ARTIFICIAL INTELLIGENCE AS AN EFFECTIVE CLASSROOM ASSISTANT[1]

Benedict du Boulay
Human Centred Technology Group, Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK
B.du-Boulay@sussex.ac.uk

Introduction

The field of Artificial Intelligence in Education (AIED) has been in existence for about 40 years and operated under various other names, the most common of which is Intelligent Tutoring Systems. The field was initially brought to wider attention by papers in a special issue of the International Journal of Man-Machine Studies (see e.g., Brown, Burton, & Bell, 1975), by papers in a book based on that special issue (see e.g., O'Shea, 1982) and in Artificial Intelligence books of the era (see e.g., Brown & Burton, 1975). This field used and continues to use techniques from artificial intelligence and cognitive science to attempt to understand the nature of learning and teaching and so build systems to assist learners to master new skills or to understand new concepts, in ways that mimic the insightful and adaptive tutoring of a skilled human tutor working one-to-one with the learner. That is to say, such systems attempt to adapt the way that they teach to the existing and developing knowledge and skill of the learners, to their preferred ways of going about learning, and to take into account the affective trajectory of the learners as they deal with the expected set-backs and impasses of mastering new material. There is clearly some overlap with other uses of computing technology in education, though the commitment to individual adaptation through modelling different parts of the educational process is a key defining characteristic.

In order for such systems to adapt to the learner and so provide a personalised learning experience, a typical conceptual architecture has evolved. This consists of (i) a model of the domain being learned so that the system can reason about and judge whether a student's answer or indeed a problem-solving step is appropriate; (ii) a model of the current level of the learner's understanding or skill, so that tasks of appropriate complexity can be posed; (iii) a model of pedagogy so that the system can make sensible tutorial moves such as providing effective feedback or adjusting the nature of the next task; and (iv) one or more interfaces through which the system and the learner can communicate to explore and learn about the domain in question.

[1] This column is an adapted and enlarged version of a letter to the Editor of the International Journal of Artificial Intelligence in Education (du Boulay, 2016) .

Over the years many systems using a variety of pedagogical techniques and topics have been built and evaluated. To give a sense of the wide scope of the work, four systems are mentioned. These have been chosen for their diversity and range from classic teaching in a formal subject and a procedural skill, through learning by creating externalised forms of knowledge for a highly conceptual learning task to rich, natural user interaction via speech for a learning complex culture-laden skills.

The four examples of AIED systems are: (i) a system to help learners understand basic algebra by being set problems and provided with step by step feedback and guidance on their solution (Koedinger, Anderson, Hadley, & Mark, 1997); (ii) a system to help learners gain a conceptual understanding of river ecosystems by building a concept map of that domain, as if for another learner, and having that simulated other learner take tests on the adequacy of the concept map so far built (Leelawong & Biswas, 2008); (iii) a system to help military personnel both learn and speak Arabic as well as understand the social and cultural norms needed to interact with people in the country within which they are operating (Johnson, 2010).

The fourth example system illustrates the increasing importance of the interface in AIED systems and their use in informal, such as museums, as well as formal learning environments. The screenshot in Figure 1 shows Coach Mike, a pedagogical agent designed to help children visiting a museum to learn about robotics. This kind of application extends the role of classroom teaching: "it means that such systems need to go beyond simply focusing on knowledge outcomes. They must take seriously goals such as convincing a visitor to engage, promoting curiosity and interest, and ensuring that a visitor has a positive learning experience. In other words, pedagogical agents for informal learning need to not only act as coach (or teacher), but also as *advocate* (or salesperson)" (Lane et al., 2013, p. 310).

Figure 1 Coach Mike in three different poses, taken from (Lane et al., 2013)

Coach Mike was designed to emulate some of the work of the human museum

curators, including helping to orientate visitors, encouraging them to explore and providing problem-solving challenges and support.

