How does affective robot feedback influence learner experience in a real-world treasure hunt?

Mary Ellen Foster[1], Amol Deshmukh[2], Srini Janarthanam[2], Mei Yii Lim[2],
Helen Hastie[2], and Ruth Aylett[2]

[1] School of Computing Science, University of Glasgow
18 Lilybank Gardens, Glasgow G12 8RZ, United Kingdom
MaryEllen.Foster@glasgow.ac.uk
[2] School of Mathematical and Computer Sciences, Heriot-Watt University,
Riccarton, Edinburgh EH14 4AS, United Kingdom
{A.Deshmukh,sc445,M.Lim,H.Hastie,R.S.Aylett}@hw.ac.uk

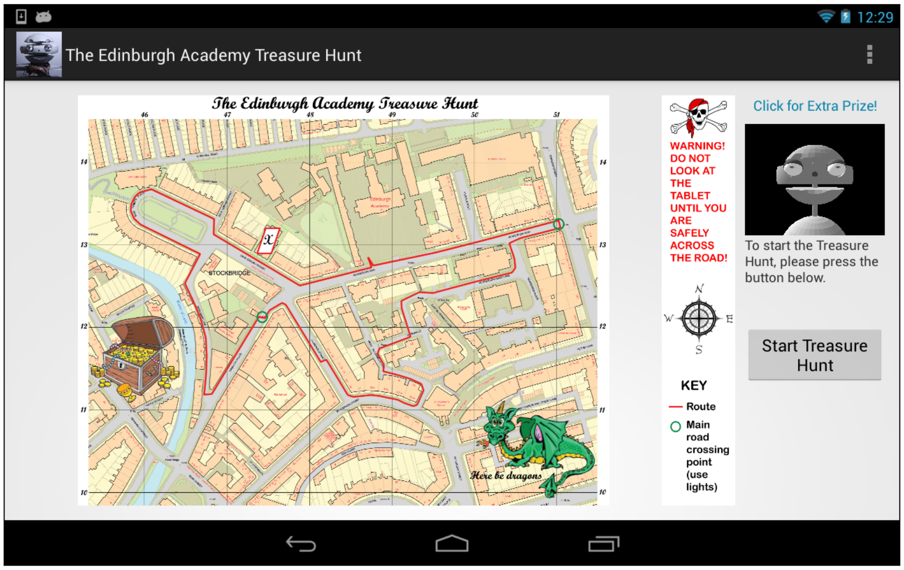
Abstract. We explore the effect of the feedback strategy used by a virtual robot agent in the context of a real-world treasure-hunt activity carried out by children aged 11–12. We compare two versions of a tablet-based virtual robot agent, which provides either neutral or affective feedback during the treasure hunt. The results suggest that the use of the tablet app increased the perceived difficulty of the instruction-following task compared to a paper-based version, while the affective robot feedback increased the perceived difficulty of the questions.

1 Introduction

Emotions play an important role in human-human interaction [6], and robotic agents that exhibit human-like emotions have now become commonplace in the domain of human-computer interaction. Starting from the pioneering work of Bates [3] and Picard [21], emotional agents now exist in various applications to serve different purposes, including military [12], health [5], commerce [11], tourism [18], and video games [14]. One rich application area for such agents has been education [8,9,13,19,20,22], where emotional expressions have been incorporated into embodied teaching agents with the aim of improving learning experience in users. Although the inclusion of emotional expressions into virtual tutors rarely leads to negative interaction, there have not always been positive effects on learning outcomes [4]. This might be due to the fact that the task of learning requires concentration: so if an agent offers assistance at an inappropriate time or in an inappropriate manner, the result may be more of a distraction than a help.

To establish successful human-robotic interactions in an educational context, it is therefore essential to understand the impact on the learner of affective behaviour from an embodied agent. Note that it is not sufficient to simply ask whether emotional agents are "better" or "worse" than unemotional agents [7]; rather, the relevant issues are: (1) the kinds of emotional expression that have an effect on users; (2) the elements of the user's attitude and/or performance that are affected; and (3) the impact of different forms of emotional expression.

