Retrieved January 14th, 2022 from https://www.brookings.edu/research/enrollment-algorithms-are-contributing-to-the-crises-of-higher-education/

Engström, J., Bärgman, J., Nilsson, D., Seppelt, B., Markkula, G., Piccinini, G. B., & Victor, T. (2018). Great expectations: a predictive processing account of automobile driving. Theoretical issues in ergonomics science, 19(2), 156-194.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Feathers, 2019. Flawed Algorithms Are Grading Millions of Students' Essays. Retrieved January 22nd, 2022 from https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays

Flach, J. M., Bennett, K. B., Jagacinski, R. J., Mulder, M., & van Paassen, M. M. (2013). The closed-loop dynamics of cognitive work. The Oxford handbook of cognitive engineering, 1-18.

Fogg, B. J., & Tseng, H. (1999, May). The elements of computer credibility. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 80-87).

Garrett, P., Coupland, N., & Williams, A. (2003). Investigating language attitudes: Social meanings of dialect, ethnicity and performance. University of Wales Press.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. California law review, 94(4), 945-967.

Ha, T., Kim, S., Seo, D., & Lee, S. (2020). Effects of explanation types and perceived risk on trust in autonomous vehicles. Transportation research part F: traffic psychology and behaviour, 73, 271-280.

Hill, K. (2020, December 29th, updated January 6th, 2021). Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. The New York Times. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407–434. https://doi.org/10/f68kpx

Hou, M., Ho, G., & Dunwoody, D. (2021). IMPACTS: A trust model for human-autonomy teaming. Human-Intelligent Systems Integration. https://doi.org/10/gh5dx9

Huang, L., Cooke, N. J., Gutzwiller, R. S., Berman, S., Chiou, E. K., Demir, M., & Zhang, W. (2021). Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In C. S. Nam & J. B. Lyons (Eds.), Trust in Human-Robot Interaction (pp. 301–319). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00013-7

Itoh, M., & Tanaka, K. (2000). Mathematical modeling of trust in automation: Trust, distrust, and mistrust. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 44, No. 1, pp. 9-12). Sage CA: Los Angeles, CA: SAGE Publications.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 624-635).

Knowles, B., & Richards, J. T. (2021, March). The Sanction of Authority: Promoting Public Trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 262-271).

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. Psychological Review, 127(1), 47.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 31. https://doi.org/10.1518/hfes.46.1.50_30392

Lev Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. Journal of Experimental Social Psychology, 46(6), 1093.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), Handbook of affective sciences (pp. 619–642). Oxford University Press.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277–301. https://doi.org/10/d4sv4f

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. Communications Monographs, 66(1), 90-103.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35-59.

Lippi-Green, R. (2012). English with an accent: Language, ideology, and discrimination in the United States. Routledge.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. The Academy of Management Review, 20(3), 709. https://doi.org/10/fs6wzz

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human Factors: The Journal of the Human Factors and Ergonomics Society, 48, 656–665. https:// doi. org/ 10. 1518/ 001872006779166334

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K., & Louis, S. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57, 34–47. https:// doi. org/ 10. 1177/ 0018720814561675

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

National Academies of Sciences, Engineering, and Medicine. 2021. Human-AI Teaming: State of the Art and Research Needs. Washington, DC: The National Academies Press.https://doi.org/10.17226/26355.

Noble, S. U. (2018). Algorithms of oppression. New York University Press.

O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Peirce, C. S. (1974). Collected papers of charles sanders peirce (Vol. 5). Harvard University Press.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). Four principles of explainable artificial intelligence. Gaithersburg, Maryland.

Picard, F., & Friston, K. (2014). Predictions, perception, and a sense of self. Neurology, 83(12), 1112-1118.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. Journal of applied social psychology, 34(2), 243-281.

Renn, O., & Levine, D. (1991). Credibility and trust in risk communication. In Communicating risks to the public (pp. 175-217). Springer, Dordrecht.

Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. International Journal of Human-Computer Studies, 62(3), 381-422.

Roscoe, R. D., Craig, S. D., & Douglas, I. (Eds.). (2017). End-user considerations in educational technology design. IGI Global.

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. Computers in Human Behavior, 70, 207-221. https://doi.org/10.1016/j.chb.2016.12.076

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Sadigh, D., Landolfi, N., Sastry, S. S., Seshia, S. A., & Dragan, A. D. (2018). Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state. Autonomous Robots, 42(7), 1405-1426.

Schroeder, N. L., Chiou, E. K., & Craig, S. D. (2021). Trust influences perceptions of virtual humans, but not necessarily learning. Computers & Education, 160, 104039. https://doi.org/10/ghzb34

Sheridan, T. B. (2002). Humans and Automation: System Design and Research Issues. John Wiley & Sons, Inc.

Sheridan, T. B . (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. Frontiers in Psychology, 10, 1–7.doi:10.3389/fpsyg.2019.01117

Shermis, M. D., & Burstein, J. (2013). Handbook of automated essay evaluation. NY: Routledge.

Stab, C., & Gurevych, I. (2014, August). Annotating argument components and relations in persuasive essays. In Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers (pp. 1501-1510).

Stanton, B., & Jensen, T. (2021). NISTIR 8332: Trust and Artificial Intelligence. National Institute of Standards and Technology, U.S. Department of Commerce

Thornton, L., Knowles, B., & Blair, G. (2021, March). Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 64-76).

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & Van Moorsel, A. (2020, January). The relationship between trust in AI and trustworthy machine learning technologies. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 272-283).

Wagner, A. R., & Arkin, R. C. (2011, July). Recognizing situations that demand trust. In 2011 RO-MAN (pp. 7-14). IEEE.

Wickens, C. D., & Carswell, C. M. (2021). Information processing. Handbook of human factors and ergonomics, 114-158.

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. Journal of Educational Computing Research, 58(1), 87-125.

Woolard, K. A., & Schieffelin, B. B. (1994). Language ideology. Annual review of anthropology, 23(1), 55-82.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 295-305).