A number of papers recently have argued the case for the benefit of artificial intelligence systems in education (see e.g., Luckin, Holmes, Griffiths, & Forcier, 2016; Woolf, Lane, Chaudhri, & Kolodner, 2014), while others have been more sceptical (see e.g., Enyedy, 2014). This column looks at the evidence derived from meta-reviews and meta-analyses conduced over the last 5 years. Its main focus is on the comparative effectiveness of AIED systems *vs* human tutoring. We note in passing the a meta-review of the use of pedagogical agents (not necessarily in AIED systems) "produced a small but significant effect on learning" (Schroeder, Adesope, & Gilbert, 2013).

This column is absolutely not intended as support for an argument about getting rid of human teachers, but is intended as support for blended learning where some of the human teacher's work can be off-loaded to AIED systems, as if to a classroom assistant.

Meta-analysis and Meta-reviews

Since 2011 a number of meta-reviews and meta-analyses have attempted to determine the degree to which a whole host of systems such as those mentioned above have been educationally effective. Typically this has meant comparing them in terms of learning gains with other instructional methods, such whole class teaching by a human teacher or use of a text-book without a teacher.

Van Lehn's meta-analysis

Under the title "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems", VanLehn (2011) analysed papers comparing five types of tutoring. These were: no tutoring (e.g. learning with just a textbook), answer-based tutoring, step-based tutoring, substep-based tutoring and human tutoring.

The difference between Answer-based, step-based and substep-based is in terms of the granularity of the interaction between tutor and student, see Sidebox 1. Answer-based systems are capable of providing hints and feedback only at the level of the overall answer. Step-based systems are capable of providing hints, scaffolding and feedback on every step that the student makes in the problem-solving. By contrast substep-based systems work at a finer level of granularity still and "can give scaffolding and feedback at a level of detail that is even finer than the steps students would normally enter when solving a problem" (VanLehn, 2011, page 203). Artificial Intelligence techniques are required to underpin both step-based and substep-based tutors, while answer-based systems would typically otherwise fall under the heading of computer-based or computer-assisted instruction.

Sidebox 1.

For example, imagine that the problem is to solve the equation

$2(14 - x) = 23 + 3x$

An answer-based system would expect the student to do all the working offline and then provide the answer $x = 1$. If asked for a hint prior to the answer being provided, the tutor can suggest broad ways of going about the problem, such as collect all the terms in x on one side of the equation, but has no way of knowing that this advice is being followed. If the answer provided is wrong e.g. $x = 1.25$, the tutor may be able hypothesise that perhaps the student multiplied out the bracket incorrectly, but if the answer provided is $x = 14$, it probably will not be able to offer much in the way of specific help.

In a step-based system, the student might be invited to multiply out the bracket expression as a first step and so types in $28 - 2x$ as the answer to that step. If a hint is requested or a wrong answer given to this step, then help can be given about the working of that step. Once the step is completed correctly, the tutor would invite an answer to the next step, e.g. reordering terms in the equation, and then on through further steps to the final answer.

In a substep-based system there might be a remedial dialogue at a finer level than an individual step, for instance about what expressions such as $2x$ or $3x$ mean, if that seems warranted by the request for a hint or by a wrong step answer.

Given the above levels of granularity, VanLehn derived 10 pairwise comparisons of effect sizes, see Table 1. In this table the rightmost column shows the proportion of the results for that row where the individual study comparison was statistically reliable at the level $p < 0.05$.

Table 1. Effect sizes adapted from (VanLehn, 2011). Row[1] was taken by VanLehn from a separate study (C.-L. C. Kulik & Kulik, 1991).

	Comparison	No of studies	Mean Effect Size	%reliable
1	Answer-based *vs* no tutoring[1]	165	0.31	40%
2	Step-based *vs* no tutoring	28	0.76	68%
3	Substep-based *vs* no tutoring	26	0.40	54%
4	Human *vs* no tutoring	10	0.79	80%
5	Step-based *vs* answer based	2	0.40	50%
6	Substep-based *vs* answer-based	6	0.32	33%
7	Human *vs* answer based	1	-0.04	0%

8	Substep based *vs* step based	11	0.16	0%
9	Human *vs* step based	10	0.21	30%
10	Human *vs* substep based	5	-0.12	0%

For the purposes of this review, the comparison of most interest is row 9, that between one-to-one human tutoring and step-based tutors (effect size = 0.21). By collating all the results in Table 1, VanLehn found that step-based tutors were, within the limitations of his review, "just as effective as adult, one-to-one tutoring for increasing learning gains in STEM topics" (VanLehn, 2011, page 214). He also found that while increasing the granularity of instruction from answer-based to step-based yielded significant gains, going to the finer level of substep-based tutoring did not add further value. Note that this latter finding was based on a small number of studies only.