This work takes place in the context of the EU project EMOTE[3] (EMbOdied-perceptive Tutors for Empathy-based learning), which has the overall goal of developing artificial tutors that have the perceptive and expressive capabilities to engage in empathic interactions with learners in school environments, grounded in psychological theories of emotion in social interaction and pedagogical models for learning facilitation. Previous studies on robotic companions in real-world classroom environments [17] have shown that robotic platforms are promising tools for experimental learning. We hypothesise that a robot tutor that is able to detect the user's affective state and respond appropriately will result in increased motivation and better learning outcomes—and a crucial aspect of this overall goal is the specification of appropriate robot behaviour.


12:29
The Edinburgh Academy Treasure Hunt
The Edinburgh Academy Treasure Hunt
Here be dragons
Click for Extra Prize!
N
W
E
S
KEY
Route
Main road crossing point (use lights)
Start Treasure Hunt

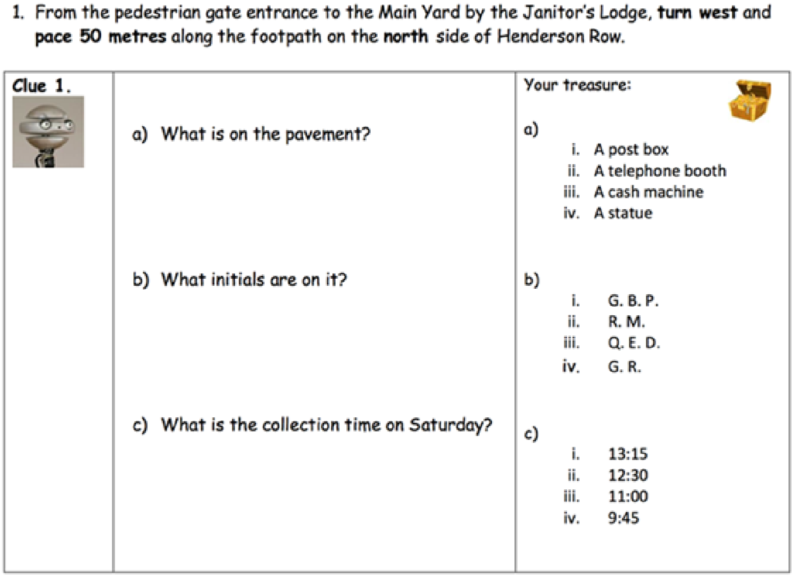
WARNING! DO NOT LOOK AT THE TABLET UNTIL YOU ARE SAFELY ACROSS THE ROAD!

To start the Treasure Hunt, please press the button below.

Fig. 1: The treasure hunt application

As one way of exploring this issue, we carried out a user study investigating how the nature of feedback from a virtual robot affects a child's perception, experience and performance in the context of a real-world treasure hunt activity. This study involved 37 students aged 11-12, divided into three experimental conditions. Two-thirds of the students used an Android-based tablet application (Figure 1) which displayed a digital map, along with a virtual robot head which presented the navigation instructions and posed the questions. The virtual robot also provided the students with feedback on the correctness of their answers; depending on the experimental condition, the feedback was either **neutral** ("correct", "incorrect") or **affective** (e.g., "well done", "too bad"). As a basis for comparison, a third group of students carried out the treasure hunt using the paper-based map and questionnaire that have been used in previous years.

[3] http://www.emote-project.eu/



1. From the pedestrian gate entrance to the Main Yard by the Janitor's Lodge, **turn west** and **pace 50 metres** along the footpath on the **north** side of Henderson Row.

