Say What? Learner Reactions to Unexpected Agent Dialogue Moves

Andrew J. Hampton, Jaclyn J. Gish-Lieberman, Jessica Gatewood, & Andrew A. Tawfik

For maximally efficient and effective conversation-based intelligent tutoring systems, designers must understand the expectations carried by their intended learners. Strategies programmed into the agents may be interpreted differently by the learners. For example, conversational heuristics in these systems may be biased against false alarms in identifying wrong answers (potentially accepting more incorrect answers), or they may avoid directly answering learner-generated questions in an attempt to encourage more open-ended input. Regardless of pedagogical merit, the learner may view these agents' dialogue moves as bugs rather than features and respond by disengaging or distrusting future interactions. We test this effect by orchestrating situations in agent-based instruction of electrical engineering topics (through an intelligent tutoring system called AutoTutor) where the pedagogical agent behaves in ways likely counter to learner expectations. To better understand the learning experience of the user, we then measure learner response via think-aloud protocol, eye-tracking, and direct interview. We find that, with few exceptions, learners do not reason that the actions are meant as instructional or technical strategies, but instead broadly understood as errors. This indicates a need for either alteration of agent dialogue strategies, or else additional (implicit or explicit) introduction of the strategies to productively shape learners' interactions with the system.

Introduction

In intelligent learning technology, the effectiveness of pedagogical interventions rests in part on establishing and leveraging learners' trust that the interventions are appropriate. As such, unexpected actions may be met with critical thinking, skepticism, confusion, or outright dismissal. Some of these outcomes may prove desirable—others disastrous. Natural language paradigms present additional challenges, with questions open to interpretation, answers defying singular expression, and opportunities for confusion at both the system and learner ends. Intelligent agents are one application of natural language instructional tools that can promote engagement and deep learning (Graesser et al., 2003). This paper explores some challenges a designer may confront related to effectiveness versus the substantial cost of natural language processing (NLP) interventions, with an emphasis on the learning experience as the focal point of design decisions (Tawfik, 2021).

The impact of conversation-based intelligent tutoring systems (ITS) on student learning has been subject to numerous and varied investigations. Researchers have explored the physical design of pedagogical agents to examine the difference between animated and human-like agents (Moreno et al., 2001), the embodiment of agents through the use of gestures, facial expressions, and eye movement (Li et al., 2019; Louwerse et al., 2009; Lusk & Atkinson, 2007), and the role of the agent's gender (Krämer et al., 2016). Others have focused on attributes ranging from voice quality (Craig & Schroeder, 2017) to personalization through levels of politeness (Wang et al., 2008) and rapport-building efforts (Krämer et al., 2016). These studies have expanded the knowledge base through specific guidelines for designing aesthetic and identity aspects of the artificial agents themselves.

As research and practice move deeper into dialogue spaces shared by agent and learner, investigation correspondingly expands into conditions that transcend the superficial aspects of design. The conversational actions of the agent—or dialogue moves—may impact the learner directly through metacognitive prompting (McCarthy et al., 2018; Wu & Looi, 2012) and indirectly through confusion (Lehman et al., 2012; Lehman et al., 2013). Lehman et al. (2013) ascribed confusion to "contradiction, anomaly, or system breakdown" engendered by the agent for the purposes of triggering cognitive disequilibrium within the learner (p. 86). Learners presented with agent-induced confusion did, indeed, show improved learning, prompted by the need to resolve that cognitive disequilibrium and create an internal model of the world that matches the information provided. However, that confusion in the learning environment must be appropriately regulated to produce learning gains, and these regulations still require examination and

cataloging (Lehman et al., 2012).

Meaningful learning in the face of deliberate confusion may be tempered by a learner's inherent trust in artificial intelligence. Pedagogical agents in learning environments are designed to provide a social presence to positively affect learning either as an emulation of a teacher or a co-learner (Chae et al., 2016; Lee et al., 2007). Researchers have studied determinants of learners' perception of trust of the agents, including the physical appearance of (Burgoon et al., 2016; Chae et al., 2016), emotional connection to (Savin-Baden et al., 2015), and the perception of caring from (Lee et al., 2007) the agent. The trust emanating from these factors appears to have a causal impact on learner participation intention, disclosure of information, and learning, respectively. Though these findings shape our understanding of how trust impacts communication and learning, they do not bear directly on pedagogical technique. A complex interaction of expected linguistic proficiency, perceived agent intention, and learner understanding of pedagogical technique likely has a significant impact on how contradictory or confusing agent dialogue moves impact perceptions of trust. Simply put, whether or not the learner knows what the agent is doing likely has an impact on communication patterns and subsequent learning outcomes, as much or more so than merely cosmetic agent characteristics. However, that interaction is not well understood.