Four Meta-reviews

Since VanLehn's meta-analysis, four meta-reviews have been published, as well as a large-scale study of a specific tutor, see Table 2. In this table the No. of Studies column shows the number of instances for the given comparison in that row, not the total number of studies in the overall meta-review.

Table 2. Six meta-reviews and a large scale study.

***The standard error in row 1 is based on all 10 studies, not just the 30% that produced reliable results, see Table 1.**

$Standard errors computed by the author of this paper.

	Meta-review	Comparison	No. of Comparisons	Mean Effect Size	Standard Error
1	VanLehn (2011)	Step based vs one-to-one human tutoring	10	-0.21	0.19*$
2	Ma et al. (2014)	Step based *vs* one-to-one human tutoring	5	-0.11	0.10
3		Step based *vs* "large group human instruction"	66	0.44	0.05
4	Nesbit et al. (2014)	Step based *vs* "teacher led group instruction"	11	0.67	0.09
5	Kulik et al. (2016)	(Step based and Substep based) *vs* "conventional classes"	63	0.65	0.07$
6	Steenbergen-Hu et al. (2014)	Step based vs one-to-one human tutoring	3	-0.25	0.24
7		Step based vs	16	0.37	0.07

		"traditional classroom instruction"			
8	Steenbergen-Hu et al. (2013)	(Step based and answer based) vs "traditional classroom instruction"	26	0.09	0.01
9	Pane et al. (2014)	Blended learning including a step-based system *vs* traditional classroom instruction	147 schools	-0.1	0.10
				0.21	**0.10**
				0.01	0.11
				0.19	0.14
10	Weighted mean	AIED system vs one-to-one human tutoring	18	-0.19	
11	Weighted mean	AIED system vs conventional classes	182	0.47	

In a meta-review of 107 studies, Ma, Adesope, Nesbit, and Liu (2014) found similar results to VanLehn for step-based ITSs both when compared to no tutoring condition (i.e. just a textbook; mean effect-size = 0.36) and, more positively than VanLehn, when compared to large group human teacher led-instruction (mean effect size = 0.44), but no differences when compared to small group human tutoring or one–to–one tutoring.

The same authors analysed 22 systems for teaching programming and also found a "a significant advantage of ITS over teacher-led classroom instruction and non-ITS computer-based instruction" (Nesbit, Adesope, Liu, & Ma, 2014). A larger version of a similar study involving 280 studies is currently in progress (Nesbit, Liu, Liu, & Adesope, 2015).

In a meta-review of 50 studies involving 63 comparisons, J. A. Kulik and Fletcher (2016) found similar sized improvements (mean effect size = 0.65) but distinguished between studies that used standardised tests from those where the tests were more specifically tuned to the system providing tuition, with smaller effect sizes when standardised tests were employed. Overall they concluded that "This meta-analysis shows that ITSs can be very effective instructional tools . . . Developers of ITSs long ago set out to improve on the success of CAI tutoring and to match the success of human tutoring. Our results suggest that ITS developers have already met both of these goals" (J. A. Kulik & Fletcher, 2016, page 67). They also found better results for substep based systems than VanLehn, which they ascribed to differing comparison methodologies.