Clue 1.		Your treasure:
	a) What is on the pavement?	a) i. A post box ii. A telephone booth iii. A cash machine iv. A statue
	b) What initials are on it?	b) i. G. B. P. ii. R. M. iii. Q. E. D. iv. G. R.
	c) What is the collection time on Saturday?	c) i. 13:15 ii. 12:30 iii. 11:00 iv. 9:45

Fig. 2: Excerpt from paper-based questionnaire

2 Treasure hunt activity

The real-world treasure hunt activity, which has been carried out at a local school for several years, requires the students to carry out a series of navigation steps in the real world, following one of two predetermined routes on a map. Each navigation step first requires the students to walk a few yards while making use of their map-reading skills, and then to answer a series of questions regarding their new location: for example, they might need to identify the colour of a nearby door. For the basic version of the treasure hunt, the students are given a map on paper, along with a paper-based questionnaire (Figure 2) listing the steps to follow and the multiple-choice questions to answer.

For the current study, we have developed an Android treasure hunt application, keeping the features as close to the paper version as possible. In particular, all images, fonts and layout are comparable between the two versions, and the application (Figure 1) displays a map corresponding to its paper counterpart and presents a sequence of the same steps as in the paper version to be carried out by the students. As the screen of the target tablet device (Galaxy Nexus 7) is smaller than a piece of A4-size paper, the map permits dragging and zooming to enable the students to explore it as they would with the paper version; note that the map cannot be zoomed to larger than 100% its actual size. Also, we chose not to display the student's current GPS location on the map, again to keep the tablet-based task as similar as possible to the paper-based version.

As shown in Figure 1, the tablet application also includes a virtual robot character, and each navigation step begins when the virtual robot presents the next navigation task to be carried out through speech. Subtitles are displayed on screen in case the students could not hear the speech (traffic noise was problematic at times on the route), and the students can also replay the speech at any point if necessary. When the students reach the target location, they press an on-screen button, and the app presents the next set of questions one at a time, along with the multiple choice answers (Figure 3). After the students choose an answer, the virtual robot indicates whether it is correct or incorrect, using one of two strategies: in the **neutral** strategy, the robot simply says “correct” or “incorrect”, while in the **affective** strategy, the robot uses phrases such as “well done” or “too bad”. Note that all other aspects of the robot behaviour were kept the same in both conditions, including all non-verbal signals; the difference was only in the actual words used.

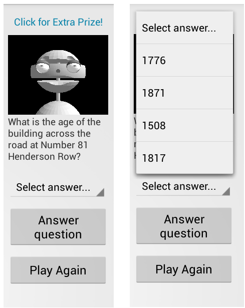
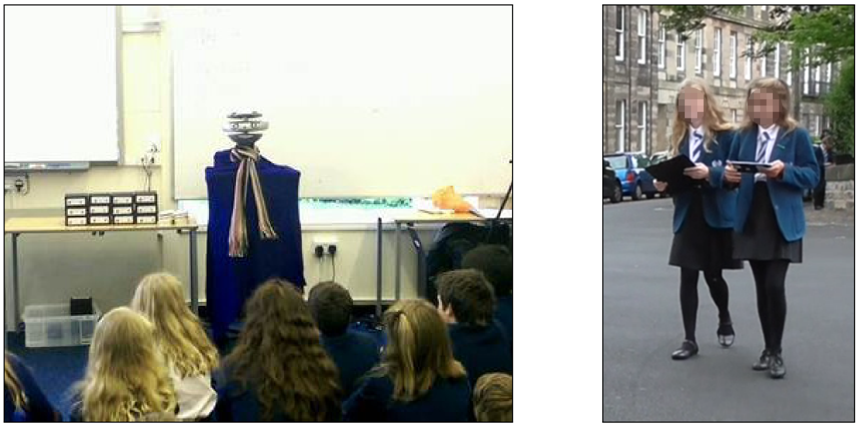

What is the age of the building across the road at Number 81 Henderson Row?

Fig. 3: App details

3 User evaluation

3.1 Participants and procedure


(a) The robot introducing the activity

(b) Hunting

Fig. 4: Steps in the treasure hunt activity

37 students aged 11–12 (7 female) participated in the user study. At the start of the study, a physical robot—the EMYS head [15], here called “Susie”—introduced the activity to the group of students (Figure 4a) and conducted a short Q&A session to check the students’ readiness for the activity; this session also served to introduce the students to the voice and appearance of the robot agent. The students were then sent off on the treasure hunt in groups of two[4] (Figure 4b) at regular intervals. The students’

[4] One group had three members.

progress around the route was recorded using a GPS logger. After a group returned from the treasure hunt, each group member individually completed a short subjective questionnaire.

