Supporting and challenging learners through pedagogical agents: Addressing ethical issues through designing for values

Pedagogical Agents (PAs) that would guide interactions in intelligent learning environments were envisioned two decades ago (Johnson, Rickel, & Lester, 2000). These early animated characters had been shown to increase positive perceptions of the learning experience, perceived credibility of the task and motivation for the activity, known as the persona effect, leading to learning benefits (Lester et al., 1997). However, little was understood regarding what aspects were beneficial for learning and what sort of learning PAs were suitable for (Johnson & Lester, 2018). This article will consider the current and future use of PAs to support and challenge learners from three perspectives. Firstly, we will look at PAs from a *practical* perspective to consider what PAs are, the roles they play in education and beyond and the underlying technologies and theories driving them. Next we take a *pedagogical* perspective to consider the vision, pedagogical approaches supported and new possible uses of PAs. This leads us to the *political* perspective to consider the values, ethics and societal impacts of PAs. Drawing all three perspectives together we will present a design for values approach to designing ethical and socially responsible PAs.

Practical: What are Pedagogical Agents, how are they used and created?

We distinguish PAs from virtual characters or avatars where a graphical representation is controlled by a human, such as those we find in multi-user virtual environments (MUVEs) and virtual worlds created using Second Life or OpenSim for purposes such as practicing negotiation skills between physically remote law students (Butler, 2008). Here we restrict PAs to mean virtual characters driven by software that use rules and agent technologies to guide the reasoning and/or behaviours of the avatar or virtual character. Thus, a PA must control its own reasoning and behaviour, not a human. An agent is a piece of software that typically has its own goals, beliefs and plans, commonly according to the belief, desires (goals) and intentions (plans) (BDI) model (Rao & Georgeff, 1991) widely used by agent researchers. This article concerns a subset of the field of agent technology, known as Intelligent Virtual Agents (IVAs), in which the agents are embodied, typically, but not always, representing a humanlike character. IVA researchers describe their technology using terms that reflect the IVA's main behaviour, purpose or application domain. Hence they are known as Embodied Conversational Agents when dialogue is the focus, Relational Agents in supportive medical or health applications and Pedagogical Agents in the education domain. Sometimes the role the IVA is playing is the main focus and thus the IVA would be referred to as a Virtual Human, Virtual Nurse, Virtual Patient, Virtual or Intelligent Tutor, Virtual Coach or simply given the title of a Peer Learner, Learning Companion, Teachable agent and so on. The term "Virtual Human" is commonly used in social training simulations involving scenarios and role-plays, for example to gain interpersonal skills (Schmid Mast, Kleinlogel, Tur, & Bachmann, 2018). A medical student might interact with a Virtual Patient or Virtual Nurse as part of their training (e.g. (Hubal, Kizakevich, Guinn, Merino, & West, 2000), see further medical applications below). A Virtual Coach might educate the learner about good behaviours (e.g. (Grolleman, van Dijk, Nijholt, & van Emst, 2006)).

Johnson and Lester (2018) discuss a number of current PA roles such as virtual role-players, learning companions or peers as well as less common roles to support navigation within a virtual world, provide an interactive demonstration of a skill. Importantly, they note that learning can be negatively impacted if the role of the agent is not clear. The role of the PA will influence the nature of the relationship developed between the learner and the PA.

Walker and Ogan (2016) stress the importance of carefully designing that relationship; clearly important in applications such as the use of virtual and robotic PAs to train people with autism to learn social skills (Beaumont, Rotolone, & Sofronoff, 2015; Tanaka, Negoro, Iwasaka, & Nakamura, 2017).

Outside school and university environments we see the use of PAs to aid senior citizens conduct e-commerce transactions (Chattaraman, Kwon, Gilbert, & In Shim, 2011); for cultural sensitivity training (Swartout et al., 2006) and language learning (Johnson, Friedland, Schrider, Valente, & Sheridan, 2011) in the military and for job interview training in the hospitality industry (Muralidhar et al., 2016). The most varied and extensive use of PAs can be found in the health and wellbeing domain. In these contexts, PAs have been used for instruction, education and training where the patient, trainee nurse, trainee doctor or the public in general are the targeted learners. Medical education often involves PAs in the role of virtual patients (Carnell, Halan, Crary, Madhavan, & Lok, 2015; Danforth, Procter, Chen, Johnson, & Heller, 2009) to allow medical students to practice breaking bad news in oncology (Berney et al., 2017); teach empathy for patients (Halan, Sia, Crary, & Lok, 2015); train clinicians to collect family histories (Wang et al., 2015) and improve patient assessment and interpersonal communication skills (Rizzo, Kenny, & Parsons, 2011). In a few cases, PAs have been created to provide education and support to both the patient and the health professional, as in the study by Kowatsch et al. (2017) on childhood obesity.

PAs can vary widely in their appearance, intelligence, capabilities and purpose. The seminal article on animated pedagogical agents (Johnson et al., 2000) described key capabilities of PAs and provided examples of PAs ranging from Herman the Bug to teach school children about plants to STEVE who demonstrates to trainees how to operate equipment. Since the early work of Disney animators, it has become apparent that visual fidelity or realism of the character's appearance has not been an important feature for IVAs (Richards & Szilas, 2012). Of greater concern has been the achievement of IVA *believability*, a plausible model that does not detract from overall focus (Norling, 2009). Believability is closely connected with perceived social ability. IVAs exhibit a range of social capabilities and a continuum of intelligence ranging from fully scripted to those driven by cognitive agent architectures and emotion appraisal systems, such as FAtiMA (Fearnot Affective Mind Architecture)(Dias, Mascarenhas, & Paiva, 2014).