Much smaller effect sizes were found by Steenbergen-Hu and Cooper (2013) in their meta-analysis of pupils using ITSs in school settings. J. A. Kulik and Fletcher (2016) put this down to the weaker study inclusion criteria (e.g. the

inclusion of answer based systems as if they were step based systems) used by Steenbergen-Hu and Cooper who also noted that lower-achievers seemed to do worse with ITSs than did the broad spectrum of school pupils, though this result is again disputed by Kulik and Fletcher. However, in a parallel study of university students, Steenbergen-Hu and Cooper (2014) found more positive effect sizes (in the range 0.32 – 0.37) for ITSs as compared to conventional instruction. They conclude that "ITS have demonstrated their ability to outperform many instructional methods or learning activities in facilitating college level students' learning of a wide range of subjects, although they are not as effective as human tutors. ITS appear to have a more pronounced effect on college-level learners than on K-12 students" (Steenbergen-Hu & Cooper, 2014, page 344).

Rows 10 and 11 summarise the results of the meta-reviews, excluding the evaluation of the Cognitive Algebra Tutor, and show a weighted mean effect size of 0.47 for AIED systems *vs* conventional classroom teaching. We use the term AIED system to cover all the systems, step-based, substep-based and answer-based looked at in the meta-reviews. The comparison with one-to-one human tutoring shows that AIED system do slightly worse with a mean effect size of -0.19. In both cases the means are weighted in terms of the number of comparisons in the meta-review, not in terms of the original N values in the studies themselves.

The Cognitive Tutors

The Cognitive Tutor family of tutors "are found in about 3000 schools and over a half million students use the courses each year" (Koedinger & Aleven, 2016) and represent the most successful transition, in terms of numbers of students, of Artificial Intelligence in Education work from the laboratory to the classroom. They provide scaffolded help with step-by-step problem-solving in a variety of domains, mostly mathematical, and are designed to be used in a blended learning manner, thus freeing up the teacher to work with other children while some work with the tutors. Care is taken to ensure that teachers are trained to make the best of the arrival of these systems into their classrooms in terms of how to manage all the pupils in the classroom before, during and after the use of the tutors (Koedinger et al., 1997). Individual evaluations of various Cognitive Tutors are included in the reviews already described.

A large-scale study in the USA of the Cognitive Tutor for Algebra, (Pane, Griffin, McCaffrey, & Karam, 2014) undertook a between-schools project involving 73 high Schools and 74 middle Schools across 7 states. The schools were matched in pairs and half received the Cognitive Algebra Tutor and adjusted their teaching to include it as they saw fit, while the others carried on as before in terms of their normal method of teaching algebra. The study ran over two years and found no significant differences on post-test scores in the first year of the study but a small but significant effect size of 0.21 in the high schools in favour of the schools which used the Cognitive Tutor in the second year of the study (see data in bold, in row 9 of Table 2).

Note that how the Cognitive Tutor was actually used in the classrooms was not controlled, though post-hoc analyses showed that teachers did not generally use the Tutor exactly as recommended by its developers.

Conclusions

The overall conclusion of these meta-reviews and analyses is that AIED systems perform better than CAI systems and also better than human teachers working in large classes. They perform slightly worse than one-to-one human tutors. Note that most of the systems were teaching mathematics or STEM subjects, as these are the kinds of subjects for which it is easier to build the domain and student models mentioned in the Introduction. It should be noted that there was a degree of overlap between these meta-reviews and analyses in terms of the collections of individual evaluations from which they have drawn their conclusions.

The specific study of the Cognitive Tutor for Algebra evaluated its use as a blended addition to the normal algebra teaching in the schools where it was tried rather than as a total replacement for the teachers, and found good results in high schools, as opposed to middle schools, and in the second year of the evaluation, as opposed to the first year. For a whole variety of reasons, the way forward for AIED systems in the classroom must be the blended model, classroom assistants if you like, so as to provide detailed one-to-one tutoring for some of the students while the human teacher attends to others as well as having overall responsibility for all the students' progress.

Of course good post-test results are not the only criteria for judging whether an educational technology will be, or indeed should be, adopted (Enyedy, 2014). However the overall message of these evaluations is that blending AIED technology with other forms of teaching is beneficial, particularly for older pupils and college level students studying STEM subjects.