3.2 Independent measures

We independently manipulated two features of the treasure hunt activity during this study, both in a between-groups design. First, we controlled the level of interaction: one-third of the groups used the tablet application with neutral feedback from the virtual robot, one-third used the tablet application with affective feedback, and the remaining third used the basic paper map and questionnaire. In addition, half of the groups across all conditions followed the treasure hunt route in a clockwise direction (Route 1), while the other half followed the route in an anticlockwise direction (Route 2). This is the process normally used by the school in previous years to allow more groups to be sent out at the same time without getting in each other's way.

3.3 Dependent measures

We gathered two classes of dependent measures: estimated task success based on the GPS logs, as well as a range of subjective measures computed from the questionnaires.

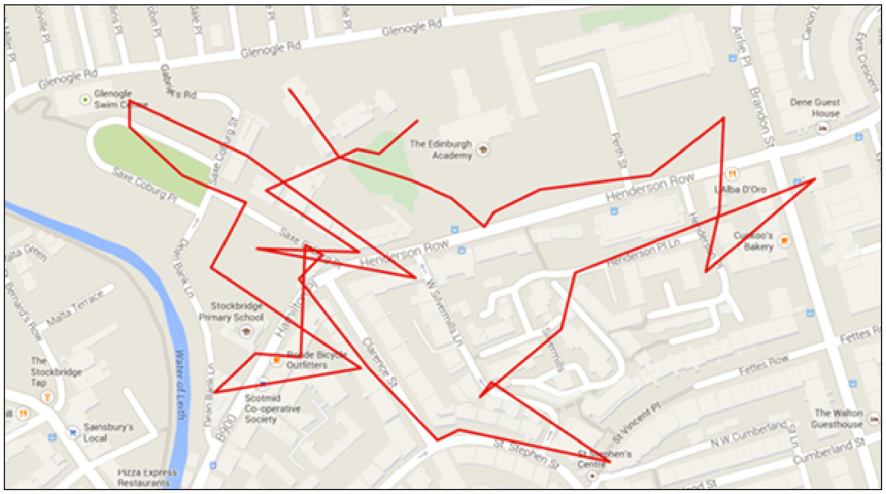


Fig. 5: Sample GPS trace

We used the GPS traces of the groups' progress around the treasure hunt route to estimate their overall success at the treasure hunt task; an example trace is shown in Figure 5. Using the traces, we computed two measures of task success: the number of the intended waypoints that the group reached (from a maximum of 18), and the total time taken to complete the treasure hunt route. We had also intended to assess the task performance more directly by analysing the subjects' responses to the treasure-hunt questions; however, due to a technical failure, that data could not be analysed.

The students' subjective experience of the treasure hunt was measured through a three-part questionnaire (the full set of questions is given in Appendix A):

- Three questions regarding the student's opinion of the robot before the treasure hunt;
- Four questions regarding the treasure hunt itself; and
- Seven questions addressing the students' opinion of the virtual robot during the treasure hunt.

The items from the last part of the questionnaire were based on the Godspeed questionnaire series [2], which is designed to be a standard user measurement tool for human-robot interaction. The items were drawn primarily from the "likeability" portion of the questionnaire and were rephrased to make them clear for the target age group. This portion of the questionnaire was given only to students from the groups from the two tablet conditions. All questions were presented using a five-point Smileyometer (Figure 6), which has been shown to be a good instrument for evaluating child-computer interactions [25]. Based on our prior experience of childrens' responses to a Smileyometer-like instrument, we did not include any frowning faces, as we anticipated that the students would be reluctant to choose an overtly negative subjective response.

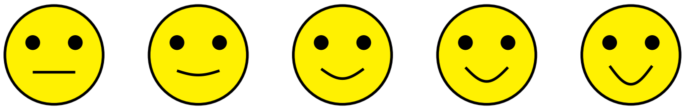


Fig. 6: The five-point Smileyometer

3.4 Results

We first summarise the overall results on the objective and subjective measures, and then measure the influence of the two experimental manipulations on these results.