A number of articles providing the state of the art of Pedagogical Agents in education have been published in recent years (see, Johnson and Lester (2018), Kim and Baylor (2016), Schroeder, Adesope, and Gilbert (2013) and Sottilare and Hart (2012)). The meta-analysis undertaken by Schroeder et al. (2013) reviewed 43 studies covering agents represented as stick figures to full-bodied virtual humans and learner populations ranging from K-12 students, university students, trainees in business, defence, intelligence, and medical education. They note that statistically significant findings were reported when learning was measured after interventions involving agents compared with learning environments with no agents. They also found some subject domains, such as Mathematics and Science, reported greater learning benefits compared to subjects from the humanities. Perhaps not surprisingly given the game-like appearance of many systems, the study found that K-12 learners benefited more from using PAs than older learners. Kim, Baylor, and Group (2006) found that for adult education it was best for a PAs to act as peer learners. In a study with 20 older adults using a learning companion, Davies and Eynon (2013) found that more academically self-confident participants were less willing to engage in conversations with PAs compared with those who were less academically self-confident. Gender and cultural differences have

also been found. In general and across age groups, females appeared more open and favourable to using PAs than males (Arroyo, Murray, Woolf, & Beal, 2003; Kim et al., 2006). Other studies show that females preferred young and cool agents over older and uncool agents with more expertise, even though significant learning differences based on gender were not found (Rosenberg-Kima, Baylor, Plant, & Doerr, 2008), and females favoured learning companions that sought to build a social relationship over ones that were just focussed on the task (Haake & Gulz, 2009). From an ethnicity perspective, high-school students (Plant, Baylor, Doerr, & Rosenberg-Kima, 2009) and college-aged students of color (Moreno & Flowerday, 2006) preferred PAs that represented a similar ethnicity to themselves over human tutors. These findings seem to indicate that those who feel marginalized prefer to deal with an artificial person over a real person that they do not identify with. This suggestion is supported by a study with middle grade students involving learning algebraic concepts that found, in contrast to white males, females and ethnic minorities reported ease of learning with the PA and significantly improved their self-efficacy and attitude towards algebra (Kim and Lim 2013). Clearly, the teaching role played by the PA must be carefully aligned to the learner population.

PAs utilise a range of technologies including intelligent tutoring systems, agent cognitive architectures that drive the reasoning of the agent, games technology and affective technologies that detect and express emotion. Gaming elements are used to improve motivation (Richards & Caldwell, 2017b); while games technology, such as the Unity3D game engine, is used for development and rendering of avatars and the virtual environments they inhabit (Bouvier, Sehaba, & Lavoué, 2014). Harmon (2016) is drawn to the use of games technology and virtual reality as a means to allow people to interact with artificial humans as they evaluate social dilemmas and consider alternatives in decision-making. Affective technologies involving multimodal inputs are used to persuade learners to engage in computer-based training (Bosch et al., 2015) and detect the learner's affective states (Gwo-Dong et al., 2012). Affective technologies are underutilised currently in education applications. For example, the detection of epistemic emotions such as engaged, bored, frustrated and confused is a relatively new and under-researched field within the broader and more established AI fields of computer vision and emotion recognition (Nezami & Richards, 2017; Nezami, Richards, & Hamey, 2017).

Theories from cognitive science, educational psychology and the learning sciences have underpinned PA research. Much of the work using IVAs in scenarios and simulations involves concepts from narrative, story-telling and drama management (Richards & Szilas, 2012). The focus on believability and emotion within IVA research was initially motivated by the findings of animators from Disney and continues to be driven by theories of human behaviour and media (e.g. Hoorn et al. (2004)). In the early days of PAs, Kim et al. (2006) were pioneers in examining the effect of specific PA design features on specific learning outcomes. They drew on the work of (Bandura, 1986; Rosalind Wright Picard, 1995; Rosalind W Picard, 1997; Spence, 1995) to provide a social-cognitive lens to analyse the effect of learners' self-efficacy and affective experience on cognitive engagement and deep learning.

In this section we have considered what a PA is, what roles it can play in a range of learning applications, responses of certain learner populations to PAs, what technology is used to create PAs, emerging related technologies and the theories driving their development.

Pedagogy: Vision and New Possibilities.

In a recent review of how AI can reform education, PAs were envisioned to "Give every learner their own personal tutor, in every subject; Provide every teacher with their own AI teaching assistant; Lifelong learning companions to advise, recommend, and track learning" (Luckin, Holmes, Griffiths, & Forcier, 2016) p. 47. Johnson and Lester (2018), p. 33-34 echo and expand part of this vision:

> One can imagine a future in which every learner has her own pedagogical agent—or perhaps a cast of pedagogical agents—that accompanies her from the time she is young through adulthood and on into senescence. Her agents could provide highly customized support, delivered ubiquitously in all of her activities …..with the lines between education and training blurring and eventually disappearing altogether.

Both of these visions recognise that each learner is an individual with their own preferences and personal needs for support. Despite the personalised assistance that PAs could provide to teachers and students, we rarely find PAs in today's physical, virtual or informal classrooms. We are even less likely to find PAs that offer individualised and context-specific support due to limited data, understanding of individual differences and their impact on student learning {Makhija, 2018 #303} a lack of student specific data and models of individual learning are still much . However, current IVA research and applications in the health domain suggest, in line with the above visions, that PAs have the potential to blur the distinction not only between learning and training but between learning and all human activities. Reinforcement and reframing of the educational role of PAs are potentially in order.

PAs support a social constructivist view of learning, where the learning process is both an individual and social activity involving artefacts but also other people (Greeno, Collins, & Resnick, 1996). We suggest that the goal and role of PAs has less to do with management and delivery of content but rather their function should be to provide social support and social training. For example, in school contexts, PAs could be used for social training relating to learning empathy. An elementary school study found learning through observation of dilemmas being played out can invoke empathy for one or more of the parties (Upright, 2002). Using IVA technology, the FearNot! Project allowed students to explore alternative strategies to cope with bullying scenarios. The project used the FAtiMA cognitive agent architecture (Dias et al., 2014) to implement the OCC emotion appraisal model (Ortony, Clore, & Collins, 1990). This work was inspired by earlier work, known as Carmen's Bright Idea, using an approach called Interactive Pedagogical Drama, where parents of children with cancer encounter IVAs enacting potential social dilemmas, such as whether/how to tell the neighbour and relatives about the child's condition (Marsella, Johnson, & LaBore, 2003). As described, "The goal of Interactive Pedagogical Drama (IPD) is to exploit the edifying power of story while promoting active learning. An IPD immerses the learner in an engaging, evocative story where she interacts with realistic characters. The learner makes decisions or takes actions on behalf of a character in the story, and sees the consequences of her decisions" p. 1. Active participation in a dilemma allows the individual to engage with the narrative in a way that is less constrained by bias or stereotyping (Kemmer, 2014). In the Orient Project (Aylett et al., 2009), the FAtiMA architecture was further extended to allow modeling of cultural norms using Hofstede (2011) cultural dimensions, to provide scenarios to evoke empathy for migrant/refugee children amongst school children in the UK and Germany.

