Accountability in human and artificial decision-making as the basis for diversity and educational inclusion

Kaśka Porayska-Pomsta[1] and Gnanathusharan Rajendran[2]

[1]University College London, UCL Institute of Education, UCL Knowledge Lab, 23-29 Emerald Street, London WC1 3QS, UK. K.Porayska-Pomsta@ucl.ac.uk
[2]Heriot-Watt University, Department of Psychology, Edinburgh Centre for Robotics, Edinburgh, EH14 4AS, Scotland, UK. T.Rajendran@hw.ac.uk

Abstract. Accountability is an important dimension of decision-making in Human and Artificial Intelligence (AI). We argue that it is of fundamental importance to inclusion, diversity and fairness of both the AI-based and human-controlled interactions and any human-facing interventions aiming to change human development, behaviour and learning. Less well debated, however, is the nature and the role of biases that emerge from theoretical or empirical models that underpin AI algorithms and the interventions driven by such algorithms. While, the biases emerging from the theoretical and empirical models also affect human-controlled educational systems and interventions (e.g. hindsight and unconscious biases), the key mitigating difference between AI and human decision-making is that human decisions involve individual flexibility, context-relevant judgements, empathy, as well as complex moral judgements, missing from AI. In this chapter, we argue that our fascination with AI, which pre-dates the current craze by centuries, resides in its ability to act as a 'mirror' reflecting our current understandings of human intelligence. Such understandings also inevitably encapsulate the biases emerging from our intellectual and empirical limitations. We make a case for the need for diversity to mitigate against biases becoming inbuilt into systems (in both Education and AI) and, with reference to specific examples of AI approaches and applications, we outline one compelling future for inclusive and accountable AI and Educational research and practice.

> *"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less" Maria Skłodowska-Curie*

1. Introduction

Artificial Intelligence (AI) presently receives a lot of press, both for its potential to tackle challenges, from policing, to healthcare, to education, and for the perceived threat that it poses to our (Human) identity, autonomy and future functioning. There is a tension in the current perception of AI between its utopian and dystopian overtones, which Stiglitz has synthesised recently as one between AI as an *Human replacing machine (AI)* vs. as an *Human assisting machine (IA)*[1]. He placed this distinction at the heart of the questions about the implications of AI for society, and for the future of human self-determination, wellbeing and welfare. As Reisman et al. (2018) point out, the recent transition of AI from a purely scientific domain to real-world applications has placed AI at the centre of our decision-making without our having had a chance to develop a good understanding of the nature of those implications, or to define

[1] https://www.youtube.com/watch?v=aemkMMrZWgM

appropriate accountability measures to monitor and safeguard against any harms. Crawford[2] refers to this situation as an inflection point at which we are starting to comprehend how AI can reinforce a whole plethora of socio-cultural biases that are inherent in our existing social systems, and where there is a pressing need for us to question and to hold to account the AI solutions and the decisions that are based on them. Thus, in the present context where AI technologies and their impact are still largely unknown, questions (and actions) related to social and educational inclusion, and education more broadly, are critical to how we develop and utilise AI in the future, and how we develop a system of accountability that is able to guide us in doing so in socially responsible, and empowering ways.

There is a growing awareness that current AI systems tend to expose and amplify social inequalities and injustice, rather than address them (Crawford & Calo, 2016; Curry & Reiser, 2018). There are two known reasons for this. First, the socio-cultural biases that are inherent in the data consumed by the AI models, make those models also socially skewed. Such biases may originate from (i) our historic and current socio-cultural prejudices (be-it related to race, gender, ethnicity, etc., e.g. police records that are skewed towards particular social groups such as young black males as more likely to commit crimes), (ii) lack of data that is representative of the society as a whole (Crawford & Calo, 2016), or (iii) they may be an artefact of the specific classification algorithms used and the ways their success is being measured (e.g. Lipton & Steinhardt, 2018). Although many AI 'solutions' are well intentioned, given the current state of both the AI technologies and our own limited understanding of the ways in which they impact human decision-making, their deployment in real high-stakes contexts, such as arrest decisions, seems premature (Reisman et al., 2018).

The fact that many AI solutions – especially machine learning – are seldom open to being inspected, or contested by humans represents the second reason why AI is thought to reinforce social biases. Specifically, the so-called black-box AI often prevents humans (engineers and users) from even knowing that biases are present, or from fully understanding how they arise (e.g. from data or from classification algorithms, or both - see e.g. Brinklorf & Hammer, 2018). This is known interchangeably as the explainability or interpretability problem. Addressing this problem is increasingly seen by the AI community as a remedy to data and AI interpretation bias, speaking to questions of accountability and trustworthiness of the AI-driven decisions (Lipton & Steinhardt, 2018, Conati et al., 2018). The AI community is also beginning to recognise that to be genuinely beneficial, AI-driven decisions must be contestable and open to being changed by the users (Brinklorf and Hammer, 2018; Bull & Kay, 2016).

In this chapter, our treatise is that in order to conjure a positive future of educational inclusion involving AI requires us first to appreciate that our own human systems (educational, clinical, social justice, etc.) and models of inclusion do not represent absolute truths. Instead, those systems are inherently biased representations of the world, which are limited by our current knowledge and social structures, which determine who and how may influence those representations. Second, in order to genuinely understand the potential of AI in the context of human learning, development and functioning, and to safeguard against misuse, there is a need for an informed differentiation between human and artificial intelligence. Such differentiation is needed to make us stop "worshiping at the altar of technology"[3] and to admit diverse stakeholders, who are not AI experts, into partaking actively in the design of AI technologies and their use for education.

[2] https://royalsociety.org/science-events-and-lectures/2018/07/you-and-ai-equality/

[3] https://royalsociety.org/science-events-and-lectures/2018/07/you-and-ai-equality/

The rest of the chapter is structured as follows. In sections 2 and 3 respectively, we briefly introduce the concepts of accountability and inclusion. We outline the definitional challenges, highlighting how the two concepts relate to one another, to scientific and technological innovation, and to dominant approaches to inclusion. In section 4 we define AI in order to provide an informed basis for considering its true potential in the context of educational inclusion. In particular, we introduce AI not solely as a solution to some "curable" problem, but as a conceptual framework for formulating pertinent questions about learning, development and inclusion, and as a method for addressing those questions. In this section we also outline the key differences between Human Intelligence (HI) and Artificial Intelligence (AI). As will be argued in section 5, acknowledging and understanding this difference explicitly allows us to appreciate how AI can be designed and used to assist learners and educators through (i) providing relative safety zones for learners to accommodate and even reduce any pronounced differences or difficulties, e.g. social communication anxiety in autism (AI as *a stepping stone*), (ii) acting as a mirror in self-exploration and development of self-regulation competencies (AI as *a mirror*) and (iii) offering a medium for understanding and sharing of individual perspectives and subjective experiences, as the basis for nurturing tolerance, compassion and for developing appropriately tailored educational support (AI as *a medium*). We employ examples from our own research, which are of relevance to social and educational inclusion in which we used AI: one involving a genetically determined case of autism and the other – a socio-economic one of youth unemployment. Section 6 will conclude the chapter by summarising the interdependency of the key concepts considered (inclusion, accountability, and AI), and will outline the steps that are needed to achieve our vision of AI as a technology for social good.