Table 1: Overall objective results

Measure	Mean	Median	Min	Max
Waypoints met	16.8	18	13	18
Duration	58:42	54:52	44:20	1:28:23

Summary Table 1 summarises the task-success measures derived from the GPS traces; note that the data for one group (neutral feedback, route 1) is not included due to a technical failure. In general, all groups hit most of the 18 waypoints on the route, and took around an hour to complete the entire treasure hunt. There was no significant correlation between these two measures of task performance: $r = 0.04, p > 0.2$ (using Pearson correlation).

To test the internal consistency of the subjective responses, we first computed Cronbach's Alpha on each of the three question groups. For the first three questions about the robot, $\alpha = 0.62$; for the questions about the treasure hunt, $\alpha = 0.52$; while for the Godspeed-style questions about the virtual robot, $\alpha = 0.83$. Based on these results, we have combined the Godspeed responses by averaging them into a single measure for further analysis, but have considered the responses to the remaining questions individually. The subjective responses are summarised in Table 2.

Table 2: Overall subjective results

Question	Mean	Median	Max	Min
1	4.1	4	5	3
2	3.8	4	5	2
3	3.9	4	5	2
4	4.1	4	5	2
5	4.0	4	5	3
6	3.9	4	5	2
7	4.6	5	5	3
8–14	4.0	4	5	3.1

As can be seen in Table 2, the overall subjective responses tended to be quite high. Indeed, no student ever chose the lowest score, while the majority of the scores were 3 or above. This suggests that at age 11–12, our participants may have been slightly too young to make the fullest possible use of the Smileyometer, even with our modifications: scores with this type of instrument have been found to decrease and diversify with age [24], and other studies have found that age 11 or 12 seems to be the critical turning point [23].

Although we used a robot agent with two different embodiments—physical for the introduction (Figure 4a) and virtual during the study (Figure 3)—we did not directly test whether the students considered the two agents to be the same, as previous studies of such agent migration (e.g., [1]) indicate that people generally consider this to be the case. However, we did repeat the three questions from the first part of the questionnaire, which asked students to assess the robot before the treasure hunt, in the final section where the students assessed it during the treasure hunt. In all cases, there was a significant (Pearson) correlation between the responses to the question pairs: if a student liked the physical robot, they tended also to like the virtual robot, and vice versa:

Questions 1&8 (Friendliness) $r = 0.51, p < 0.0005$
Questions 2&9 (Understandability) $r = 0.42, p < 0.001$
Questions 3&14 (Liking) $r = 0.58, p < 0.0001$

The influence of route and condition To test the influence of the two experimental manipulations on the above results, we used a two-way ANOVA analysis to determine the significant factors, and then used two-way post-hoc Mann-Whitney *U* tests to assess the influence of the factors. The main findings of the ANOVA analysis were as follows:

- The route had a marginally significant effect on the overall duration of the interaction ($F(1, 11) = 3.85, p \approx 0.08$).
- The interaction level had a significant impact on responses to Question 5, which assessed how easy the instructions were to follow ($F(2, 30) = 5.49, p < 0.01$).
- Both the route and the interaction level had a significant impact on responses to Question 6, which assessed how easy the questions were to answer. Interaction level: $F(2, 30) = 3.60, p < 0.05$; route: $F(1, 30) = 10.65, p < 0.001$; no significant interaction ($F(2, 30) = 0.21, p > 0.8$).