The "active learning" that Marsella et al. (2003) describe as the goal of IPD, in contrast to learning theories that emphasize learning by doing, does not mean that active participation in

the drama is required. Most IVA applications use the popular "first-person shooter" game genre where players do not see themselves in the game; instead they look at the environment and others as in real life a person looks and interacts with the world in the first person. Unlike a role-playing game, by design in Carmen's Bright Ideas, FearNot! and Orient the learner is an observer, not a participant. The learning approach seeks to encourage reflective learning and the development of tacit, rather than declarative or procedural knowledge. In line with this approach, PAs have been used to significantly improve episodic memory by using a reminiscing agent to debrief the student following exploration of a virtual world (Nicholas, Van Bergen, & Richards, 2015). Similar to the use of IVAs to assist the elderly to reminisce (Nikitina, Callaioli, & Baez, 2018), PAs could help students with remembering and act as a form of external memory aid that does more than store or retrieve information. Another direction to encourage reflection is the use of teachable agents where the human (co)-learner teaches the agent what to do or provides the agent with answers to its questions (Pareto, 2014).

Hoorn et al. (2004) remark that while studies have found that fictional characters can be annoying and distracting and that their use to deliver information does not necessarily improve recall or comprehension; the value of fictional characters is their ability to improve motivation to engage with the content. Beyond engagement and motivation we also see their potential for behaviour change (Kowatsch et al., 2017; Lisetti et al., 2012), empowerment (Richards & Caldwell, 2017a) and to deliver education and contact strategies to change college students' stigmatised attitudes to mental illness in their peers (Sebastian & Richards, 2017). IVAs provide a conversational humanlike way of delivering information that overcomes literacy barriers (Bickmore et al., 2010).

This section began with visions conceived for PAs and explored innovative roles and ways that PAs can enhance learning including for social skills training, decision-making, reflection, empowerment and as a humanlike means to increase engagement and motivation.

Politics: values, ethics, societal impacts

The extent to which various technologies are adopted and supported in our educational institutions is often politically driven or at least constrained. We might have expected by now that technology would have disrupted learning and teaching in our classrooms following the launch of grand sounding initiatives such as the "Digital Education Revolution" (Buchanan, 2011) in Australia. However, on closer review we see that its aims were to 1) provide all students with computers and schools with access to networks and digital resources and 2) equip teachers, students and parents to use digital technologies. These goals are not so ambitious and do not encompass state of the art technology or new pedagogies. Similarly, two decades on from the promised "Classroom of the 21st Century" envisioned in the previous century, we see little use of advanced technologies in the classroom or even extensive facilitation of the classroom using digital technologies. To make matters worse, from our own experience, gaining access to classrooms to expose students and teachers to potentially disruptive EdTech is difficult even when the materials are co-designed with teachers to address concepts poorly grasped via existing pedagogies and align with the national curriculum (Jacobson, Taylor, & Richards, 2016).

It is not surprising in such climates that PAs are far from pervasive in current classrooms. This situation begs the question of why in contrast we see a proliferation of IVAs being used in many and varied health related learning contexts. Funding is probably the answer. Spending on health and health-related research in Western economies significantly outstrips spending on education. Health expenditure in Australia is around 10% of GDP and medical

research expenditure in the 2018-19 Australian Federal Government budget was promised $2Billion. In contrast, after the closure of the Office of Learning and Teaching in 2016, there has been no specific budget allocation for research in learning and teaching in the Australian Federal Government budget.

This draws us to consider the beliefs and values of our societies particularly concerning the priority given to education; what education, teaching and learning are understood to be; and what AI can contribute to a revised or enlightened view of these. To reach the vision of Johnson and Lester (2018) where the distinction between training and education are blurred and we become societies of lifelong learners, governments and societies should reevaluate the importance they place on education and broaden their view of education beyond K-12 curricula and higher education programs. What can assist with changing narrow-thinking that has led to limited use of AI, Robotics, XR (Virtual/Augmented/Mixed Reality), Games technology, Educational Virtual Worlds and PAs in classrooms is an understanding of how these technologies could transform not just education but also our society. Fears of AI replacing teachers are founded on beliefs that teaching is about instruction and presenting students with content. A useful place to start is to consider what these technologies might provide that current approaches, including human-delivered teaching, do not.

The SimSensei (DeVault et al., 2014) state of the art system that can detect the emotional state of the user and respond using verbal and non-verbal empathic cues has been used successfully with sufferers of Post Trauma Stress Disorder (PTSD). SimSensei was found to be preferred by patients over a human counsellor because patients believed Ellie the Virtual Interviewer was less judgemental than a human would be (Gratch, Lucas, King, & Morency, 2014). Are there parallels for use in our classrooms?

Lisetti (2012) outlines the following 10 advantages of using avatars over humans in patient-centered computer-based interventions for behavior change: increased accessibility, increased confidentiality and divulgation, tailored information, diminished variability, avoidance of righting reflex with infinite patience, addresses low literacy, lower attrition rates, allows patient-physician concordance/matching, provide working alliance and express empathy. Similarly, human teachers have limited accessibility and are not exempt from bias and fatigue leading to varied levels of empathy and patience toward different students. Reevaluating what teaching and learning fundamentally are and identifying current gaps and what is done well or poorly will help to inform and motivate reform and identify the possible role of AI within the complex learning landscape.