2. Accountability

Accountability is a key dimension of decision-making in Human and Artificial Intelligence, and it is crucial to any democratic, tolerant and inclusive society. This is because accountability is fundamentally about giving people the autonomy of action through knowledge. However, although accountability has become *de facto* a cultural term, it is not always clear what it actually means in practice.

To date, two main ontological perspectives on accountability have been adopted in law and policy (Dubnick, 2014). The first perspective relates to the *post-factum* accountability, involving a blame-*able* agent whose attempts to manipulate another agent's actions according to their wishes require them to be held responsible for those actions and for the consequences thereof, e.g., the blameable agent *after* the 2008 financial crisis was the financial sector. The second dominant perspective is the *normative* one, representing some preferential solutions to a range of aspirational problems such as justice, democracy, racial discrimination etc., where societal, political, or institutional organisations are the decision-makers. This is referred to as the *pre-factum* type of accountability, involving an *a priori* blame-*worthy* agent or agents. Here, it is assumed that the aim of accountability measures is to reach a societal change or mass acquiescence in anticipation of some possibly blameworthy actions or events. For example, after the 2008 financial crisis, a set of accountability measures were imposed by the UK's Financial Conduct Authority (FCA) on the financial sector to prevent similar crises in the future, and to offer transparency in the sector's decision-making and actions.

Recently these two dominant stances have been critiqued as being too rigid to allow for an operationalisation of accountability as an ongoing social process that it is (Dubnick, 2014). The main issue here is that although the pre and post-factum definitions provide a moral and legal framing of accountability, they do not specify how accountability can be actioned in an agile way in diverse and often changeable contexts, involving different stakeholders, and given our perpetually changing understanding of the world. For example, by exposing the existing biases in our pre-AI representations of the world (e.g. in policing), AI has also demonstrated a substantial gap between the aspirational social rhetoric of inclusion, tolerance and welfare and the reality on the ground. Specifically, it showed not only that our systems are still based on historical and socially skewed data, but also that our predominant accountability measures and laws struggle to catch up with our social aspirations and changing norms, and with our developing scientific and practical knowledge related to inclusion.

A more flexible approach is offered by an ethics-based theory of accountability (Dubnick, 2014), where accountability is defined as *a social setting* and *a social negotiation*. Here, the rules and the moral codes which define how it is operationalised in practice can be adjusted according to the changes and needs occurring within the individual stakeholder groups in tandem with and in response to the developments in our scientific, economic and social circumstances and understandings. In this view, accountability is of *relational nature*, involving 'multiple, diverse and often conflicting expectations (MDCE)', priorities, and investments of different stakeholders, along with temporal fashions that determine *who* is accountable for decisions and actions to *whom*, with respect to *what,* and *when* (Dubnick, 2014, p.4). In this account, *the who*, *the whom*, *the for-what* and *the when* represent context dependent variables that are instantiated based on the salience assigned to the specific MDCEs, with accountability becoming an exchange and an ethically regulated, tractable and auditable compromise between different competing interests and gains of the decision makers.

This relational approach is of particular relevance both in the context of AI and educational inclusion. In particular, this interpretation acknowledges that there is no one-size-fits-all, best way to make the decisions of others auditable and that, fundamentally, the judgements related to the *blameability* or *blameworthiness* of decisions are based on the relative needs and goals of the stakeholders affected. This means that if the system is designed in such a way that it hinders or by definition excludes some groups of stakeholders from being able to inspect and influence it, for example by obstructing their participation in making decisions in matters that affect them, or by preventing them from acquiring appropriate skills to engage in such auditing, then social inclusion, equity and fairness are compromised.

In contrast, the relational definition of accountability: (i) allows us to appreciate it as a social construct that assumes different and often conflicting interests and prioritisations thereof that affect people's decisions; (ii) presupposes the existence of stakeholders who are empowered intellectually, financially, etc., to generate and respond to the different expectations, and invest in enhancing their salience, therefore also highlighting accountability practices themselves as being neither perfect or neutral; (iii) it can be used directly to examine the role of AI in first exposing this lack of neutrality (as already discussed in the introduction), and second, to highlight the continuing need to empower different potential stakeholders to invest in generating and lobbying for their priorities. Thus, the uniqueness of AI in this context lies not only in its ability to act as a moral mirror and a magnifying glass for examining our pre-existing conceptions of social inclusion and social justice, tolerance and welfare. It also lies in its ability to provide concrete, tractable, interactive, and scalable means for genuinely democratising accountability mechanisms, including those related to the explainability and contestability of educational interventions and assessments. As will be elaborated and exemplified further in section 5, this latter affordance of AI represents one of the most exciting avenues for AI in educational inclusion.

3. Inclusion

So far, we discussed the potential of the relational framing of accountability in the context of inclusion in allowing us to devise accountability policies in ways that respond to our developing knowledge and changing social norms. We also highlighted the link between accountability and AI and the latter's ability to expose pre-existing biases. Although the relational view of accountability may describe how the accountability processes play out, its present operationalisations rely predominantly on the pre- and post-factum framings. This is problematic from the point of view of inclusion in two ways. First, the two non-relational stances on accountability de-emphasise the need for empowering all potential stakeholders to influence decisions that affect them, instead surrendering the responsibility for enforcing accountability to those who are endowed with appropriate governing powers, but may lack the experiential, contextual and intellectual basis for their decisions. Second, they are prescriptive top-down approaches which reinforce existing definitions of inclusion, rather than assuming a priori that those definitions are likely to evolve with the changing scientific knowledge, social norms and aspirations.