AutoTutor

AutoTutor and its family of related systems (e.g., Graesser, 2016; Nye et al., 2014) provide an invaluable test bed for this inquiry. In this conversational intelligent tutoring system, (typically) adult learners encounter one or more talking head agents that present conceptual questions. The agents play the role of a tutor agent or a peer agent, often with both roles presented to afford complex interaction dynamics. Referring to a visual aid (e.g., a circuit diagram for electrical engineering problems), the agent(s) introduces a concept and then asks a question about it that requires multipart answers. The diagram contains hotspot-enabled "Point & Query" interaction for common questions (Graesser et al., 2018). For example, hovering over a critical component would trigger the appearance of questions like "What is this component?" and "What does this component do?", each of which containing a secondary hotspot with an answer. After the AutoTutor main question appears, learners attempt to answer via typed natural language input, interacting with and disengaging from the Point & Query as needed.

AutoTutor analyzes learner input on several factors. Based on a complete and correct answer provided and validated by several domain experts, the system

extracts knowledge components that form discrete parts of the complete answer. Each knowledge component is processed for conceptually linked alternative articulations via latent semantic analysis and given a degree of leeway in spelling and colloquialism via regular expressions. This expanded and flexible version of a systematically segmented answer forms the basis for comparison. If the learner provides a complete and correct answer (i.e., one that rests above a critical threshold based on latent semantic analysis and regular expression similarity to the ideal answer), the agents will acknowledge that with positive feedback, summarize the key points, and invite the learner to move on to another question. If the learner provides only part of a correct answer, the agents will encourage additional information or reasoning via one of three techniques. A hint will introduce a key concept that the learner omitted (e.g., “But what will happen to the voltage?”). A prompt encourages the inclusion of a single key content word from a missing knowledge component (e.g., “Increasing the resistance will decrease what?”). Finally, a pump provides generic encouragement to add more information (e.g., “Can you tell me anything else about it?”).

In combination, these AutoTutor responses handle a variety of learner input and can create a relatively detailed assessment of the learner’s understanding. It constitutes a “diagnostic” interactive learning resource based on its ability to discern conceptual understanding at gradient levels (i.e., requiring support to provide a full answer versus offering all parts on initial inquiry) (Hampton, 2019). This general approach has demonstrated learning gains in a wide range of fields (Nye et al., 2014), including reading comprehension and physics.

However, each of the interventions detailed above requires a degree of confidence in understanding unconstrained learner input. Absent perfect comprehension, imperfect heuristics must guide system response in the presence of heightened uncertainty. As far as possible, these heuristics should align with pedagogical goals. We next highlight several such heuristics that may be deployed in the event of uncertainty, that we categorize as edge cases.

Edge Cases

Acceptable Faults

As noted, the criteria for differentiating correct from incorrect input is mathematical. However, this discrete threshold masks a fuzzy distinction that questions when the system should act to resolve an incorrect answer. Following from statistics principles laid out by Neyman and Pearson (1967), there exist four possible outcomes in deciding whether or not to deploy corrective feedback: the

answer is correct and the system judges it as correct (correct rejection); the answer is wrong and the system judges it as wrong (hit); the answer is wrong but the system judges it as correct (miss); and the answer is correct by the system judges it as wrong (false alarm). Total proportions of hits and false alarms derive from accuracy of classification mechanisms within a system.