Before we commit to a new future where AI pervades learning, as educationalists and technologists we need to guide society and governments to enter this new world with eyes wide open concerning the social and ethical ramifications. Rather than worrying about PAs taking over the world (or at least the classroom), as in the technological singularity fear (Korotayev, 2018), the main concern should be humans' readiness to blindly accept the technology as a replacement. Human use of technology can include: misuse, an overreliance on technology; disuse, ignoring or underutilization; or abuse, use without concern for the consequences (Parasuraman & Riley, 1997). Early studies by psychologists of human behaviour towards machines revealed that humans tend to use the same human politeness strategies with machines (Reeves & Nass, 1996). IVAs are much more humanlike in appearance and behaviour than the software used in those early studies. Kim and Baylor (2016) also attribute the social, emotional and often unexpected responses by learners to PAs to the human tendency to treat media socially and anthropomorphise objects. From a social

and ethical standpoint, the first author believes the greatest threat posed by AI in general, and IVAs and PAs in particular, is their overuse, which goes beyond misuse, and overreliance if humans allow technology to replace, rather than mediate, human relationships. It is one thing to use relational agents, virtual learning companions, mentors, coaches or confidantes, in contexts that a human is not available or suitable, but if students chose not to speak with their classmates, teachers or parents because they prefer the company of a machine, that is a concern. To highlight the concerns, consider the following quotes from the book "Close engagements with artificial companions" (Peltu & Wilks, 2008) "There is openness to seeing computational objects as 'other minds'; there is willingness to consider what a computer and human mind have in common; and, in a different register, there is evidence of a certain fatigue with the difficulties of dealing with people." p.27 "The question is not whether children will grow up to love their robots more than other toys, or indeed, their parents, but what will loving come to mean?"p.30. Classrooms, and life, are full of social relationships. Learning is a social activity (Vygotsky, 1980) and making technology more social could improve learning. McLaren, DeLeeuw, and Mayer (2011) found that when polite language is used by a web system, learners learnt more. But that does not mean the web system should replace the humans in that learner's life. While dealing with humans can be difficult, this is a life skill to be learnt, not avoided. Similarly, learners will need to gain life skills on how to understand and deal with technology, including humanlike technology. A healthy society will be a society of people who know and enforce their boundaries with technology. Education will play an important role in teaching society to establish those boundaries.

Roxas, Richards, Bilgin, and Hanna (2018) found that just watching a faceless dancing IVA was able to evoke and change human emotions. Similarly, a virtual coach seeks behaviour change by using empathic, motivational and persuasive strategies that appeal to the trainee's emotion. PAs could be used in exposure therapies, like the use of virtual reality to treat PTSD (DeVault et al., 2014), to gain cultural awareness (Aylett et al., 2009; Johnson et al., 2011), build resilience or overcome fears and anxieties. But do we have safeguards to ensure the student is not pushed too far? Or will exposure lead to desensitisation rather than dealing with the problem or inducement of empathy. Just as exposure therapy should only be delivered by a trained therapist, the use of PAs in such circumstances needs to be under the careful control of a trained instructor. This philosophy was adopted in the eADVICE website that provides medical treatment advice and the opportunity to discuss your treatment with the IVA known as Dr Evie while patients are on the hospital waiting list and under the care of their referring GP (Richards & Caldwell, 2017c). Concerning social relationships between PAs and the learner, Walker and Ogan (2016) ask "Is it acceptable if technology lies to students? If it is purposefully manipulative? Is it the designer's responsibility to avoid encouraging students to get too involved with the technology?" p.726.

Current awareness of ethical issues relating to AI and Learning Analytics are mostly restricted to privacy, security and appropriate uses of personal data. Luckin et al. (2016) raise the issue of a virtual teaching assistant who monitors the teacher's performance. Learning analytics poses a similar concern. From an educational perspective, Luckin et al. (2016) also warn of a learning companion that remembers and reminds the student of their past failures, making it difficult for them to experience or believe in success. Indeed a study showed worse results with the empathic version of an agent that reminded students about what they had learnt (Hastie et al., 2016). Parents will want to protect the personal data of their children, but may also want access to that data themselves, while children and young people may want privacy from their parents. Dealing with these conflicts of interest and protection of the rights of vulnerable people such as children and the mentally disabled, will need carefully designed

standards and legislation. Further discussion of data privacy will be left to other articles in this special issue whose focus is on learning analytics (Johanes and Thille, 2019; Kitto and Knight, 2019; Tsai et al., 2019; Prinsloo, 2019; Williamson, 2019).

Hudlicka (2016) identifies several ethical issues that go beyond the general concerns of data privacy and which are specific to virtual agents. These concerns include affective privacy (the right to keep your thoughts and emotions to yourself), emotion induction (changing how someone feels), and virtual relationships (where the human enters a relationship with the agent). A visit to the gym, therapist or supermarket, or a quiet afternoon at home, could produce psychometric, physiological, financial, emotional and social information that can be used to build an affective user model. Such intimate data could provide personalised and appropriate responses (Richards, 2017). However, sharing of user models that capture our inner thoughts and feelings could potentially impact that individual if revealed to their employer, family, friends or wider public. When we consider the context of learning and education, such monitoring on the one hand could be seen as a way to identify and handle bullying. However, in the wrong hands, public or targeted distribution of this private information would be devastating and exacerbate the bullying problem. Is the potential risk greater than the potential benefit?

Herein lies our predicament. As discussed in this section, as we add more social capabilities into our educational technologies we add more potential social and ethical issues. Can we design technology that includes our values towards building socially responsible and ethical AI systems, and PAs in particular?

Ethical PAs: Design for Values

In order to design PAs that are sensitive to moral principles and human values, methods are needed that identify and implement these values in an open, participatory and responsible manner. To ensure *openness*, the purpose and motives that led to the development of the PA needs to be clear and explicit, enabling enquiry and the possibility to understand positioning in a given context. PA systems will be taking decisions which we would consider to have an ethical flavour if they were made by people. Means are needed to support the design of such capabilities. This requires models and algorithms to represent and reason about, and take decisions based on, human values, and to justify their decisions according to their effect on those values. Where it concerns *participation*, it is necessary to understand the possible roles of PAs and how different people work with and live with AI technologies across cultures. In fact, the use of PAs must be understood as part of a socio-technical environment. Finally, responsibility includes the political context of the PA. This is an issue of regulation and legislation, but also of codes of conduct and of education. It is up to society to determine how issues of liability should be regulated. For example, who will be to blame if a PA suggests an incorrect or inappropriate action to the user? The builder of the hardware (e.g. of the sensors used by the PA to perceive the environment)? The builder of the software that enables the PA to decide on a course of action? The authorities that allow the use of the PA? The user that personalized the PA's settings to meet her preferences? All these, and more, questions must be informing the regulations that societies put in place towards responsible use of such systems.

