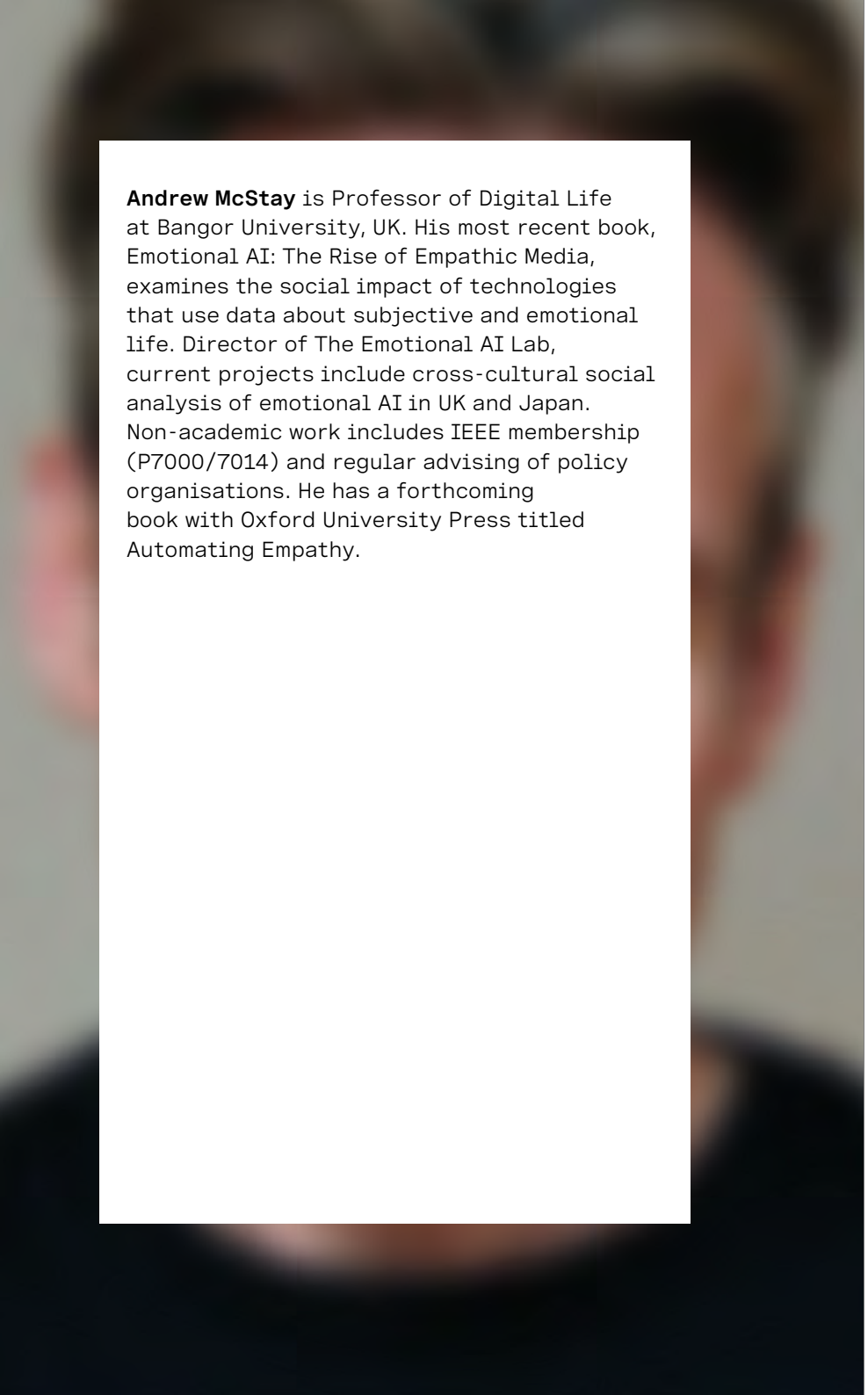

Andrew McStay is Professor of Digital Life at Bangor University, UK. His most recent book, Emotional AI: The Rise of Empathic Media, examines the social impact of technologies that use data about subjective and emotional life. Director of The Emotional AI Lab, current projects include cross-cultural social analysis of emotional AI in UK and Japan. Non-academic work includes IEEE membership (P7000/7014) and regular advising of policy organisations. He has a forthcoming book with Oxford University Press titled Automating Empathy.