Beyond those limits, designers must choose whether false alarms or misses are the preferred outcome. Any threshold decision necessarily implies a value assignment between the two outcomes. Is it worse to be wrong and be told you are right, or to be right and told you are wrong? In AutoTutor, a relatively low threshold argues that the latter constitutes the worse outcome. Learner-generated answers that may be right (but probably are not) will be treated as right. AutoTutor avoids perpetuating misconceptions by reviewing the correct answer immediately after providing positive feedback. The low threshold value argues that this situation represents an acceptable fault relative to the alternative. In that alternative, a learner provides the right answer only to have the system provide corrective feedback, followed by a summary in which her original answer appears in paraphrase. This may well inspire unproductive confusion or, worse, distrust and disengagement that precludes further study. The heuristic, then, boils down to giving the learner the benefit of the doubt.

Question Rerouting

The trialogue paradigm that uses both tutor and peer agent may inspire a looser conversational dynamic than other natural language ITS approaches. As such, learners may feel inclined to pose their own questions for clarification or background. These questions likely reflect good-faith attempts to answer the main question. However, direct responses may prove suboptimal for several reasons. First, answering questions is technically demanding. Interrogatives entail different sentence structure comprehension to accurately parse and different response patterns to reply fluidly. Further, the open input mechanism does not integrate a marker to indicate the difference a priori, requiring a purpose-built function to identify interrogatives before parsing and responding. Though learners posing questions is certainly a possibility, it is not common practice in the conversational ITS paradigm. Therefore, these complex programming demands are unlikely to prove cost effective.

Second, the Point & Query system integrated into the visual aid should make learner-generated interrogatives largely redundant. The presence of that referential information should preclude more basic questions, as the system largely anticipates what those would be and makes the answers readily available. Third, pedagogically, the existing conversational and diagrammatic interface is

designed to provide all the information necessary to answer the question except for what must be provided by the learner. Any information or clarification provided only serves to confound evaluations that will inform the learner model and determine downstream learning activity.

Following these arguments, AutoTutor does not directly respond to learner questions. Instead, it parses questions the same as answers. Likely the questions include relevant content words but lack relational language that would rise above the threshold of correctness. As such, interrogatives would generally trigger responses similar to partially correct answers. These responses include hints, prompts, and pumps intended to encourage more complete answers, targeting specific knowledge components as deemed necessary. Essentially, the system answers learner questions with questions. This strategy fits well within the overall pedagogical approach. However, it does not account for how the learner interprets a non-answer to their direct question. Understanding how learners interpret these two heuristics will inform our understanding of the learner–system dynamic more generally and provide insight on how best to design conversational interactions for optimal learner outcomes.

Method

To better understand this dynamic, we orchestrated situations within a learning platform focused on electrical engineering topics with participants from our target learner demographic. By controlling the starting point and providing only general instructions rather than specific phrasing to be input, we balance the need for realistic testing circumstances and active engagement with the benefits of reasonable comparison across participants. An array of measures for learner response (think-aloud protocol, eye-tracking, and direct interview) attempted to gather as much meaningful information as possible.

Participants

We recruited participants from the electrical and computer engineering department of a large university in the Midsouth region of the United States. Participants had to have completed at least one course in electrical engineering to register. A total of nine students participated, of whom we eliminated two: one due to communication issues stemming from speaking English as a second language, and the second due to corruption of the audio file. This left seven participants. Though this constitutes a small *n*, we conceive of this inquiry primarily in terms of user experience optimization. In that experimental paradigm, as few as five participants may be considered sufficient to uncover a strong majority of exigent

design deficiencies (Nielsen & Loranger, 2006).

Materials

Participants interacted with a federated learning system consisting of several constituent learning resources. These included both intelligent and conventional materials, with adaptivity varying in type and degree across resources. The system, ElectronixTutor (Graesser et al., 2018; Hampton & Graesser, 2019), uses a single interface to present all resources. Of these resources, AutoTutor figures prominently and constituted the largest portion of testing. All learning content focused on electricity and electronics topics common to undergraduate university or basic military education in the area, derived primarily from the curriculum set forth in the Navy Electricity and Electronics Training Series (U.S. Navy, 1998). The Tobii Eye Tracking system provided unobtrusive attentional measurement as well as voice recording. Interview questions delivered throughout the study by a member of the research team (in person) supplemented these data.