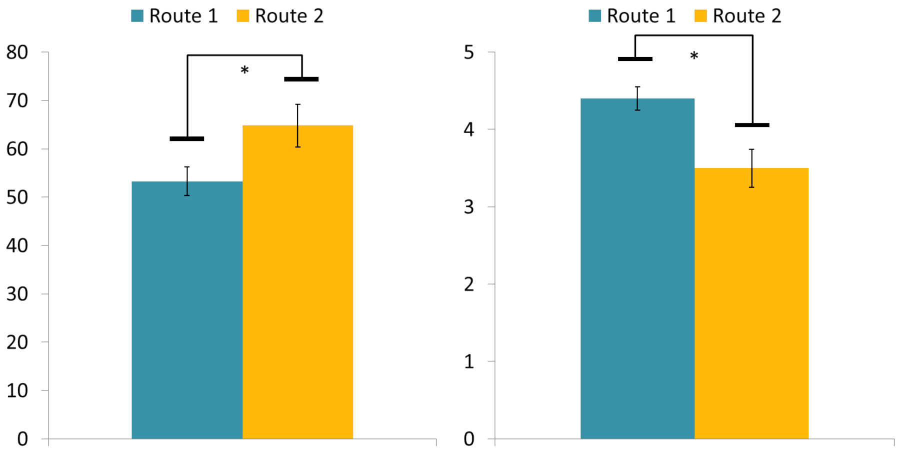
In the post-hoc tests, we found that participants who followed Route 1 completed the hunt significantly more quickly than Route 2 participants, and also rated the questions as significantly easier (both $p < 0.05$; see Figure 7). Also, the students who used the paper treasure hunt rated the instructions as easier than all of the tablet users (Figure 8a), while the students who received affective feedback from the robot rated the questions as harder than students in the other two groups (Figure 8b).

3.5 Discussion

The results of this study suggest that the interaction level had no influence on the students' objective performance on the treasure hunt; the main factor affecting performance was in fact the choice of route, and and teachers later confirmed later that Route 2 had fewer directional signs and was therefore known to be more challenging.

On the other hand, the choice of interaction level did have some significant effects on the responses to the subjective questionnaire, in particular on the two items that measured the perceived difficulty of the treasure-hunt task. The participants who used the original, paper-based questionnaire found the instructions significantly easier to follow than did any of the tablet participants. This is likely because the paper presentation allowed the students to access more context during the treasure hunt by looking ahead at the instructions. Note that increased context has also been found to improve instruction-following performance in other settings, such as human-robot joint action [10].

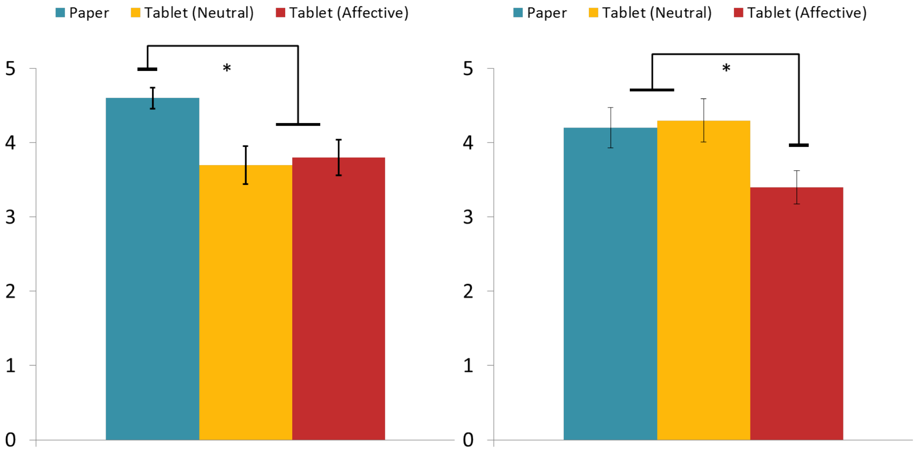
The other effect—where the participants who received affective robot feedback found the questions harder to answer than did those who received neutral feedback—is more difficult to explain. Note also that the reported difficulty from the students who received neutral feedback was not distinguishable from that reported by the paper-based participants. One possible explanation is that the affective feedback on incorrect answers (e.g., "Too bad") magnified their impact, while the affective feedback on correct answers ("Fantastic!") somehow made the participants feel like answering correctly was more of an accomplishment. However, since the actual question-answering performance of the different groups is not available, it is difficult to draw any definite conclusions from this effect.