Automated empathy in education: benefits, harms, debates

Andrew McStay, Bangor University†

This essay assesses what might appear a niche interest: the use of technologies to gauge emotion expressions, and whether a child is attentive and engaged. Variously labelled 'affective computing', 'emotion AI' and 'emotional AI', I use the meta-label, *automated empathy*, to cluster a variety of systems programmed to identify, quantify, judge, respond and interact with emotions, affective states, cognitive states, attention and intention (McStay, 2022: forthcoming). The essay considers claimed benefits and problems of these technologies, disputed usefulness in learning and educational development, and ethical questions about acceptability of using technologies in education this way, progressing to discuss these issues in the context of General Comment No. 25 (GC25) on the United Nations Convention on the Rights of the Child (UNCRC).

EdTech innovation in machine social and emotional learning

Educational technologies (EdTech), tutoring apps and related systems have scope to improve life chances through education, especially in regions where formal schooling is difficult (e.g.,

† This work is supported by the Economic and Social Research Council [grant number ES/T00696X/1].

because of distance from schools, violence in conflict areas, gender discrimination, cost of materials, overcrowding, poor curricula and inadequate teachers). Yet, where EdTech systems and automated empathy overlap, there are significant concerns, as this essay unpacks. As social and emotional learning balances cognitive elements in education (knowledge acquisition, analysis, reasoning and memory) with management of feelings and emotions, perseverance to achieve goals and ability to work with others, EdTech can help with the non-cognitive dimensions of learning. A stated benefit is that in the context of growing classrooms, all children receive close and recorded attention.

Companies building these systems include many start-ups and established technology companies. For example, the education branch of the global technology company Intel states that they are researching how recognition of emotion and disposition may help personalised learning. While not yet present in UK school classrooms, Intel Education's (2022) take on automated empathy involves three inputs to a classroom computer that records and predicts engagement during a class session. Inputs include *appearance*, where cameras extract facial landmarks, upper body and head movement and pose; *interaction* and how the student uses input devices such as a keyboard and mouse; and *time to action*, or how long the student is taking to complete tasks or act on a learning platform.

In addition to interest in deploying automated empathy in the physical classroom is interest in the virtual classroom, something of increased attention since the beginning of the COVID-19 pandemic. The EdTech industry saw an opportunity in home learning in that it recognised that embodied and interpersonal dynamics of in-classroom empathy were missed during the height of the COVID-19 pandemic. As a remedy, many start-ups and legacy technology firms suggested that their services could be used to gauge emotion, attention and interest. Intel (again), for example, have teamed with Classroom Technologies to develop 'Class' that runs on top of Zoom. This is claimed to detect whether students in the USA are bored, distracted or confused by assessing student facial expressions in relation to the educational content they are studying (Class, 2022).

Looking slightly further ahead to other automated empathy technologies that have realistic scope to be used in schools is metaverse-based interest in education, with Microsoft actively developing in this area. Scepticism of the metaverse is fair, but non-biometric mixed reality (through virtual reality [VR] and augmented reality [AR] systems) already has a growing presence in education, enabling students to 'feel-into' places, pasts, presents and futures, and even the constitution of objects (Microsoft Education, 2022). There is genuine value for students and teachers in this sort of non-biometric digital empathy in education, be this regarding the tangible impacts of climate change in faraway places, of biological systems, or of historical situations.