Procedure

Participants received a brief overview of the ElectronixTutor system, its purpose, intended use, and basic structure (i.e., a federated collection of learning systems). Following eye-tracking explanation and calibration, the participants received a list of tasks to complete in ElectronixTutor, and instructions to narrate their perceptions and thought process via think-aloud protocol. Research assistants instructed participants to navigate to specific functions corresponding to tasks in line with the intended classroom integration and independent study usage. Each instruction came with a brief scenario description to guide their usage. Total time for individual participants ranged from 30 minutes to one hour. After initial navigational tasks (e.g., find the home page), participants were presented with an AutoTutor problem and progressed through it normally. Next, the researchers presented a second problem and instructed the participants to provide a designated incorrect answer in their own words. Specifically, the question asked for the relationship between total current of a circuit and three branch currents. The correct answer is that the branch currents add up to the total current, but in the scenario, participants were instructed that they believed the relationship was multiplicative. That answer may or may not have met the criteria for correctness depending on how participants phrased it. This manipulation created the first possibility for confusion, corresponding to the ‘Acceptable Faults’ heuristic. Participants were then instructed to pose questions to the system about the main question. This manipulation created the second possibility for confusion, corresponding to the ‘Question Rerouting’ heuristic. Participants then proceeded

through other stages relevant to the broader evaluation of ElectronixTutor, but not to the current study. Each stage entailed its own follow-up interview questions administered by a member of the research team in person.

Data Analysis

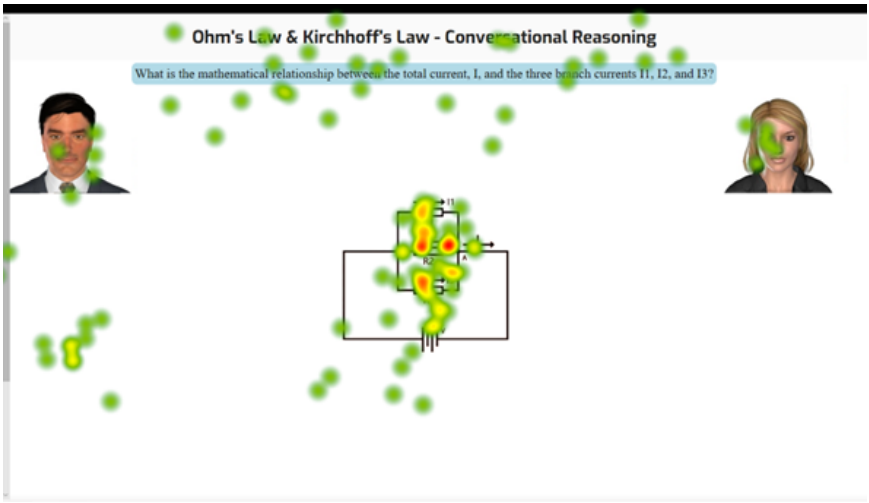
Following the study period, two graduate research assistants transcribed all audio recordings. Transcripts were then organized into idea units (Weinberger & Fischer, 2006). These idea units had break points when (1) the participants spoke about an interaction with the interface (2) the participants spoke about a learning interaction with a tutor or multiple-choice questions, or (3) participants completed a task. Approximate time codes supplement the transcript data.

Results & Discussion

The seven participants demonstrated relatively stable patterns across the three critical tasks. In the unconstrained AutoTutor interaction (i.e., when they provided good-faith attempts to answer questions), we see patterns exemplified by Figure 1. Here, participants move relatively quickly over the text of the main question (blue box near the top), briefly look at the agents (tutor agent male in the top left, peer agent female in the top right), and focus primarily on the circuit diagram (center).

Figure 1

Typical AutoTutor Visual Fixation Pattern



Visual fixation heat map showing most fixation on the circuit diagram with less on the question text and talking head avatars.