Historically inclusion tends to be defined in terms of specific pronounced differences from what may be currently considered the 'norm', i.e. the definitions of inclusion tend to be exclusive by default. For example, the OED defines inclusivity as:

> *"The practice or policy of including people who might otherwise be excluded or marginalized, such as those who have physical or mental disabilities and members of minority groups. 'you will need a thorough understanding of inclusivity and the needs of special education pupils'"*[4]

This definition explicitly uses special education highlighting as its illustration those who otherwise would be marginalised. This is also indicative of the definitional problems with inclusion that arise at the systemic level, where inclusion is treated as a *solution* to integrating and assimilating those who are considered at the margins, rather than as a *process* through which differences can be used to extend our understanding of 'normality' or 'typicality', and where societies can expect to be influenced by and to benefit from diversity. The definitional limitations of this framing of inclusion also both reflect and are reflected in many educational and clinical intervention approaches. In particular, the history of psychiatry and psychology is one of exclusion and marginalisation, and it is based explicitly on notions of *abnormality*. Although laudable in its aims to understand conditions through aetiology by taxonomy, the resultant diagnoses, classifications, and practices have been historically decided by the majority and imposed on the minority.

To illustrate this, the Diagnostic and Statistical Manual for the American Psychiatric Association was first published in 1952 (DSM-I). This manual has framed our clinical and scientific understanding of abnormality, and it has remained disorder-focused. For example, in the case of autism, the very name in DSM-V (2013) Autism Spectrum Disorder (ASD) reflects this perspective. In 1952 DSM-I listed 106 disorders. In 2013, DSM-V listed 300 disorders (Baron-Cohen, 2017). Homosexuality was classified as a disorder in DSM-I and DSM-II and

[4] https://en.oxforddictionaries.com/definition/inclusivity

was only removed from DSM-III in 1980. With respect to autism, O'Neill (2008) argues that in much the same way as homosexuality was no longer considered a disorder, the classification of autism should also be reconsidered.

One criticism of framing developmental differences as disorders rather than as conditions is that it misses the point of functionality that reflects an evolutionary purpose. For example, conditions like Tourette syndrome, ASD and ADHD include behavioural phenotypes of executive control, such as behavioural inhibition or inability to start or stop oneself from engaging in certain behaviours. An evolutionary perspective asks about the function of poor inhibitory control and about the purpose of keeping certain traits and behaviours in the gene pool, i.e. leading to *neurodiversity* in the population. This is in contrast to viewing abnormality as an aberration either from a social ideal of 'normal', or from a statistically derived norm, like intellectual disability[5]. This is supported by animal behaviour research, such as that by Dobson and Brent (2013), who postulate a mechanism by which neurodiversity might be functional and beneficial, i.e. that variations in the genome can help animals be adaptive and that such, differences are part of natural selection and fitness, rather than abnormalities to be eradicated. Thus, the neurodiversity perspective takes a different position from the pathological one. For example, in the context of autism, increasingly researchers have been investigating how social and educational environments can be co-created with stakeholders to represent and empower, rather than segregate neuro-diverse learners both in traditional practices (e.g. Baron-Cohen, 2017; Rajendran, 2013; Remington, 2018), and those involving the application of AI (Porayska-Pomsta et al., in press).

The take-home message here is that, abnormality is *socially* rather than biologically constructed, and thus, any developments related to educational and social inclusion, including AI use in education, must take this into account explicitly. Importantly, a closer look at the history of social and educational practices in this context, reveals that the questions of bias and discrimination pre-date the emergence of AI, by a long way. This highlights that the questions of accountability, inclusion and the role of AI in shaping our collective understanding of ourselves are intricately intertwined and that for AI to serve educational inclusion and best educational practices, they need to be considered together.

4. Artificial Intelligence

In order to appreciate how AI technologies may interplay with the constructs of accountability and inclusion and to help us understand how AI can be used to deliver more inclusive education, it is important to consider the original conception of AI as:

1. An *applied philosophy* allowing us to formulate key questions about different aspects of (human) intelligence (Davis et al., 1993; Davis, 1996; Russell ad Norvig, 2003; Woolf, 2008);
2. A *method* for testing our different theories about intelligence by operationalising them in computational models which produce observable and measurable behaviours without our having to take real action (e.g. Davis et al., 1993; Porayska-Pomsta, 2016);
3. A *solution* to specific real-world challenges (like policing or medical diagnosis), but which are nevertheless artefacts of our questioning and experimentation, based on the

[5] An IQ of less than 70 is considered intellectually disabled because it falls 2 standard deviation from the population mean of 100. The assumption is that IQ is normally distributed and abnormality can be statistically determined in an objective was

current, and hence by definition incomplete, state of our knowledge and understanding.

With AI having now crossed over from a purely scientific domain to practical mainstream applications, AI as *a solution* has taken the centre stage. However, we believe that this single-lens limits our view on the actual strengths and weaknesses of AI in the context of socially embedded practices such as in education, and more broadly as a tool for scientific enquiry into what makes us human. It obscures the need for our asking *what society do we want*, instead permitting technological advances (and the few tech specialists behind those) to dictate what society we end up with.

In contrast, the broader three-lens view of AI makes us appreciate that both the questions and the answers formulated with the help of AI are relative to the current state of our knowledge. Importantly, this definition helps us further in approaching inclusion and education not merely as some fixed state for which there is a set of equally fixed solutions that can be administered like medicine, but as a social process and a state of mind, which requires our own investment, enquiry and willingness to change. Seen in this way, the necessary pre-requisites of a socially inclusive AI, that caters for and involves the human in its decision-making, become readily apparent. As will be illustrated in section 5, AI can uniquely provide both the intellectual and physical means that are manipulable and scalable, allowing for an exploration, speculation and rigorous experimentation (e.g. through simulated scenarios) about what it means to be inclusive along with the mechanisms that may be conducive and effective to fostering inclusion through education and educational practices. To appreciate this point, it is necessary also to understand the ways in which AI differs from human intelligence by considering how it operates at a lower-level of description.

Two broad schools of thought define how AI has been implemented to date: (i) the so-called ‘good old fashioned’ AI (or GOF AI) and (ii) machine learning. The GOF AI requires explicit representation of knowledge, which reflects an ontological conceptualisation of the world and actions that are possible therein, along with some well-defined measures of success in terms of concrete goals and goal satisfaction constraints. For example, in the context of maths tutoring, the ontological representations will relate to the specific sub-domains of maths, say – misconceptions in column subtraction, and rules which define the possible operations on this subdomain. The goal satisfaction in this case may be in terms of student’s correct or incorrect answers. As such, GOF AI is by definition limited, with the concepts and rules being hard-coded into the systems, often during laborious and time-consuming design stages (Porayska-Pomsta and Bernardini, 2013). Such rules are typically elicited through questioning of human experts in a given domain, by observing their expertise in real contexts or by hand annotating data (video recordings, interaction logs, etc.) of humans engaging in specific tasks. From the point of view of human learning and accountability of AI decision-making, the key advantage of such knowledge-based systems is that they require a detailed understanding of the domain, in order for knowledge ontologies to be constructed, thus also potentially leading to a greater understanding of the domains represented, and the fact that the resulting ontologies are transparent, inspectable, and often understandable by humans (Davis, 1993; Russell & Norvig, 1995).