However, biometric profiling of facial expressions and student in-world interaction is looking likely through Microsoft's 'Mesh for Teams' that uses biometrics to map bodily behaviour and physical facial expressions onto in-world avatars, for novel 'Teams' meetings. Although 'Mesh for Teams' is aimed at the world of work (Roach, 2021), the potential for education is clear, through the existing presence of Microsoft Teams for online teaching. The goals of Intel, Microsoft and others are multiple, including quantification of what were qualitative phenomena, and providing an evidence base for education established on numbers, novel interactivity and performance metrics other than assignment scores and attendance.

For educators unfamiliar with biometric-automated empathy these technologies may have appeal, although lessons may be gleaned from countries that have practical actual experience with it. China's pilot tests with emotion profiling and automated empathy in the classroom are instructive. In reference to the Class Care System (CCS) from Hanwang Education, an extended report by human rights organisation ARTICLE 19 (2021) found that students feign interest and game the system to receive rewards. Self-policing may become an everyday occurrence, but 'chilling effects' (i.e., self-censorship) and being 'always-on' is not the sort of mindfulness we should be introducing. This is especially so given that students will perform for how they think the camera sees them (Andrejevic & Selwyn, 2019).

Indeed, van der Hof et al. (2020) see this through the prism of human dignity, because automated systems risk making de-individualised decisions without respecting the full and intrinsic worth of a human being. In China itself, tests with emotion-based automated empathy show it to be neither popular with students nor teachers, in part due to privacy questions, but also the lack of actionable feedback (ARTICLE 19, 2021). Does inattentiveness to part of a lesson, for example, signal boring content or boring delivery? Furthermore, the systems offer no suggestion on how to improve these. This is not to foreground practical over ethical concerns, but to provide a sense of limitations in practice.

Historical development

Use of technologies to gauge emotion and human disposition has a surprisingly long history, originating in the 1800s. Technically, this entailed pre-digital tracking of emotion by measuring temperature differentials, changes in heartbeat, blood pressure, breathing, conductivity of the skin and brain activity measures and facial coding, among other signals (Dror, 2001). How interiority was represented is also notable, as emotions were formalised into tables, charts and curves. Debates on positivism versus socially grounded understanding of emotion are beyond the limit of this essay, but emphasis on visualisation resonates with the modern educator usage of dashboards. Skipping centuries, cybernetic and computational apps to emotion have theoretical roots in the 1970s, with Manfred Clynes who argued for physical laws of emotion and its communication that could be rendered by computers. 'Sentics' for Clynes would help children 'be in touch with their emotions' and allow 'different races and backgrounds to experience their common basis in humanity' by being sensitive to the emotions of others (1977, p. xxii).

Rosalind Picard, the originator of the term and practice of 'affective computing', tried to put this into practice by building a 'computerized learning companion that facilitates the child's own efforts at learning' (Picard et al., 2001). The goal of the companion was to improve pedagogical techniques by using computer vision techniques to watch and respond to the affective states of children. By the 2010s, Sidney D'Mello's 'Affective AutoTutor' would detect and respond to learners' boredom, confusion and frustration. Through facial coding, and tracking of interaction patterns and body movement, this system sought to provide motivational feedback to students through appropriate facial expressions and voice emotion (D'Mello & Graesser, 2012). Related work was based on voice that, in addition to assessing whether verbal answers are correct, also seeks to detect learners' certainty or uncertainty (Forbes-Riley & Litman, 2011). Other work focuses on attention rather than emotion. Mention should also be made of teachers, as their teaching methods may also be subject to analysis through recording of in-classroom audio and automated methods to predict the level of discussions in these classes (D'Mello, 2017).

Pseudoscience?