(a) Time (minutes)

(b) Question ease

Fig. 7: The influence of route



(a) Instruction ease

(b) Question ease

Fig. 8: The influence of feedback strategy

4 Summary and future work

We have implemented a tablet-based app including a virtual robot that is designed to be used by 11-12 year old school children in the context of a real-world treasure hunt. We have carried out a user study comparing the tablet version to the original paper-based version, in terms of both objective task success and subjective user impressions. The participants in the study found the instructions easier to follow on paper than with the app; this is likely due to the increased context provided by the static paper presentation. The study participants who received affective feedback from the robot about the correctness of their answers perceived the questions to be more difficult than did the participants who received neutral feedback. We hypothesised that this may have been due to the affective feedback making the questions appear more difficult.

In the short term, we will re-run the treasure hunt study in order to obtain more accurate objective task success results, and with all students using the tablet and following the same route. This should allow the impact of feedback to be assessed more fully, and should also permit an informative investigation of the relationship between the objective and subjective measures [26]. We will also carry out additional qualitative analysis such as interviews to get a better sense of the students' opinion of the agent. On the technical side, we will update the app to incorporate more of the context provided by the paper version, which should make it easier to follow the treasure hunt instructions. Finally, we will update the robot feedback strategies in consultation with teachers and psychologists in order to make the learning experience more effective.

More generally, this study shows that introducing a robot agent may adversely affect the perceived ease of an educational task, and thus as a consequence may affect morale and the overall student experience. In particular, it clearly demonstrates that adding a conversational agent to an interface is not a simple modification: the agent's behaviour may have an unanticipated (possibly negative) effect on user experience. In other recent work, [16] found that, whilst the presence of a robot can improve learning gain, this improvement is lost when the robot is "social", using affective responses, gestures and personalisation. The authors speculate that the affective robot may be a distraction and is viewed more as a teacher in the non-social case, and warn that applying social behaviour to a robot in a tutoring context may have negative effects. Based on this study, we make a similar warning, particularly with respect to the ethics of developing educational applications for vulnerable groups such as children and people with learning difficulties. In the context of the larger EMOTE research project—which has the overall goal of developing empathic robot tutors—we will apply the findings from this study to the other agents being developed in the project, taking care to ensure that any affective feedback from the agents has the intended effect on the overall pedagogical goals.

Acknowledgements

This work was partially supported by the European Commission (EC) and was funded by the EU FP7 ICT-317923 project EMOTE. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein. We thank the pupils and teachers at Edinburgh Academy for their help in carrying out this study.

References

[1] Aylett, R., Kriegel, M., Wallace, I., Segura, E., Mercurio, J., Nylander, S.: Memory and the design of migrating virtual agents. In: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems. pp. 1311–1312. AAMAS '13 (2013)

[2] Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics 1, 71–81 (2009)

[3] Bates, J.: The role of emotion in believable agents. Communications of the ACM 37(7), 122–125 (Jul 1994)

[4] Beale, R., Creed, C.: Affective interaction: How emotional agents affect users. Human-Computer Studies 67, 755–776 (2009)

[5] Bickmore, T., Picard, R.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer Human Interaction (TOCHI) 12(2), 193–327 (2005)

[6] Damasio, A.: Descartes' Error: Emotion, Reason and the Human Brain. Gosset/Putnam Press, New York (1994)

[7] Dehn, D., Van Mulken, S.: The impact of animated interface agents: a review of empirical research. International Journal of Human Computer Studies 52(1), 1–22 (2000)

[8] Dias, J., Paiva, A.: Feeling and reasoning: A computational model for emotional agents. In: 12th Portuguese Conference on Artificial Intelligence (EPIA 2005). pp. 127–140. Springer, Portugal (2005)

[9] D'mello, S., Graesser, A.: AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems 2(4) (2012)

[10] Foster, M.E., Giuliani, M., Isard, A., Matheson, C., Oberlander, J., Knoll, A.: Evaluating description and reference strategies in a cooperative human-robot dialogue system. In: Proceedings of IJCAI 2009. Pasadena, California (Jul 2009)