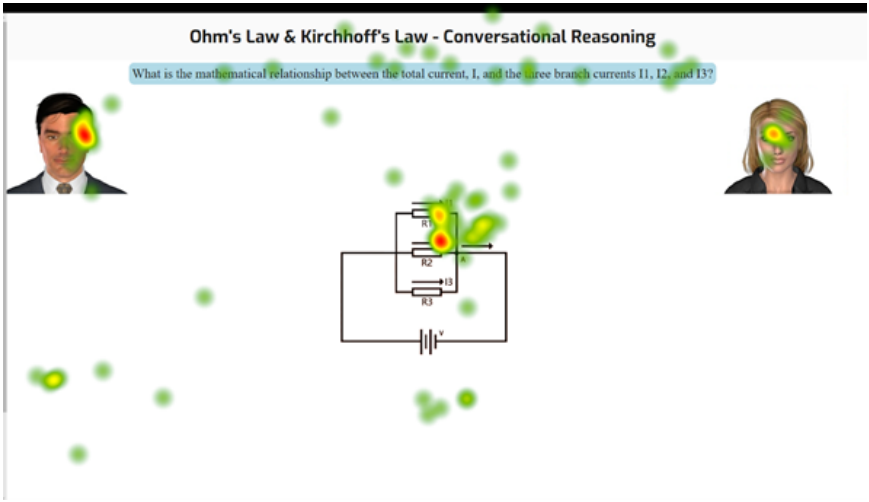
This is an appropriate fixation pattern. The visual presentation of the main question is redundant with auditory presentation and meant only for reference in case of confusion or uncertainty, therefore demanding minimal visual attention. Likewise, the agents provide social support when needed, but should not serve as a regular focal point. This pattern may serve as a point of reference for our potentially confusing situations below.

Incorrect Answers

This pattern changes substantially when participants give incorrect answers (see Figure 2). Here participants spend considerably more time viewing the agents. Notably, all the participants phrased their incorrect answers in a way that met the threshold for a correct answer (by virtue of using several relevant content words in structurally appropriate sentences). This means that AutoTutor responded in exactly the same way in both scenarios. The tutor agent posed the main question, the peer agent nominally agreed with the participant's input and then stated the correct answer, and the tutor agent affirmed, followed by a conversational transition to the next problem.

Figure 2

AutoTutor Visual Fixation Pattern Following Incorrect Input



Visual fixation heatmap showing roughly equal fixation on the circuit diagram and tutor talking head avatar, with some fixation on peer talking head avatar, and less on the question text.

Based on this visual fixation pattern and verbal participant reports, avoiding false alarms in error detection clearly does not come without cost. Many participants explicitly stated their confusion. One participant exemplified the struggle in his think-aloud transcript.

> 'Instead of correcting it afterwards and explaining it instead is straight up just changed from being multiplied to, it says the sum of the three. Yeah. I just be confused 'cause it said two different things were right. When they are very different mathematical relationships.'

Another participant suffered even more confusion.

> 'It really made me question whether or not I truly know electrical engineering or not, because last time I checked if the sum of all currents going into this one node is going to be equal to that

> output current… But, uh, I guess that's, that's false.'

We corrected this misconception along with an explanation of the technical issue during the debrief. None of the participants articulated the intended pedagogical strategy of correcting misconceptions implicitly by means of the summary at the end of the problem, or the technical concern of avoiding false alarms. Though we cannot say for certain if confusion would be higher or lower for learners who honestly gave incorrect answers, it seems clear that some alteration is necessary. Adjusting the bias likely invites a host of opposing but no less serious problems. Improvements to natural language processing diagnostics obviously offer incremental improvement, though at some (likely substantial) programming cost.

With the learning experience as the guiding principle (Tawfik et al., 2021), the most cost-effective improvement may reside in the conversational exchange framework. Perhaps designers can delineate an intermediate threshold of uncertainty that triggers a new transition. Instead of agreeing followed by a statement of the correct answer, the peer agent can indicate that she is submitting an independent answer without directly addressing the learner (e.g., "How about this answer…"). Given the relatively low match between the ideal answer and the learner's submission, the learner is unlikely to perceive this as the peer agent "stealing" a correct response. The relative cost (e.g., the learner feeling ignored) also seems low.

Questions

When instructed to pose their own questions to the agents, participants typically posed simple requests for clarification or additional information, as anticipated. These questions were almost entirely redundant with information provided in the Point & Query function or in dialogue. However, this did not prevent participants from becoming confused at the lack of direct response from the agents. The intended system response to learner-generated queries should consist of a hint, prompt, or pump. Two of these three options typically take the form of a question, resulting in a paradigm wherein AutoTutor "answers" learner questions with a question of its own.