Leading industry figures recognise the limitations of popular 'basic emotion' recognition technologies, with Microsoft publishing work in academic and technology journals saying so (McDuff & Czerwinski, 2018). Yet, despite Microsoft's own researchers publishing on this issue, for years this did not stop Microsoft from using this approach. Microsoft's Azure service, for example, labels 'basic emotion' facial expressions as happiness, sadness, neutral, anger, contempt, disgust, surprise and fear (Microsoft Education, 2022). Testament to the controversial nature of this approach to emotion recognition, Azure is slated to be discontinued in 2023 through publication of Microsoft's framework for building AI systems responsibly (Crampton, 2022). While this was widely interpreted to mean that Microsoft would desist from all work on emotion recognition, this is not what they said. Retirement of inference of emotional states applies only to their Azure Face services, with Microsoft adding that they 'need to carefully analyze *all* AI systems that purport to infer people's emotional states' (Crampton, 2022). This is a much weaker statement of intent than 'we have stopped all emotion recognition development'.

Despite Microsoft's retirement of emotion-based services in Azure, the method is popular. The Google Cloud Vision API, for example, also uses face landmark regions (e.g., mouth and eyebrows) to 'detect emotion' (Google Cloud, 2022).

There is a long line of scholars who will testify to this approach being a highly limited account of emotions, and that using 'reverse inference' to infer experience from expressions is questionable (Stark & Hutson, 2021). Adding to these voices, Barrett et al. (2019) observe that facial coding is especially poor with children, due to their immaturity and lack of development in emoting (also see McStay, 2019).

The reason why companies use simplistic approaches is simple: expedience. It is relatively easy to program systems to look for features (such as movement and actions of faces) and then match these arrangements to pre-given emotion expression labels. To question whether the full gamut of emotional life can be channelled through a suite of basic emotions, or whether expressions say much about experience, would add a lot of complexity for global technology firms seeking to deploy their products internationally. A universalist account of emotional life and subjectivity suits them well.

Despite highly vocal critique of claims of pseudoscience, this is not the core problem. A risk of a pseudoscience-based critique is that it invites *more* profiling and more granular labelling of brain, bodily and situational interactions. This would involve the connection of facial movements with factors connected to the personal and external contexts. For the person, it would include metabolic and historic dimensions (relating to the body and existing profiles of a person), and external factors including regional and societal norms on emoting, and specifics of the situation where the sensing is taking place (McStay & Urquhart, 2019). For example, is a child at home, in school, in virtual space, or in a mixed reality context? Who else is present? What is the situation? Who is teaching?

There are also accuracy problems - not only in psychological assumptions about the nature of emotion, but also in the curation of training datasets (regarding who does the labelling of an emotion expression and who is labelled). Overlapping with general concerns about AI bias against marginalised groups, market leading systems such as Microsoft and Chinese company Face++ have been found to label Black people with disproportionately negative types of emotion (notably, anger), especially if there is ambiguity of

what emotion label to give to a facial expression (Rhue, 2018). In work at our Emotional AI Lab, we tried to examine training datasets in terms of how they are constructed, who is doing the labelling, who is being labelled and the nature of this emotion profiling in relation to transport and usage in cars, but we found this to be an opaque and secretive practice as companies closely guard how their systems work (McStay & Urquhart, 2022). This is not to say that they are guarded because they are biased, but that industrial secrecy means that they are not open for public examination, despite social risks.

Child rights policies

The bundling of deeply questionable technologies with pro-social ambition risks lack of critical scrutiny. For example, internationally, pro-social emphasis on 'soft' abilities is something that influential bodies, such as UNESCO, see as 'fundamental to human creativity, morality, judgment, and action to address future challenges' (UNESCO, 2021, p. 68); but other key organisations see scope to instil these so-called soft abilities through questionable means, with the Organisation for Economic Co-operation and Development (OECD) seeing utility in measuring child sociality and emotion through affective computing (OECD, 2015). This illustrates the observation made by critical EdTech scholarship that rightly notes that datafication of emotion serves the overall education policymaking process around social and emotional learning, rather than children, through building of a psychometric evidence base (Williamson, 2019).