PAs, being AI systems, are characterized by their Autonomy, Interactivity and Adaptability (Dignum, 2019; Floridi & Sanders, 2004; Russell & Norvig, 2016). These properties enable agents to deal effectively with their environments, which are characterised by dynamic spatial and temporal variation and unpredictability, and which lead to the frequent occurrence of

previously unexperienced situations for the agents that interact with them. However, an interactive system that is autonomous and adaptable is hard to verify and predict which can lead to unexpected activity or undesirable behaviour. Currently, many organisations and nations are proposing sets of ethical principles for AI to reflect societal concerns about the ethics of AI, and ensure that AI systems are developed responsibly, incorporating social and ethical values. For the purpose of this paper, we use the ART principles: Accountability, Responsibility and Transparency (Dignum, 2017), that can be seen as a summary of most of these guidelines, that can be directly linked to the characteristics of AI systems mentioned above. Even though, many machine learning algorithms used in PAs are designed to maximize predictive accuracy, a similar approach can be taken by encoding ethical principles as objective functions. Even though this is an open research question, examples can be already be seen in the area of algorothmic fairness (Corbett-Davies & Goel, 2018).

Accountability refers to the requirement to explain and justify decisions and actions to one's interaction partners, i.e. users and others with whom the system interacts. To ensure accountability, decisions must be derivable from, and explained by, the decision-making algorithms used. This includes the need for representation of the moral values and societal norms holding in the context of operation, which the agent uses for deliberation. Accountability in AI requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values). AI explanation is necessary to ensure trust (Lyons, 2013; Theodorou, Wortham, & Bryson, 2017) and is required by EU data protection laws. Explanation should be grounded in moral and social concepts, including values, social norms and relationships, commitments, habits, motives and goals (Miller, 2018).

Responsibility refers to the role of people themselves, and to the capability of AI systems to answer for one's decision and to identify errors or unexpected results. As the chain of responsibility grows, means are needed to link the AI systems' decisions to the fair use of data and to the actions of stakeholders involved in the system's decision. Responsibility is not just about making rules to govern intelligent machines – we also need to consider how we regulate the data they create and share. It is about ensuring the proper means to respond to the novel, data-driven reality, with a constructive operational framework that is able to inform sustainable new modes of human-agent interaction.

Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data created. Many current AI algorithms, in particular those based on neural network models, are basically black boxes. However, regulators and users demand explanation and clarity about the data used. Methods are needed to inspect algorithms and their results and to manage data, their provenance and their dynamics. Transparency requires proper treatment of the learning process along with approaches to remove the so-called algorithmic black box. Trust in the system will improve if we can ensure openness of affairs in all that is related to the system. For example, being explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required from all models that use human data or affect human beings or can have other morally significant impacts. By transparency, we mean that the different factors that influence the decisions made by algorithms should be visible, or transparent, to the people who use, regulate, and are impacted by systems that employ those algorithms, i.e. the understandability of a specific model (Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018). Transparency is also seen as, and is often taken as, a requisite for algorithmic accountability (Bryson & Winfield, 2017). However, decisions made by machine learning algorithms can

be opaque due to many factors, which may not always be possible or even desirable to eliminate. These include issues of a technical nature (the algorithm may not lend itself to easy explanation), but also economic (the cost of providing transparency may be excessive, and may include the compromise of intellectual capital aspects), and social (revealing input may violate privacy expectations)[1] (Ananny & Crawford, 2018). Since human decisions can be also quite opaque, as are the decisions made by corporations and organisations, mechanisms such as audits, contracts, and monitoring are in place to regulate and ensure attribution of accountability.

A *Design for Values* approach to AI models ensures that these principles are analysed and reported at all stages of system development. Alongside these requirements, we need to rethink the optimization criteria for Machine Learning. Demanding a focus on the observance of ethical principles and putting human values at the core of system design, calls for a mind-shift of researchers and developers towards the goal of improving transparency rather than performance, which will lead to a new generation of PAs aligned with human values. A Design for Values approach provides guidelines on how AI applications should be designed, managed and deployed, such that values can be identified and incorporated explicitly into the design and implementation processes:

- Identify the relevant stakeholders;
- Elicit values and requirements of all stakeholders;
- Provide means to aggregate the values and value interpretations from all stakeholders;
- Maintain explicit formal links between values, norms and system functionalities that enable adaptation of the system to evolving perceptions and to justify implementation decisions in terms of their underlying values;
- Provide support to choose system components based on their underlying societal and ethical conceptions, in particular when these components are built or maintained by different organisations, holding potentially different values.

In the context of PAs, the framework offered by (Luckin and Cukorava, 2019) could be used to ensure the values captured and technology delivered have clear educational benefits determined by inter-stakeholder and inter-disciplinary partnerships.

Conclusion and Future of PAs

This article presents a wide range of possible roles of PAs that go beyond classroom contexts and that provide education beyond subject matter. They have the potential to be pervasive throughout one's many learning situations. Rather than replacing human teachers and human interactions, PAs could teach individuals and groups how to learn or teach more effectively and have more satisfying social and task-based human-human interactions. The value of learning via PAs, however, depends on how well learning gained in a simulated training session transfers to real-world contexts. Use of PAs using augmented reality and mixed-reality technology might help to bridge the virtual-real gap. Schmid Mast et al. (2018) call for more empirical evidence concerning learning transfer and consideration of individual factors such as personality in their discussion of the strengths and limitations of IVAs, and immersive virtual reality technology more broadly. To be able to adapt to learner differences, preferences and needs, PAs will need greater access to data about the learner. Appropriate

[1] Cf. The ACM statement on transparency: http://www.acm.org/binaries/content/assets/public-policy/2017usacmstatementalgorithms.pdf

access to learner data and tailoring of PAs to different individuals and contexts are open research questions.