By contrast, machine learning (ML) learns solutions from first principles by applying statistical classification methods to large data sets. ML is largely inspired by our current knowledge of how the brain works and by cognitive psychology theories, such as reinforcement learning (Russell & Norvig, 1996; Sutton & Barto, 2000). ML carries a substantial promise both in terms of reducing the effort required to specify knowledge ontologies and in being able to go beyond the knowledge we have ourselves, and in so doing – in driving more accurate decision-making than our own capabilities allow for. Thus, one of the most exciting aspects of ML is that it can discover new associations in the world and predict future outcomes based on prior data in complex domains which may be hard for the human to grasp and analyse efficiently.

One of the recent prominent examples of this ability of ML is the success of the AlphaGo programme by Google Deep Mind (henceforth AlphaGo-DM)[6] (Silver et al., 2017). The game of Go represents a highly complex, albeit constrained, problem space where the solutions require more than simply knowing the game's rules. It is an ancient game which takes a lifetime to master and is considered one of the most challenging games ever invented. In 2017 AlphaGo-DM beat the human Go world champion, by presenting strategies that were not known to him. Interestingly, in this context, despite his defeat, the master, expressed his excitement at the realisation that he could learn exciting new game strategies from a machine. In this, he made an explicit link between AI and its potential for human learning and creativity.

However, it is important to appreciate that ML's ability to come up with novel solutions is not a sign of its humanity or creativity, but rather of a different and in many ways a far more advanced, computational prowess and efficiency than afforded by the human brain in similar tasks. In this sense, the ML employed in AlphaGo-DM demonstrated its ability to engage in *intrepolation*, i.e. averaging information based on voluminous data, and *extrapolation*, i.e. finding new information (e.g. Sutton, and Barto, 2000). However, what ML and AI more broadly cannot do, and what differentiates it further from human learning and intelligence, is to invent new things (e.g. to invent a new game), to imagine things, to entertain fantastical scenarios, to employ counterfactual or critical thinking beyond the gain/loss measures, and crucially –to entertain moral judgement. More generally, the fundamental difference between AI and HI is that although AI aims to *emulate* our own behaviours, on the whole and for pragmatic reasons of tractability, it does not require fidelity to human cognition and functioning (Russell & Norvig, 2003). This difference is central to the present debates about AI safety, ethics and its implications for society and it explains why AI's ability to surpass (or more accurately – to bypass) some of our own talents, may lead to our sense of disempowerment and impending doom for our welfare and wellbeing[7,8], and even our status as a species[9]. However, what is far less audible in the current debates, is the fact that these same characteristics that frighten us, make AI precisely the tool that might be needed to enhance our abilities, to make us reflect on who we are and who we want to be, and to use it as an educational instrument of social change. In the next section, we use concrete examples from our own research to elaborate on how AI can act in this positive way. The key question to bear in mind here is the extent to which we want to surrender our autonomy and learning to the AI vs. to use AI to enhance our learning and decision-making capabilities (see also Stiglitz's AI vs IA introduced in section 1).

5. AI and Educational Inclusion: Beyond the Bias

The future of AI and educational inclusion is not necessarily a dystopian one. As discussed throughout this chapter, the current issues of AI bias actually provide a sharp pair of glasses onto how we create systems, and on the extent and nature of our own inherent biases. Our aim here is not to rail against systems, which often allow for patterns to be seen. We argue that as

[6] https://deepmind.com/research/alphago/

[7] https://www.theguardian.com/commentisfree/2018/feb/01/robots-take-our-jobs-amazon-go-seattle

[8] https://www.independent.co.uk/life-style/gadgets-and-tech/news/stephen-hawking-artificial-intelligence-fears-ai-will-replace-humans-virus-life-a8034341.html

[9] http://unesdoc.unesco.org/images/0026/002615/261563E.pdf

a precise philosophical and methodological tool, AI can help us first, to understand, regulate and accept ourselves, and second, to understand, and be able to access other people's experiences and points of view. According to a large body of cross-disciplinary research (e.g., Flavell, 1979; Paul, 1990; Moshman, 2011; Terricone, 2011; Prizant et al., 2003; 2006; Lai, 2011), such understanding and access represent two foundational pre-requisites to inclusion regardless of whether AI is present. In this section, using examples from our own research, we demonstrate how AI, with its ability to shine a bright light onto our own behaviours and conceptions of the world, can help us gain a better understanding of ourselves and of others, and pave the way for a more inclusive education and society. We have identified three affordances of AI in this context, which we see as key research investment areas of the future.

5.1 AI as a stepping stone

AI-driven environments are very good at providing situated, repeatable experiences to their users, offering an element of predictability and a sense of safety, while creating an impression of credible social interactions, e.g. through adaptive feedback. This is important, in contexts where the users may experience social anxiety, or when they lack self-efficacy and self-confidence. For example, the ECHOES project (Porayska-Pomsta et al., *in press*; Bernardini et al., 2014) created an AI environment for young children with autism spectrum disorders (ASD), through which they learned, practiced and explored social interaction skills.

Autism is a neurodevelopmental condition which involves difficulties in social communication and interaction, and restricted and repetitive behaviours and often includes feelings of social anxiety. The aim of autism interventions is to reduce those difficulties. One issue increasingly highlighted by interdisciplinary research (e.g. Prizant et al., 2003; 2006), is that many interventions focus on correcting the deficits in a bid to adapt the children to the environment, rather than on correcting the environments to alleviate children's difficulties. By focusing on correction rather than accommodation of differences, such interventions often fail to access children's needs, and their interpretation of the world, leading to missed opportunities for understanding and learning about each other's perspectives by both the learners and practitioners (Rajendran, 2013).