References

Brown, J. S., & Burton, R. R. (1975). Multiple Representations of Knowledge for Tutorial Reasoning. In D. G. Bobrow & A. Collins (Eds.), *Representation and Understanding* (pp. 311--349). New York: Academic Press.

Brown, J. S., Burton, R. R., & Bell, A. G. (1975). SOPHIE. A step towards a reactive learning environment. *International Journal of Man Machine Studies, 7*, 675--696.

du Boulay, B. (2016). Recent Meta-reviews and Meta Analyses of AIED systems. *International Journal of Artificial Intelligence in Education, 26*(1), 536-537. doi: http://dx.doi.org/10.1007/s40593-015-0060-1

Enyedy, N. (2014). Personalized Instruction: New Interest, Old Rhetoric, Limited Results, and the Need for a New Direction for Computer-Mediated Learning. (pp. 1-22). Boulder, Colorado: National Education Policy Center.

Johnson, W. L. (2010). Serious Use of a Serious Game for Language Learning. *International Journal of Artificial Intelligence in Education, 20*(2), 175-195.

Koedinger, K. R., & Aleven, V. (2016). An Interview Reflection on "Intelligent Tutoring Goes to School in the Big City". *International Journal of Artificial Intelligence in Education, 16*(1), 13-24. doi: http://dx.doi.org/10.1007/s40593-015-0082-8

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education, 8*(1), 30-43.

Kulik, C.-L. C., & Kulik, J. A. (1991). Effectiveness of Computer-Based Instruction: An Updated Analysis. *Computers in Human Behavior, 7*(1-2), 75-94. doi: http://dx.doi.org/10.1016/0747-5632(91)90030-5

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research, 86*(1), 42-78. doi: http://dx.doi.org/10.3102/0034654315581420

Lane, H. C., Cahill, C., Foutz, S., Auerbach, D., Noren, D., Lussenhop, C., & Swartout1, W. (2013). The Effects of a Pedagogical Agent for Informal Science Education on Learner Behaviors and Self-efficacy *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings* (pp. 309-318). Berlin: Springer.

Leelawong, K., & Biswas, G. (2008). Designing Learning by Teaching Agents: The Betty's Brain System. *International Journal of Artificial Intelligence in Education, 18*(3), 181-208.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence Unleashed. An argument for AI in Education. London: Pearson.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of educational psychology, 106*(4), 901-918. doi: http://dx.doi.org/10.1037/a0037123

Nesbit, J. C., Adesope, O. O., Liu, Q., & Ma, W. (2014). *How Effective are Intelligent Tutoring Systems in Computer Science Education?* Paper presented at the IEEE 14th International Conference on Advanced Learning Technologies (ICALT), Athens, Greece.

Nesbit, J. C., Liu, L., Liu, Q., & Adesope, O. O. (2015). *Work in Progress: Intelligent Tutoring Systems in Computer Science and Software Engineering Education*. Paper presented at the 122nd American Society for Engineering Education, Seattle.

O'Shea, T. (1982). A Self-Improving Quadratic Tutor. In D. Sleeman & J. S. Brown (Eds.), *Intelligent Tutoring Systems*: Academic Press.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis, 36*(2), 127-144. doi: http://dx.doi.org/10.3102/0162373713507480

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How Effective are Pedagogical Agents for Learning? A Meta-Analytic Review. *Journal of Educational Computing Research, 49*(1), 1-39. doi: http://dx.doi.org/10.2190/EC.49.1.a

Steenbergen-Hu, S., & Cooper, H. (2013). A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on K–12 Students' Mathematical Learning. *Journal of educational psychology, 105*(4), 970-987. doi: http://dx.doi.org/10.1037/a0032447

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of educational psychology, 106*(2), 331-347. doi: http://dx.doi.org/10.1037/a0034752

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational psychologist, 46*(4), 197-221. doi: http://dx.doi.org/10.1080/00461520.2011.611369

Woolf, B. P., Lane, H. C., Chaudhri, V. K., & Kolodner, J. L. (2014). AI Grand Challenges for Education. *AI Magazine, 34*(4), 66-84. doi: http://dx.doi.org/10.1609/aimag.v34i4.2490