Some participants found this frustrating if they attempted to persist in acquiring an answer.

> 'I feel like I ended up more confused at the end 'cause every time I'd ask it a question it just kind of ignored it and asked me a

> different question and eventually just gave up and went through it but wasn't really leading me anywhere.'

Though this participant recognized the structure, she did not find it useful or seem to identify any benefit to the approach. Another participant explicitly noted the (correctly) perceived pedagogical strategy but did not feel it was used effectively and explicitly indicated the strategy for not engaging him.

> 'It didn't directly answer my question. It answered my question with a question, which I guess is fine if it makes you want to think more. But, yeah, like if it had an answer along with a question to give you some type of reference to back it off of, that would've made it a little bit more intuitive. You know, just answering the question with another question is kind of repetitive... Not really engaging.'

These reactions, similar to the effect with incorrect answers, suggest the need for some design intervention. Once again, improved natural language processing offers improvement but at the cost of substantial programming effort and increased complexity in classification and response. However, two strategies for remediation immediately present themselves.

First, a minimal conversational transition could lessen the blow of such an abrupt transition to a new question. A relatively simple "canned" expression may indicate, if not comprehension then acknowledgement of the learner's input. "Hmm… well how about this:" before asking a conceptually related question would pose little risk of clashing with existing interactions.

Second, the almost complete overlap of questions asked and information provided elsewhere suggests that adequate design can prevent learners from constructing questions in the first place. A more detailed or interactive walkthrough of the AutoTutor Point & Query system may eliminate most questions. Further, the system already includes a function to review the full conversational exchange, but at least one participant searched for it without success. Providing affordances to convey existing functionality may essentially eliminate the shortcomings of question rerouting as a pedagogical strategy within conversational ITS.

Implications & Conclusions

One design implication relates to the use of a more computationally efficient approach to NLP without compromising the learning experience. Expense constitutes a critical barrier to entry for NLPs (Strubell et al., 2019), along with the lack of novice-friendly authoring tools needed to integrate the technology (Cai et al., 2015). In many cases, NLP in educational contexts requires a range of experts' input. Domain experts must work closely with script authoring experts knowledgeable in techniques such as latent semantic analysis to evaluate the distance in meaning between an expected answer and a novel one (e.g., Dumais, 2004), or epistemic network analysis (Shaffer et al., 2009) to granularly monitor the knowledge state of the learner. Depending on the complexity and maturity of the authoring tools, implementing these techniques in place may require ongoing coordination with computer scientists. Often the creator of the learning system serves as an indispensable component in this relationship, wearing one or more of these hats in addition to project manager, and severely limiting development at scale. In this study, the approach to NLP attempted to mitigate this bottleneck by utilizing several less computationally intensive techniques, leveraging the constraints of the task and domain to avoid depreciation of the perceived intelligence of the system. However, that perception is unavoidably subjective and requires evaluation.

This study also has implications for how instructional designers should consider prompting cognitive disequilibrium in learners through AI. Indeed, theorists have argued that failure can be an important part of the learning experience and yield specific outcomes related to problem representation (Kapur, 2018), causal reasoning (Tawfik et al., 2015), decision-making (Rong & Cho, 2018), and others. The evidence suggests that failure often causes learners to reflect on the reason for failure, which leads to a more nuanced approach for the subsequent iteration of problem-solving. That said, the present study suggests designers must be careful about finding a balance between productive disequilibrium and cognitive overload. In this study, the unexpected dialogue moves by the AutoTutor agents caused confusion, self-doubt, and frustration. Additionally, participants did not assign value to the AI actions. These outcomes threaten the dynamic interaction between the learner and the learning space that promotes continued self-paced learning (Tawfik et al., 2021). Tawfik et al. (2021) identified both assignment of value and dynamic interaction as essential pieces of learning experience design (LXD). Instructional designs working with intelligent learning technology must consider these elements of LXD when contemplating the inclusion of similar AI promptings and interactions.