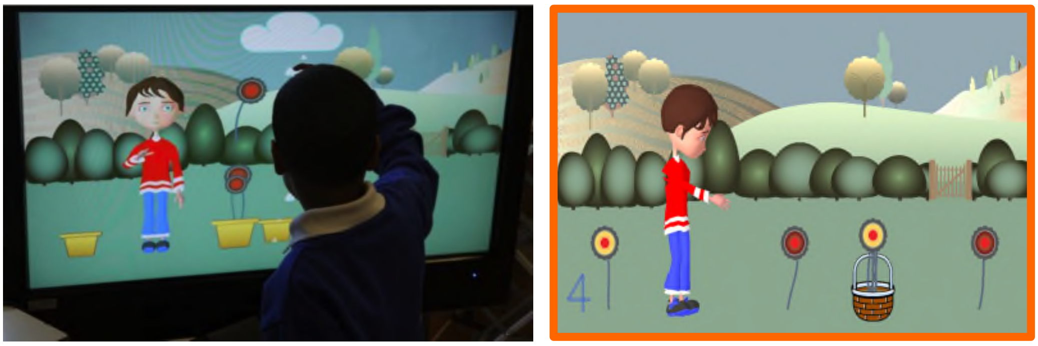
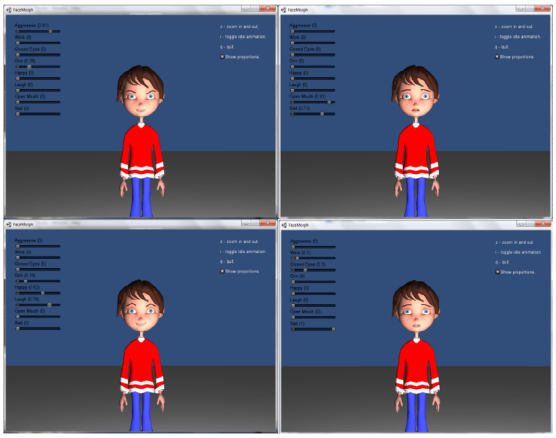


Figure 1: A child playing with the ECHOES agent through the multi-touch screen interface (Left). The agent points to a flower that it wants a child to pick and put in the basket in a bid for attention and interaction with the child (Right).

ECHOES was developed for use in schools. It utilised an AI agent as a social partner in a variety of semi-fantastical scenarios involving both exploratory, open-ended activities and well-defined closed tasks, e.g. picking flowers, or throwing a bouncy ball through a virtual cloud to change the ball's colour. Most activities were a collaboration between children and the agent, and could include a human social partner (a teacher or researcher accompanying the child), if the child wanted to involve them. As our target users were children at the lower-end of the autism spectrum who were classified as non-verbal, ECHOES employed a large multitouch screen through which they acted on the environment (see Fig. 1). The agent acted in a positive and structured way through initiating interactions with the children, and enthusiastically responding to any bids for interaction from them. Since, initiating and responding to bids for interaction is an area of particular difficulty in autism, these skills were the focus in ECHOES.

The agent's actions were aided by a GOF AI planning architecture, which determined the agent's: (i) choice of actions in real-time given its appraisal of children's behaviours, and (ii) longer-term action plans related to helping children become more used to initiating and responding to bids for interaction. The planner also catered for the emotional predispositions of the agent, e.g. its propensity for happiness and positivity (Dais & Paiva., 2005). The agent was endowed with an ability to display a wide range of complex emotions for the child to explore (see Fig.2). However, given that the agent was quite obviously not a real child (it was a cartoon character able to support social interaction contingently) coupled with children having control over the type, number and sequence of activities, provided a needed safety zone for them to engage in social interactions without having to endure the typical drill-and-practice training. It allowed them to explore the causes and effects of their actions repeatedly and without the anxiety of real-world consequences, giving them the time to get used to particular forms of interactions, to rehearse them with the agent and to decide if and when they were ready to interact with a human. This level of control and quality of interaction practice are rarely possible in classrooms or during contrived clinical environments, where many children often feel inhibited to engage in communication at all. Importantly, in adopting this approach, in line with best autism practices and contrary to the corrective approaches to inclusion, ECHOES centred around the child's needs, allowing them to reveal their abilities and strengths at their own pace and gradually.



Aggressive (0.81)
Wink (0)
Closed Eyes (0)
Grin (0.28)
Happy (0)
Laugh (0)
Open Mouth (0)
Sad (0)

Aggressive (0)
Wink (0)
Closed Eyes (0)
Grin (0)
Happy (0)
Laugh (0)
Open Mouth (0.91)
Sad (0.73)

Aggressive (0)
Wink (0)
Closed Eyes (0)
Grin (0.19)
Happy (0.62)
Laugh (0.79)
Open Mouth (0)
Sad (0)

Aggressive (0)
Wink (0.1)
Closed Eyes (0.3)
Grin (0)
Happy (0)
Laugh (0)
Open Mouth (0)
Sad (1)

Figure 2: A tool design tool demonstrating the complex emotional displays of the AI agent in the ECHOES project. The sliders to the left represent individual emotions such as anger, happiness, fright. These can be blended to display ambiguous or nuanced emotions depending on the instructions from the planner as to what emotions the agent is 'experiencing' given its interpretation of the child's actions and its own goals.

A rigorous evaluation of ECHOES revealed that the frequency of children's responses and initiations have increased over time, with a significant increase in responses to human partners during ECHOES' use. Additionally, teachers' reports suggest transfer of some critical social behaviours from ECHOES to classroom contexts, such as children's initiating and responding to greetings, transitioning between activities, and even initiating and responding verbally, which in many cases was revelatory to teachers who thought those children to be non-verbal (see Porayska-Pomsta et al., in press).

Unlike many AI environments and contrary to many teachers' fears of AI being set to replace them, in ECHOES we recognised the strength of AI residing in its imperfect, bur nonetheless a credible approximation of human social abilities. These imperfections were explicit and critical to boosting children's confidence and their own sense of social competence. The role of the human partner (a teacher) was then to build on the strengths demonstrated by the children and to reinforce the sense of confidence acquired with the AI agent in typical classroom and playground contexts. Here, the fact that AI was not the same as a human, but that it was able to approximate plausibly some human behaviours in a just-in-time socially congruent manner was key, because it allowed children to get used to the different social scenarios, with the agent providing consistent, but not fully predictable (owing to its autonomous decision-making facilitated by the AI planning architecture) interaction partnership. The recognition by children of the difference of the AI agent from a human is critical for their engagement, for lessening of their social anxiety and for increasing their sense of autonomy and control over the interaction, all of which are rarely afforded to them in real social situations. AI allows to regulate carefully this sense of autonomy and self-efficacy in preparation for the real-world situations.

5.2 AI as a mirror

AI operates on precise data and this means that it is also able to offer us precision of judgement and recall of events. With respect to inclusion, provided that there is a possibility of come-back from the human, this can be very valuable, even if in all its precision, AI does not necessarily offer us the truth. Systems that employ the so-called open learner models (OLMs) show how users' self-awareness, self-regulation and ultimately self-efficacy can be supported by allowing them to access, interrogate, and even change (through negotiation with the AI system) the data generated of them (Bull and Kay, 2016; Conati et al., 2018). For example, the TARDIS project (Porayska-Pomsta et al., 2014; Porayska-Pomsta et al., 2015; Porayska-Pomsta & Chryssafidou, 2018) successfully used the OLM approach to provide young people at risk of exclusion from education, employment or training (NEETs) with insight into their social interaction skills in job interview settings and with strategies for improving those skills. Here, data about the young people's observable behaviours is first gathered and interpreted during interactions with AI agents acting as job recruiters. This data, which relates to the quality of users' specific verbal and non-verbal behaviours (e.g. length of answer to specific interview questions, facial expressions, quality of gestures, posture, and voice respectively) is then used as the basis for detailed inspection by the learner, aided by a human coach. Such inspection is intended to provide the platform for the learners to explore their specific strengths and weaknesses in their job interview performances, and for developing a set of strategies for self-monitoring and self-regulation during further interviews.