Hudlicka (2016) concludes that we will see greater proliferation and more formal evaluations (e.g. (Pascual et al., 2008) (Richards & Caldwell, 2017c)), which will lead to better understanding of the appropriate uses of PAs. Addressing criticisms of retrograde behaviourism in the Intelligent Tutoring Systems and Intelligent Learning Environments that PAs inhabit, we will see wider ranges of pedagogy supported by these technologies (du Boulay, 2019). Changes in pedagogy Further advances will be made in developing agents that exhibit empathy and personality, improved user state recognition, affective user modeling and personalization, Scrutable User Models for Learners (Kay and Kumerfield, 2019), natural language processing (understanding and generation), new types of relationships including human-machine relationships.

The viability of self-regulated learning environments inhabited by PAs relies on the development of frameworks, standards and tools to support authoring (like FAtiMA RAGE[2]) and standards (Sottilare & Hart, 2012). These frameworks that allow the agent to reason and appraise emotion (Dias et al., 2014) and tools that allow scenarios and knowledge to be captured directly from the domain expert (Richards & Taylor, 2011) and teachers should bridge the gap between disciplines by making development viable and achievable without the continual need for programmers. Having said that, the introduction of "computational thinking and skills" starting from kindergarten that we are seeing rolled out in Australia and other countries, will also widen the base of those able to create PAs. This is consistent with Sottilare and Hart (2012) prediction that students will need greater appreciation and understanding of different mathematical, artificial computer languages to communicate and live in the future digitized societies.

A growing body of researchers are focusing on responsible design of AI, which incorporates social and ethical values, to prevent undesirable societal outcomes of this technology. Several proposals for principles to describe Responsible AI exist, including the Asilomar[3] principles, the IEEE Ethically Aligned Design recommendations[4], and the ART methodology proposed by (Dignum, 2017). These works however, refer to AI in general and as such do not provide specific guidance to the issue of ethical PAs. Ethical decision-making in AI and robotics is an emerging field, but already several approaches are available. Malle (2016) proposes a framework combining the (until recently) separate fields of robot ethics, in which ethical questions about the design, deployment and treatment of robots by humans are addressed, and machine morality, which is concerned with questions about the moral capacities of a robot, and how these should be computationally implemented. In Cointe, Bonnet, and Boissier (2016) a model is proposed through which an agent can judge the ethical aspects of its own behaviour and that of other agents in a multi-agent system. These works can be applied to the specific situation of PAs. When these capabalities will be realised is uncertain. However, it is inevitable that AI will transform many aspects of life and society in the coming decades. PAs will play a major role in learning in our hospitals, homes, workplaces and leisure spaces, and if their potential is realised and adequately funded, in our schools and universities.

DECLARATION: There are no conflicts of interest. This article did not involve data

2 https://www.gamecomponents.eu/content/582
3 https://futureoflife.org/ai-principles
4 https://ethicsinaction.ieee.org/

collection or human participants, thus ethic approval was not obtained and there is no shared dataset.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989.

Arroyo, I., Murray, T., Woolf, B. P., & Beal, C. R. (2003). *Further results on gender and cognitive differences in help effectiveness.* Paper presented at the International Conference of Artificial Intelligence in Education, Sydney, Australia.).

Aylett, R., Vannini, N., Andre, E., Paiva, A., Enz, S., & Hall, L. (2009). *But that was in another country: agents and intercultural empathy.* Paper presented at the Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. (pp. 329-336): International Foundation for Autonomous Agents and Multiagent Systems.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ, 1986*.

Beaumont, R., Rotolone, C., & Sofronoff, K. (2015). The secret agent society social skills program for children with high-functioning autism spectrum disorders: A comparison of two school variants. *Psychology in the Schools, 52*(4), 390-402.

Berney, A., Carrard, V., Schmid Mast, M., Bonvin, R., Stiefel, F., & Bourquin, C. (2017). Individual training at the undergraduate level to promote competence in breaking bad news in oncology. *Psycho-oncology, 26*(12), 2232-2237.

Bickmore, T. W., Pfeifer, L. M., Byron, D., Forsythe, S., Henault, L. E., Jack, B. W., . . . Paasche-Orlow, M. K. (2010). Usability of conversational agents by patients with inadequate health literacy: Evidence from two clinical trials. *Journal of health communication, 15*(S2), 197-210.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., . . . Zhao, W. (2015). *Automatic detection of learning-centered affective states in the wild.* Paper presented at the Proceedings of the 20th international conference on intelligent user interfaces. (pp. 379-388): ACM.

Bouvier, P., Sehaba, K., & Lavoué, É. (2014). A trace-based approach to identifying users' engagement and qualifying their engaged-behaviours in interactive systems: application to a social game. *User Modeling and User-Adapted Interaction, 24*(5), 413-451.

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer, 50*(5), 116-119.

Buchanan, R. (2011). Paradox, Promise and Public Pedagogy: Implications of the Federal Government's Digital Education Revolution. *Australian Journal of Teacher Education, 36*(2), 67-78.

Butler, D. A. (2008). Air Gondwana: teaching basic negotiation skills using multimedia. *Journal of the Australasian Law Teachers Association, 1*(1&2), 213-226.

Carnell, S., Halan, S., Crary, M., Madhavan, A., & Lok, B. (2015). Adapting Virtual Patient Interviews for Interviewing Skills Training of Novice Healthcare Students. *Intelligent Virtual Agents*, 50-59.

Chattaraman, V., Kwon, W.-S., Gilbert, J. E., & In Shim, S. (2011). Virtual agents in e-commerce: representational characteristics for seniors. *Journal of Research in Interactive Marketing, 5*(4), 276-297.

Cointe, N., Bonnet, G., & Boissier, O. (2016). *Ethical judgment of agents' behaviors in multi-agent systems.* Paper presented at the Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. (pp. 1106-1114): International Foundation for Autonomous Agents and Multiagent Systems.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Danforth, D. R., Procter, M., Chen, R., Johnson, M., & Heller, R. (2009). Development of virtual patient simulations for medical education. *Journal For Virtual Worlds Research, 2*(2).