Another design implication relates to the affective response to AI interaction. Using AI tutors to promote context-based affective responses during online learning can lead to a desire to pursue learning and promote self-efficacy. As previously stated in the paper, cognitive disequilibrium is needed to advance learning. Learners generally wish to resolve the cognitive disequilibrium they are feeling, which can lead to active engagement with the content. However, the data suggest that if the AI creates too high of an affective response, users become highly confused, lose self-efficacy in their knowledge, and begin questioning their knowledge beyond the intent of the AI response. Using AI to generate appropriate levels of cognitive disequilibrium does have practical implications for those that wish to implement AI in their design and learning environment, namely, how to balance task complexity, AI response, and learner self-efficacy.

A learning environment, like conversation itself, is most effective when the participants understand the intentions and capabilities of their opposite number. Learners come into intelligent tutoring systems, and particularly conversational ITS, with expectations for how the system will behave. Those expectations do not seem to allow designers' concerns to supplant conversational norms without explanation. Computational, statistical, and pedagogical constraints do not factor highly into learners' anticipations and subsequent evaluations. This disconnect in models of interaction will likely lead to distrust and disengagement.

Structural adjustments to the task presentation and conversational frame may have the ability to lessen these substantial stumbling blocks with minimal computational effort or risk of interfering with other finely tuned interaction patterns. Vague conversational transitions on the way to existing pedagogical strategies may smooth what learners perceive as abrupt shifts. Improved design of affordances may prevent the need for interaction types to which conversational ITS are not well suited. By anticipating learner expectation, designers can improve the perceived intelligence of their systems and let their pedagogical strategies work optimally.

Acknowledgements

This research, as well as the construction of ElectronixTutor, was enabled by funding from the Office of Naval Research (N00014-00-1-0600, N00014-15-P-1184; N0001412-C-0643, N00014-16-C-3027). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR.

References

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humphreys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies, 91*, 24–36. https://edtechbooks.org/-LhI

Cai, Z., Graesser, A. C., & Hu, X. (2015). ASAT: AutoTutor script authoring tool. *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools, 3*, 199–210. https://edtechbooks.org/-nejw

Chae, S. W., Lee, K. C., & Seo, Y. W. (2016). Exploring the effect of avatar trust on learners' perceived participation intentions in an e-learning environment. *International Journal of HumanComputer Interaction, 32*(5), 373–393. https://edtechbooks.org/-bTqh

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education, 114*, 193–205. https://edtechbooks.org/-keV

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230. https://edtechbooks.org/-oRB

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26(1)*, 124–132. https://edtechbooks.org/-qXYM

Graesser, A. C., Hu, X., Nye, B. D., VanLehn, K., Kumar, R., Heffernan, C., ..., & Baer, W. (2018). ElectronixTutor: An intelligent tutoring system with multiple learning resources for electronics. *International Journal of STEM Education, 5*(1), 1–21. https://edtechbooks.org/-FUMG

Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A., Person, N., & Tutoring Research Group. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head. In *Proceedings of Artificial Intelligence in Education* (Vol. 4754). https://edtechbooks.org/-QoTb

Hampton, A. J., & Graesser, A. C. (2019). Foundational principles and design of a hybrid tutor. In R. A. Sottilare & J. Schwarz (Eds.) *Proceedings of the First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference* (pp. 96–107), Orlando, FL, USA, July 26–31, 2019.

https://edtechbooks.org/-QPi

Hampton, A. J. & Wang. L. (2019, July). Conversational AIS as the cornerstone of hybrid tutors. In R. A. Sottilare & J. Schwarz (eds.) *Proceedings of the First International Conference on Adaptive Instructional Systems*, (pp.634–644), Springer, Cham. https://edtechbooks.org/-NiNa

Kapur, M. (2018). Examining the preparatory effects of problem generation and solution generation on learning from instruction. *Instructional Science, 46*(1), 61–76. https://edtechbooks.org/-Crpf

Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., & Gratch, J. (2016). Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers & Education, 99*, 1–13. https://edtechbooks.org/-Nbs

Lee, J.-E. R., Nass, C., Brave, S. B., Morishima, Y., Nakajima, H., & Yamada, R. (2007). The Case for caring colearners: The Effects of a computer-mediated colearner agent on trust and learning. *Journal of Communication, 57*(2), 183–204. https://edtechbooks.org/-PXak