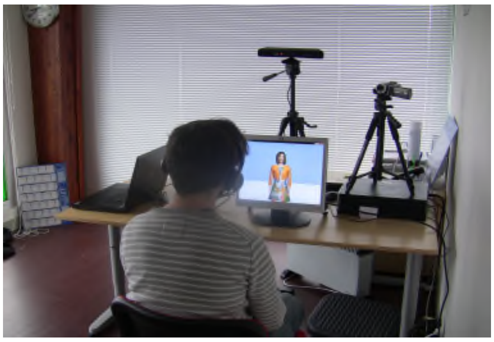


Figure 3: Interaction with TARDIS was facilitated through an off-the-shelf Micrsoft Kinect, which was used to detect users' gestures and posture as well as facial expressions, and high quality microphone to detect voice.

In TARDIS the learning interaction was facilitated through an off-the-shelve Microsoft Kinect and a high-quality microphone (Fig. 3). These collect data such as specific gestures performed by the user, voice quality and speech duration. These data provide the necessary input to the system, which allows it to create a user profile (a model) and to assess the users' performance in terms of the quality of their verbal and non-verbal behaviours. The assessments, are stored in *learner models* and are used by other modules responsible for managing the interaction scenarios, to select appropriate questions during interviews and to drive the behaviours of the AI agents acting as job recruiters (Jones et al., 2014). As in ECHOES, the agents were furnished with a wide range of behaviours, underpinned with an emotion-driven planning architecture.

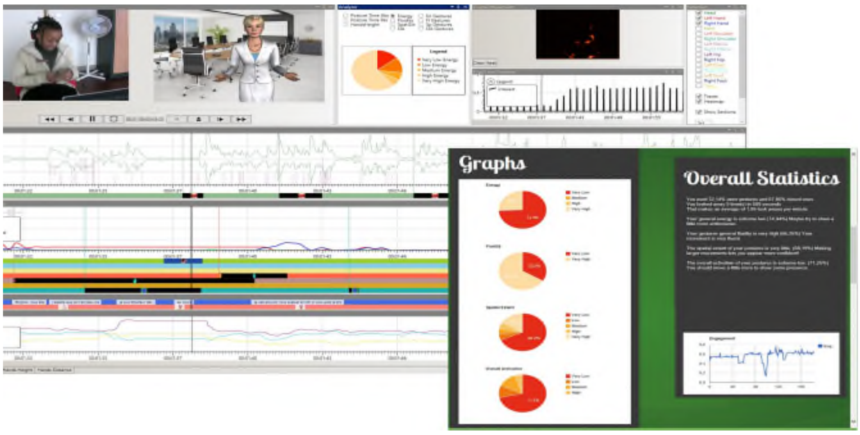


Figure 4. TARDIS scrutable OLM showing synchronised recordings of the learners interacting with the AI agents along with the interpretation of the learner's low-level social signals such as gaze patterns, gestures, voice activation in terms of higher-level judgements about the quality of those behaviours, e.g. energy in voice.

Of particular importance to the relationship between AI and data bias, accountability and social inclusion considered in this chapter, is the fact that the TARDIS learner models are opened for user inspection after the interviews with the AI agents. These models display data gathered about the users' behaviours during interview simulations, along with the system's interpretation of this data (see Fig 4). Through the TARDIS open learner model (OLM), the users have access to interactive timelines of their interview simulations, including precise information on all the actions that they and the agents performed moment-by-moment. The replay of these actions is synchronised with video recordings of the learners and the agents during interview simulations (top left of Fig.4). The learners can also inspect the AI's interpretation of the quality of their individual behaviours (top and bottom right-hand-side of Fig.4), e.g. the energy in their voice, expansiveness of their gestures, etc., together with a commentary on whether these are appropriate at any given point during the interview and what might need to be corrected in the future.

A controlled study compared TARDIS and a traditional online self-improvement programme. It revealed significant improvements for the TARDIS users in terms the quality of their interview answers, verbal and nonverbal behaviours, and self-reported measures related to their levels of anxiety, self-efficacy and quality of their answers. As well as providing a situated experience of job interviews to the young people many of whom have never experienced a job interview before, through its OLM, TARDIS offered the learners an invaluable insight into their own behaviours, triggering self-awareness, self-reflection, explanation, planning and self-monitoring in future interactions, including during human-to-human job interviews. Here, the goal was very explicitly to provide the learners with an objective mirror that they could look into, through which they could question themselves either privately, with peers, or with practitioners, and which they could use as the basis for developing informed self-understanding.

5.3 AI as a medium

Just as AI systems, such as those based on OLMs, can support the development of self-understanding and self-regulation, they can also provide unique and unprecedented insight to educational practitioners about their pupils. Gaining such insight can be game-changing in inclusion practices and individual support interventions, because it can reveal learners' behaviours and abilities that might be hard to observe or foster in traditional environments. For example, in ECHOES some children who were thought to be uncommunicative, became motivated to communicate, revealing their previously hidden potential and changing the way in which teachers supported them beyond ECHOES.

The potential of AI as a medium is not merely in the data and its classification, but also in the way that it provokes human reflection, interaction and adaptation of the existing points of view and practices, i.e. it aids self-accountability, which is of crucial importance to learning. This affordance was particularly manifest in the context of TARDIS, where the OLMs facilitated close inspection and reflection not only by the learners, but also by the practitioners. This allowed for access to the learners' experiences with AI interpretations of the job interview performances giving an objective prop for the practitioners to pump the learners for explanations, for identifying the strengths and weaknesses in their performances and for devising plans for how to build on the former while addressing the latter. One striking observation from the TARDIS studies was the change in the quality of feedback and conversations using TARDIS OLM versus relying on the learners' and practitioners' imperfect recall of the situations. As such the tool allowed to alleviate learners' sense of being judged, putting them in control over the interpretations of their own experiences and over the directions they wanted to take their debriefing conversations with the practitioners. TARDIS also provided a platform for discussions amongst the practitioners about their own practices and interpretations of the young people's job interview performances, offering them invaluable means for continuous professional development – an affordance which has been taken forward by the practitioners participating in TARDIS in their practices beyond the life of the project, (Porayska-Pomsta, 2016).