Davies, C., & Eynon, R. (2013). Believers and Non-Believers: How Potential Users Respond to the Prospect of an Onscreen Learning Assistant.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., . . . Lhommet, M. (2014). *SimSensei Kiosk: A virtual human interviewer for healthcare decision support.* Paper presented at the Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. (pp. 1061-1068): International Foundation for Autonomous Agents and Multiagent Systems.

Dias, J., Mascarenhas, S., & Paiva, A. (2014). Fatima modular: Towards an agent architecture with a generic appraisal framework *Emotion modeling* (pp. 44-56): Springer.

Dignum, V. (2017). Responsible autonomy. *arXiv preprint arXiv:1706.02513*.

Dignum, V. (2019). Responsible Artificial Intelligence. *Springer Series on Artificial Intelligence: Foundations, Theory, and Algorithms, Springer*.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines, 14*(3), 349-379.

Gratch, J., Lucas, G. M., King, A. A., & Morency, L.-P. (2014). *It's only a computer: the impact of human-agent interaction in clinical interviews.* Paper presented at the Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. (pp. 85-92): International Foundation for Autonomous Agents and Multiagent Systems.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. *Handbook of educational psychology, 77*, 15-46.

Grolleman, J., van Dijk, B., Nijholt, A., & van Emst, A. (2006). *Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation.* Paper presented at the International Conference on Persuasive Technology. (pp. 133-141): Springer.

Gwo-Dong, C., Lee, J.-H., Chin-Yeh, W., Po-Yao, C., Liang-Yi, L., & Tzung-Yi, L. (2012). An empathic avatar in a computer-aided learning program to encourage and persuade learners. *Journal of Educational Technology & Society, 15*(2), 62.

Haake, M., & Gulz, A. (2009). A look at the roles of look & roles in embodied pedagogical agents–a user preference perspective. *International Journal of Artificial Intelligence in Education, 19*(1), 39-71.

Halan, S., Sia, I., Crary, M., & Lok, B. (2015). Exploring the Effects of Healthcare Students Creating Virtual Patients for Empathy Training. *Intelligent Virtual Agents*, 239-249.

Harmon, S. (2016). *An expressive dilemma generation model for players and artificial agents.* Paper presented at the Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference.).

Hastie, H., Lim, M. Y., Janarthanam, S., Deshmukh, A., Aylett, R., Foster, M. E., & Hall, L. (2016). *I remember you!: Interaction with memory for an empathic virtual robotic tutor.* Paper presented at the Proceedings of the 2016 international conference on autonomous agents & multiagent systems. (pp. 931-939): International Foundation for Autonomous Agents and Multiagent Systems.

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture, 2*(1), 8.

Hoorn, J., Eliëns, A., Huang, Z., Van Vugt, H. C., Konijn, E. A., & Visser, C. T. (2004). Agents with character: Evaluation of empathic agents in digital dossiers. *Emphatic Agents, AAMAS*.

Hubal, R. C., Kizakevich, P. N., Guinn, C. I., Merino, K. D., & West, S. L. (2000). *The virtual standardized patient.* Paper presented at the Medicine Meets Virtual Reality. (pp. 133-138).

Hudlicka, E. (2016). Virtual affective agents and therapeutic games *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 81-115): Elsevier.

Jacobson, M. J., Taylor, C. E., & Richards, D. (2016). Computational scientific inquiry with virtual worlds and agent-based models: new ways of doing science to learn science. *Interactive Learning Environments, 24*(8), 2080-2108.

Johnson, W. L., Friedland, L., Schrider, P., Valente, A., & Sheridan, S. (2011). *The Virtual Cultural Awareness Trainer (VCAT): Joint Knowledge Online's (JKO's) solution to the individual operational culture and language training gap.* Paper presented at the Proceedings of ITEC.): Clarion Events London, UK.

Johnson, W. L., & Lester, J. C. (2018). Pedagogical Agents: Back to the Future. *AI Magazine, 39*(2).

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in education, 11*(1), 47-78.

Kemmer, M. (2014). The politics of post-apocalypse: Interactivity, narrative framing and ethics in Fallout 3. *Politics in Fantasy Media: Essays on Ideology and Gender in Fiction, Film, Television, and Games*, 97-117.

Kim, Y., & Baylor, A. L. (2016). based design of pedagogical agent roles: A review, progress, and recommendations. *International Journal of Artificial intelligence in education, 26*(1), 160-169.

Kim, Y., Baylor, A. L., & Group, P. (2006). Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and Development, 54*(3), 223-243.

Korotayev, A. (2018). The 21st Century Singularity and its Big History Implications: A re-analysis. *Journal of Big History, 2*(3), 73-119.

Kowatsch, T., Nißen, M., Shih, C.-H. I., Rüegger, D., Volland, D., Filler, A., . . . Büchter, D. (2017). *Text-based Healthcare Chatbots Supporting Patient and Health Professional Teams: Preliminary Results of a Randomized Controlled Trial on Childhood Obesity.* Paper presented at the Persuasive Embodied Agents for Behavior Change (PEACH2017).): ETH Zurich.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611-627.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). *The persona effect: affective impact of animated pedagogical agents.* Paper presented at the Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. (pp. 359-366): ACM.

Lisetti, C. L. (2012). 10 advantages of using avatars in patient-centered computer-based interventions for behavior change. *SIGHIT Record, 2*(1), 28.

Lisetti, C. L., Yasavur, U., de Leon, C., Amini, R., Visser, U., & Rishe, N. (2012). *Building an On-Demand Avatar-Based Health Intervention for Behavior Change.* Paper presented at the FLAIRS Conference.).

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education.

Lyons, J. B. (2013). *Being transparent about transparency: A model for human-robot interaction.* Paper presented at the 2013 AAAI Spring Symposium Series.).

Malle, B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and information technology, 18*(4), 243-256.

Marsella, S., Johnson, W. L., & LaBore, C. (2003). *Interactive pedagogical drama for health interventions.* Paper presented at the 11th International Conference on Artificial Intelligence in Education, Sydney, Australia. (pp. 341-348).

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education, 56*(3), 574-584.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Moreno, R., & Flowerday, T. (2006). Students' choice of animated pedagogical agents in science learning: A test of the similarity-attraction hypothesis on gender and ethnicity. *Contemporary educational psychology, 31*(2), 186-207.