[11] Gong, L.: Is happy better than sad even if they are both non-adaptive? effects of emotional expressions of talking-head interface agents. International Journal of Human Computer Studies 65(3), 183–191 (2007)

[12] Gratch, J., Marsella, S.: A domain-independent framework for modeling emotion. Journal of Cognitive Systems Research 5(4), 269–306 (2004)

[13] Harley, J.M., Bouchet, F., Azevedo, R.: Aligning and comparing data on emotions experienced during learning with MetaTutor. In: Artificial Intelligence in Education. Springer Berlin Heidelberg (2013)

[14] Isbister, K.: Better Game Characters by Design: A Psychological Approach. Morgan Kaufmann (2006)

[15] Kędzierski, J., Muszyńki, R., Zoll, C., Oleksy, A., Frontkiewicz, M.: Emys – emotive head of a social robot. International Journal of Social Robotics 5(2), 237–249 (2013)

[16] Kennedy, J., Baxter, P., Belpaeme, T.: The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. pp. 67–74 (2015)

[17] Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings. In: Proceedings of HRI 2012. pp. 367–374 (2012)

[18] Lim, M.Y.: Emotions, Behaviour and Belief Regulation in An Intelligent Guide with Attitude. Ph.D. thesis, School of Mathematical and Computer Sciences, Heriot-Watt University, Ediburgh, Edinburgh (2007)

[19] Maldonado, H., Lee, J., Brave, S., Nass, C., Nakajima, H., Yamada, R., Iwamura, K., Morishima, Y.: We learn better together: enhancing elearning with emotional characters. In: Koschmann, T., Suthers, D., Chan, T. (eds.) Computer Supported Collaborative Learning 2005: The Next 10 Years!, pp. 408–417 (2005)

[20] Okonkwo, C., Vassileva, J.: Affective pedagogical agents and user persuasion. In: Stephanidis, C. (ed.) Proceedings of the 4th International Conference on Universal Access in Human Computer Interaction. pp. 5–10. Springer, Beijing, China (2001)

[21] Picard, R.W.: Affective Computing. MIT Press (1997)

[22] Prendinger, H., Mayer, S., Mori, J., Ishizuka, M.: Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. In: Fourth International Working Conference On Intelligent Virtual Agents (IVA 03). pp. 283–291 (2003)

[23] PuppyIR Project: Fun toolkit: The smileyometer and again-again table. http://hmi.ewi.utwente.nl/puppyir/results/user-evaluation-toolkit/fun-toolkit-the-smileyometer-and-again-again-table/, accessed 16/09/2015

[24] Read, J.C., MacFarlane, S.: Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In: Proceedings of the 2006 Conference on Interaction Design and Children. pp. 81–88. IDC '06 (2006)

[25] Read, J., Macfarlane, S.: Endurability, engagement and expectations: Measuring children's fun. In: Proceedings of the 2002 Conference on Interaction Design and Children. pp. 189–198. IDC '02 (2002)

[26] Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. Natural Language Engineering 6(3&4), 363–377 (2000)

A Subjective questionnaire

The following are the questions that were included in the subjective questionnaire for the treasure hunt study (Section 3). All were presented using a five-point Smileyometer (Figure 6). Students who used the paper-based treasure hunt answered questions 1–7 only, while students in the two tablet conditions also answered questions 8–14.

Questions about Susie before the treasure hunt

1. Susie was: Unfriendly … Friendly
2. Susie was: Hard to understand … Easy to understand
3. I liked Susie: Not at all … A lot

Questions about the treasure hunt

4. The treasure hunt was: No fun at all … Lots of fun
5. The instructions were: Hard to follow … Easy to follow
6. The questions were: Hard to answer … Easy to answer
7. I think my group did: Very badly … Very well

Questions about Susie during the treasure hunt

8. Susie was: Unfriendly … Friendly
9. Susie was: Hard to understand … Easy to understand
10. Susie was: Unkind … Kind
11. Susie was: Unpleasant … Pleasant
12. Susie was: Awful … Nice
13. Susie was: Not helpful … Helpful
14. I liked Susie: Not at all … A lot