7. Discussion and Conclusions

In 1941 Fromm argued that the rise of the Nazis was helped by the human tendency to not want to have too many choices, preferring to surrender the responsibilities for making decisions to the few and thus, leaving humans open to authoritarianism and ultimately fascism (Fromm, 1941). Throughout history, the consequences of such a surrender were profound for inclusion, tolerance, democracy, and for human life. Presently, with the rise of the 'intelligent machine' our social biases already ingrained in our systems have been acutely exposed. Feeding on the pre-existing data, AI has exposed our shockingly exclusive systems. As such it has also been shown to reinforce those biases and even as a tool to fuel social and political divide (Crawford, 2018). The application of AI as such a tool is aided, it seems, by the same ease as described by Fromm, with which we delegate our decision-making and choices to others.

This surrender of choice is not necessarily premeditated. Instead, we seem predisposed by nature to making decisions based on what we already know, rather than to processing new information. We are predisposed to choosing simpler strategies over those that require more effort to implement, i.e. we are by our very design lazy (Satpathy, 2012, Gavalas, 2014). According to Houdé (2013) we seem to lack cognitive inhibition in strategy selection between perceptual to logical brain, which on the whole requires us to make a heroic effort to engage in logical thinking, and often leads us to making decisions based on first impressions, to jumping to conclusions, and to acting parsimoniously (Epstein, 1984). If reinforced, our resistance to changing and to anything that opposes our beliefs and knowledge (Strebel, 1996; Gavalas, 2014) is bad news for inclusion, for our learning and development, and for our AI enhanced future. Given this view, the hazards of AI for society do not reside in AI *per se*. Instead, they are located in our propensity for parsimony in complex decision-making that seems amplified by the AI's unwavering ability to find optimal, rather than simplest, strategies in complex domains, releasing us from having to make an effort. With this in mind, accountability presents itself as a key pre-requisite of inclusion in human and AI-enhanced contexts, rendering the process of making oneself or itself accountable a mechanism for overcoming our parsimonious tendencies.

This is also where AI brings new and exciting opportunities in helping us challenge and question ourselves concretely, as a matter of habit, and also across time (since AI can make predictions about future events based on past occurrences). Such questioning has been shown to require advanced meta-cognitive competencies which are particularly beneficial to learning (Aleven & Koedinger, 2002; Richardson et al., 2012). As we discussed in this chapter, such competencies are also fundamental to inclusion, to our development of ethically balanced moral judgement and to our self-determination (Paul et al., 1990; Moshman, 2011), with positive implications for the excluding and the excluded. In section 5 we offered concrete examples from our own research, showing how the application of particular forms of AI (AI humanoid agent technology and open learner models) can act as a catalyst in our understanding of ourselves and of others, and how they can provide a much-needed mirror onto our systems, established ways of thinking, prejudices, and ultimately ignorance.

The purpose of such a mirror is not to shame us, but to support us in becoming more informed about ourselves, more confident at recognising when our systems fail to cater for our needs, and in taking cognisant steps to change. Sometimes, all that is needed is a safe space in which to rehearse situations that make us anxious or to provide such safe spaces to others in which we can witness their full potential. Sometimes we need a stepping stone or a medium to help us achieve this, something that can act as an unthreatening trigger for us to try out our strengths. AI, with is ability to emulate our own behaviours, while clearly being different from us, can give us just this, provided that we acknowledge that change has to come from us and not from AI's application alone. The outcomes can be revelatory to all concerned and may lead to changes in attitudes and support practices, as was the case in ECHOES. At other times, like in TARDIS, guided self-inspection is needed to empower learners to become self-efficacious, to self-reflect and to shed their inhibitions to share their reflections with others, while also offering the others a chance to see and to understand different perspectives and interpretations of the world. AI represents an increasingly powerful tool in this respect, through precise data and uncompromising, but nevertheless devoid of personal criticism (it's a machine after all!) interpretation thereof that aids concrete inspection and questioning of ourselves, and a platform for planning and rehearsing next steps. To be such a tool, however, AI must be designed in ways that allow its decisions to be explainable and interpretable by humans. Furthermore, to be educationally efficacious, it also needs to allow for an appropriate adaptive management of human vs artificial autonomy, with humans being given the possibility to challenge and to edit AI's interpretations of their data (Conati et al., 2018).

It is important to appreciate that the success of AI in the context of educational inclusion, as in the examples offered in this chapter, depends critically on an understanding that AI does not offer a solution *per se*. It is not a magic bullet to cure our ills, but rather, and more usefully, it offers a very strong lens through which we can study the extent to which our ideas of ourselves as an inclusive society match the reality on the ground, and a tool for simulating and rehearsing different states of the world and behaviours therein. In this, AI both facilitates our accountability and requires to be accountable itself to be truly an empowering learning tool for all.

As we have discussed extensively, accountability and inclusivity are *prima facie* frequently used concepts and have clear dictionary definitions, but delved deeper, truly workable definitions are not only hard to find, they also are *de facto* socially exclusive. By those definitions, the ways that we view and implement accountability and inclusion at the front line, are inflexible and slow to reflect our changing scientific knowledge, social understandings and aspirations. AI shows us that accountability and inclusivity are *processes* rather than 'set states', challenging our knowledge orthodoxies and putting to question our 'ground truths' (e.g. abnormality as a social construct vs. objective transparent criteria). It also offers ways in which inclusion as a social process can be democratised through empowering all stakeholders to own their data and influence how it is interpreted and shared. Viewed from this perspective, AI and educational inclusion shares a potentially compelling, mutually informing future worth investing in. However, in order for this future to become a reality AI cannot be a purely engineering solution. Instead it needs to be co-created by multiple stakeholders in a human-centred, socially contextualised way, whereby accountability of human and AI decision-making is built-in explicitly not only into AI, but also into the educational and social system within which AI is being applied.

References

Aleven, V.A., Koedinger, K.R.: An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. Cognit. Sci. **26**(2), 147–179 (2002).

Baron-Cohen, S. (2017). Editorial Perspective: Neurodiversity - a revolutionary concept for autism and psychiatry. *Journal of Child Psychology and Psychiatry, 58*(6), 744-747. doi:10.1111/jcpp.12703.

Bernardini, S., Porayska-Pomsta, K., Smith T J. (2014). ECHOES: An intelligent serious game for fostering social communication in children with autism, Information Sciences, 264, 41-60.

Brinkrolf, J. and Hammer B. (2018). Interpretable machine learning with reject option, De Gruyter Oldenbourg at – Automatisierungstechnik, 2018, Vol.66(4), pp. 283-290.

Bull, S. and Kay, J. SMILI: a framework for inter- faces to learning data in open learner models, learning analytics and related fields. *International Jour- nal of Artificial Intelligence in Education*, 26(1):293– 331, Mar 2016. ISSN 1560-4306. doi: 10.1007/s40593-015-0090-8. URL https://doi.org/10. 1007/s40593-015-0090-8.