Muralidhar, S., Nguyen, L. S., Frauendorfer, D., Odobez, J.-M., Schmid Mast, M., & Gatica-Perez, D. (2016). *Training on the job: Behavioral analysis of job interviews in hospitality.* Paper presented at the Proceedings of the 18th acm international conference on multimodal interaction. (pp. 84-91): ACM.

Nezami, O. M., & Richards, D. (2017). *Introducing a Multiple Model for Evaluating User Engagement in Educational Virtual Worlds.* Paper presented at the Proceedings of the 9th International Conference on Computer and Automation Engineering. (pp. 16-20): ACM.

Nezami, O. M., Richards, D., & Hamey, L. (2017). Semi-Supervised Detection of Student Engagement.

Nicholas, M., Van Bergen, P., & Richards, D. (2015). Enhancing learning in a virtual world using highly elaborative reminiscing as a reflective tool. *Learning and Instruction, 36*, 66-75.

Nikitina, S., Callaioli, S., & Baez, M. (2018). Smart conversational agents for reminiscence. *arXiv preprint arXiv:1804.06550*.

Norling, E. (2009). On evaluating agents for serious games *Agents for Games and Simulations* (pp. 155-169): Springer.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions*: Cambridge university press.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors, 39*(2), 230-253.

Pareto, L. (2014). A teachable agent game engaging primary school children to learn arithmetic concepts and reasoning. *International Journal of Artificial intelligence in education, 24*(3), 251-283.

Pascual, M., Salvador, C. H., Sagredo, P. G., MÁrquez-Montes, J., Gonzalez, M. A., Fragua, J. A., . . . Muñoz, A. (2008). Impact of patient-general practitioner short messages based interaction on the control of hypertension in a follow-up service for low-to-medium risk hypertensive patients. A randomized controlled trial. *significance, 5*, 6.

Peltu, M., & Wilks, Y. (2008). Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues.

Picard, R. W. (1995). Affective computing.

Picard, R. W. (1997). *Affective computing*. Cambridge, Massachsusetts: MIT Press.

Plant, E. A., Baylor, A. L., Doerr, C. E., & Rosenberg-Kima, R. B. (2009). Changing middle-school students' attitudes and performance regarding engineering with computer-based social models. *Computers & Education, 53*(2), 209-215.

Rao, A. S., & Georgeff, M. P. (1991). *Modeling Rational Agents within a BDI-Architecture*. Paper presented at the Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning. (pp. 473-484): Morgan Kaufmann publishers Inc.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*: Cambridge university press.

Richards, D. (2017). *Intimately intelligent virtual agents: knowing the human beyond sensory input.* Paper presented at the Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents. (pp. 39-40): ACM.

Richards, D., & Caldwell, P. (2017a). *An empathic virtual medical specialist: It's not what you say but how you say it.* Paper presented at the Virtual System & Multimedia (VSMM), 2017 23rd International Conference on. (pp. 1-8): IEEE.

Richards, D., & Caldwell, P. (2017b). Gamification to Improve Adherence to Clinical Treatment Advice. *Health Literacy: Breakthroughs in Research and Practice: Breakthroughs in Research and Practice*, 80.

Richards, D., & Caldwell, P. (2017c). Improving health outcomes sooner rather than later via an interactive website & virtual specialist. *IEEE Journal of Biomedical and Health Informatics, 22*(5), 1699-1706.

Richards, D., & Szilas, N. (2012). *Challenging reality using techniques from interactive drama to support social simulations in virtual worlds.* Paper presented at the Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System. (pp. 12): ACM.

Richards, D., & Taylor, M. (2011). Scenario Authoring by Domain Trainers *Multi-Agent Systems for Education and Interactive Entertainment: Design, Use and Experience* (pp. 206-232): IGI Global.

Rizzo, A., Kenny, P., & Parsons, T. D. (2011). Intelligent virtual patients for training clinical skills. *JVRB-Journal of Virtual Reality and Broadcasting, 8*(3).

Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., & Doerr, C. E. (2008). Interface agents as social models for female students: The effects of agent visual presence and appearance on female students' attitudes and beliefs. *Computers in Human Behavior, 24*(6), 2741-2756.

Roxas, J. C., Richards, D., Bilgin, A., & Hanna, N. (2018). Exploring the influence of a human-like dancing virtual character on the evocation of human emotion. *Behaviour & Information Technology, 37*(1), 1-15.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*: Malaysia; Pearson Education Limited.

Schmid Mast, M., Kleinlogel, E. P., Tur, B., & Bachmann, M. (2018). The future of interpersonal skills development: Immersive virtual reality training with virtual humans. *Human Resource Development Quarterly, 29*(2), 125-141.

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.

Sebastian, J., & Richards, D. (2017). Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. *Computers in Human Behavior, 73*, 479-488.

Sottilare, R., & Hart, J. (2012). Cognitive and affective modeling in intelligent virtual humans for training and tutoring applications. *Advances in Applied Human Modeling and Simulation*, 113.

Spence, S. (1995). Descartes' error: Emotion, reason and the human brain. *BMJ, 310*(6988), 1213.

Swartout, W., Hill, R., Gratch, J., Johnson, W. L., Kyriakakis, C., LaBore, C., . . . Moore, B. (2006). *Toward the holodeck: Integrating graphics, sound, character and story*. Retrieved from

Tanaka, H., Negoro, H., Iwasaka, H., & Nakamura, S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one, 12*(8), e0182151.

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science, 29*(3), 230-241.

Upright, R. L. (2002). To tell a tale: The use of moral dilemmas to increase empathy in the elementary school child. *Early Childhood Education Journal, 30*(1), 15-20.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*: Harvard university press.

Walker, E., & Ogan, A. (2016). We're in this together: Intentional Design of Social Relationships with AIED systems. *International Journal of Artificial intelligence in education, 26*(2), 713-729.

Wang, C., Bickmore, T., Bowen, D. J., Norkunas, T., Campion, M., Cabral, H., . . . Paasche-Orlow, M. (2015). Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. *Genetics in Medicine, 17*(10), 822.