Lehman, B., D'Mello, S., & Graesser, A. C. (2012). Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education, 15*(3), 184–194. https://edtechbooks.org/-zbuh

Lehman, B., D'Mello, S., Strain, A., Mills, C., Gross, M., Dobbins, A., … & Graesser, A. C. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education, 22*(1-2), 85-105. https://edtechbooks.org/-Cxsh

Li, W., Wang, F., Mayer, R. E., & Liu, H. (2019). Getting the point: Which kinds of gestures by pedagogical agents improve multimedia learning? *Journal of Educational Psychology, 111*(8), 1382–1395. https://edtechbooks.org/-zStu

Louwerse, M. M., Graesser, A. C., McNamara, D. S., & Lu, S. (2009). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology, 23*(9), 1244–1255. https://edtechbooks.org/-Vacz

Lusk, M. M., & Atkinson, R. K. (2007). Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? *Applied Cognitive Psychology, 21*(6), 747–764.

https://edtechbooks.org/-VcJC

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education, 28*(3), 420–438. https://edtechbooks.org/-bmF

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The Case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19*(2), 177-213. https://edtechbooks.org/-UzRr

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20*(3/4), 263–294. https://edtechbooks.org/-tGiE

Nielsen, J., & Loranger, H. (2006). *Prioritizing web usability*. Pearson Education.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*(4), 427–469. https://edtechbooks.org/-NFx

Rong, H., & Choi, I. (2018). Integrating failure in case-based learning: a conceptual framework for failure classification and its instructional implications. *Educational Technology Research & Development, 67(3),* 617–637. https://edtechbooks.org/-mxYx

Savin-Baden, M., Tombs, G., & Bhakta, R. (2015). Beyond robotic wastelands of time: Abandoned pedagogical agents and “new” pedalled pedagogies. *E-Learning and Digital Media, 12*(3–4), 295–314. https://edtechbooks.org/-bkMh

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., … & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media, 1*(2). https://edtechbooks.org/-ebY

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *The 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy. https://edtechbooks.org/-JHtj

Tawfik, A. A., Gatewood, J., Gish-Lieberman, J. J., & Hampton, A. J. (2021) Towards a definition of learning experience design. *Technology, Knowledge and*

Learning, 1-26. doi:10.1007/s10758-020-09482-2

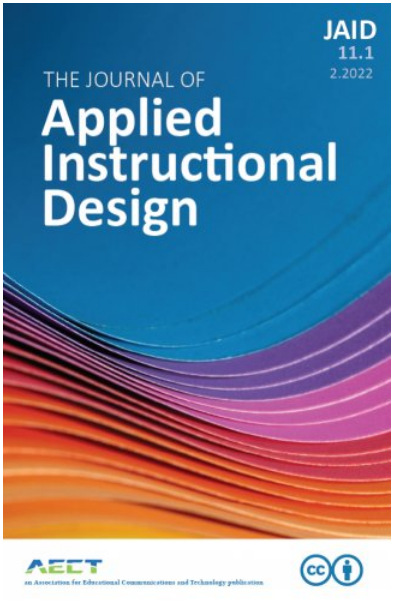
Tawfik, A. A., Rong, H., & Choi, I. (2015). Failing to learn: Towards a unified design approach for failure-based learning. *Educational Technology Research and Development, 63*(6), 975–994. doi:10.1007/s11423-015-9399-0

U.S. Navy. (1998). *Navy electricity and electronics training series (Vols. 1–24).* Pensacola, FL: Naval Education and Training Professional Development and Technology Center.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66*(2), 98–112. https://edtechbooks.org/-ZPs

Weinberger, A., & Fischer, F. (2006). A Framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95. https://edtechbooks.org/-SZv

Wu, L., & Looi, C.-K. (2012). Agent prompts: Scaffolding for productive reflection in an intelligent learning environment. *Journal of Educational Technology & Society 15*(1), 339–353. https://edtechbooks.org/-ZdcW



Hampton, A. J., Gish-Lieberman, J. J., Gatewood, J., & Tawfik, A. A. (2021). Say What? Learner Reactions to Unexpected Agent Dialogue Moves. *The Journal of Applied Instructional Design, 11*(1). https://edtechbooks.org/jaid_11_1/say_what_learner_rea