Conati, C., Porayska-Pomsta, K., Mavrkis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling, CML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.

Crawford, K., and Calo, R. (2016). There is a Blind Spot in AI, Nature Comment, 538(7625).

Curry, AC and Reiser, V. (2018). #MeToo Alexa: How Conversational Systems Respond to Sexual Harassment, Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing, pp. 7-14, New Orleans, Louisiana, June 5 2018.

Davis, R., Shrobe, H., Szolovits, P., 1993. What is knowledge representation? AI Magazine *14*(1), 17–33.

Davis, R, J. (1996). What are Intelligence? And Why? 1996 AAAI Presidential Address, The American Association for Artificial Intelligence.

Dias, J. and Paiva, A. (2005). Feeling and Reasoning: A Computational Model for Emotional Characters. In Progress in Artificial Intelligence. Lecture Notes in Computer Science, Vol. 3808. Springer Berlin, Heidelberg, 127–140.

Dobson, S. D., & Brent, L. J. (2013). On the evolution of the serotonin transporter linked polymorphic region (5-HTTLPR) in primates. *Frontiers in human neuroscience*, *7*, 588.

Dubnick, M J. (2014). Toward an Ethical Theory of Accountable Governance. International Political Science Association meeting, July 19-24, Montreal

Epstein, R., 1984, "The Principle of Parsimony and Some Applications in Psychology", *The Journal of Mind and Behavior*, Volume 5, No. 2, pp. 119-130

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34, 906-911.

Gavalas, A. (2014). Brain Parsimony and its effects on decision making, *OKS Review, Vol. 3, No.1, EN, 1-14, March 2014*

Houdé, O. (2013). The *Psychology of a Child*, Vesta Editions, Thessaloniki

Jones, H., Sabouret, N., Damian, I., Baur, T., André, E., Porayska-Pomsta, K., Rizzo, P. (2014). Interpreting social cues to generate credible affective reactions of virtual job interviewers, IDGEI 2014, ACM, arXiv preprint arXiv:1402.5039

Lai, E R. (2011). Metacognition: A literature review, Research Report, Pearson, https://images.pearsonassessments.com/images/tmrs/Metacognition_Literature_Review_Final.pdf.

Lipton, Z. and Steinhardt, J. (2018). Troubling Trends in Machine Learning Scholarship, ICML 2018: The Debates, 2018arXiv180703341L.

Moshman, D. (2011). Adolescent Rationality and Development, Routledge.

O'Neil, S. (2008) The meaning of autism: beyond disorder, Disability & Society, *23*(7), 787-799, DOI: 10.1080/09687590802469289.

Paul, R W., and Binkler, J A. (1990). Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World. Rohnert Park, CA: Center for Critical Thinking and Moral Critique.

Porayska-Pomsta, K., Alcorn, A M., Avramides, K., Beale, S., Bernardini, S., Foster, M-E., Frauenberger, C., Pain, H. Good, J., Guldberg, K., Kea-Bright, W., Kossyvaki, L., Lemon, O., Mademtzi, M., Menzies, R., Rajendran, G., Waller, A., Wass, S., Smith, T J. (2018). Blending human and artificial intelligence to support autistic children's social communication skills, ACM Transactions on Human-Computer Interaction, in press.

Porayska-Pomsta, K. and Chryssafidou, E. (2018), Adolescents' Self-regulation During Job Interviews Through an AI Coaching Environment, International Conference on Artificial Intelligence in Education, 281-285, Springer Cham.

Porayska-Pomsta, K. (2016). AI as a methodology for supporting educational praxis and teacher metacognition, International Journal of Artificial Intelligence in Education, Vol.26(2), 679-700.

Porayska-Pomsta, K., Rizzo, P., Damian, I., Baur, T., André,E., Sabouret, N., Jones, H. Anderson, K., Chryssafidou, E. (2014), Who's afraid of job interviews? Definitely a question for user modelling, International Conference on User Modelling, Adaptation and Personalization, 411-422, Springer Cham.

Porayska-Pomsta, K and Bernardini, S. (2012). Learner Modelled Environments, Handbook of Digital Technology Research, S. Price, C. Jewitt, B. Brown (eds.), Sage.

Prizant, B.M., Wetherby, A.M., Rubin, E. and Laurent, A.C. (2003). The SCERTS model: A Transactional, Family-Centered Approach to Enhancing Communication and Socioemotional Ability in Children with Autism Spectrum Disorder. Infants and Young Children 16, 4 (2003), 296–316.

Prizant, B.M., Wetherby, A.M., Rubin, E., Laurent, A.C. and Rydell, P.J. (2006). The SCERTS® Model: A Comprehensive Educational Approach for Children with Autism Spectrum Disorders, Brookes.

Rajendran, G. (2013). Virtual environments and autism: a developmental psychopathological approach. *Journal of Computer Assisted Learning*, *29*(4), 334-347. doi:10.1111/jcal.12006.

Reisman, D., Schultz, J., Crawford, K., Whittacker, M. (2018). Algorithmic Impact Assessments: A Practical Fremwork for Public Agency Accountability, AI Now Institute Report, April 2018.

Remington, A. (11 July 2018). Autism can bring extra abilities and now we're finding out why. *New Scientist*. https://www.newscientist.com/article/mg23931860-200-autism-can-bring-extra-abilities-and-now-were-finding-out-why/

Richardson, M., Abraham, C., Bond, R.: Psychological correlates of university students' academic performance: a systematic review and meta-analysis. Psychol. Bull. **138**(2), 353 (2012).

Russell, S.J. and. Norvig, P. (2003). Artificial Intelligence: A Modern Approach. Prentice Hall. Second Edition.

Satpathy, J., 2012, "Issues in Neuro-Management Decision making", *Opinion: International Journal of Business management*, Vol. 2, No 2, pp 23-36.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y. Lillicrap, T., Hui, F. Sifre, L. van den Driessche, G., Graepel, T. and Hassabis, D. (2017). Mastering the Game of Go Without Human Knowledge, *Nature* volume550, pages354–359 (19 October 2017)

Strebel, P., 1996, "Why Do Employees Resist Change?", edit. in Harvard Business Review on Change, USA, pp.139-157.

Sutton, R S. and Barto, A G. (2000). Reinforcement Learning: An Introduction, The MIT Press.

Terricone, P. (2011). The Taxonomy of Metacognition. Psychology Press.

Weizenbaum, Joseph (1976). Computer power and human reason: from judgment to calculation. W. H. Freeman.

Woolf, B, (2008). Building Intelligent Tutoring Systems. Morgan Kaufman.