There are, of course, wider ethical and governance concerns. With an explicit focus on emotion and affect-based technologies, historically these have been under-served by tools such as the EU and UK General Data Protection Regulation (GDPR), which make no reference whatsoever to emotions. Similarly, the European proposal for the ePrivacy regulation rarely mentions emotions. Only Recitals 2 and 20 of the ePrivacy preamble mention emotions although, importantly, Recital 2 defines them as highly sensitive (McStay & Rosner, 2021). However, this lacuna is on absence of emotion profiling regulation being noted. In 2021 the United Nations Human Rights Council (UNHRC) formally adopted

the Resolution titled 'Right to privacy in the digital age' where §3 notes the need for safeguards for emotion recognition (UN General Assembly, 2021). More regionally, and with application to children, the Council of Europe (2021) likewise calls for strict limitations and bans in areas of education and the workplace. Also in 2021, the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS) issued a joint statement declaring the use of AI to infer emotions of a natural person as highly undesirable, and that it should be prohibited, except for specified cases, such as for some health purposes (EDPB, 2021).

Relatedly, 2021 also saw the release of the proposed EU AI Act, a risk-based piece of legislation that classifies emotion recognition usage with children (such as in toys as well as education) as high risk (European Commission, 2021). Notably, Recital 28 of the proposed EU AI Act names the UNCRC and General Comment No. 25 that expands on rights regarding the digital environment (also see Articles 5.1b and 9 of the proposed EU AI Act). The UK itself does not have bespoke regulation on emotion profiling and children, although the Centre for Data Ethics and Innovation (part of the Department for Digital, Culture, Media & Sport) sees use of biometric data such as eye tracking, facial expressions and affective states as a way of improving understanding of levels of engagement and educational resource design (GOV.UK, 2021).

Of special interest to this essay on automated empathy and education is General Comment No. 25 (GC25) 2021 update to the UNCRC, especially because it recognises opportunities in new technologies, as well as seeking to define and defend rights. Lacklustre in name only, GC25 details how child rights in the digital environment should be interpreted and implemented by States around the world. The Emotional AI Lab responded to the call for evidence for GC25 (McStay et al., 2021). Unique among evidence provided to the call, we focused on harms associated with datafied emotion in education and toys. It appears that we were heard as GC25 contains multiple mentions of emotion analytics (§42, 62, 68), finding this to interfere with children's right to privacy, and freedom of thought and belief, also flagging the importance 'that automated systems or information filtering systems are not used to affect or influence children's behaviour or emotions or to limit their opportunities or development' (UNCRC, 2021, §62).

In relation to education itself, the digital environment is seen in the published GC25 as providing scope for 'high-quality inclusive education, including reliable resources for formal, non-formal, informal, peer-to-peer and self-directed learning' (§99), also with potential 'to strengthen engagement between the teacher and student and between learners' (§99). Surface consideration might see this as making the case for biometric-automated empathy for generalised use in the classroom, in platform-based learning and in an immersive VR/metaverse context. This would involve rendering physical facial expressions onto in-world avatars and profiling in-world interactions between students and teachers for future reference. Yet, as detailed, there are deep methodological and discriminatory problems that mitigate against this reading of GC25 in relation to emotion, immersive media and biometrics.

However, non-biometric mixed reality (through VR and AR systems that enable students to 'feel-into' places, pasts, presents, objects and imagined futures) is seen here as having pedagogic value, especially when used to deepen and enrich understanding of a topic (Daniela, 2020). In this regard, §101 is also notable, seeking to ensure that the 'use of digital technologies does not undermine in-person education and is justified for educational purposes', which points to an intrinsic belief of the value of in-person learning (and human-teacher empathy therein) and that promises of automated empathy for platform-based learning should not be allowed to undermine in-person embodied interaction.

Finally, §103 is also of keen relevance, specifying that standards for digital educational technologies should ensure that child personal data is not misused, commercially exploited or otherwise infringes their rights. Concern about datafied exploitation of children is longstanding, especially in relation to marketing and advertising (van der Hof et al., 2020), but this is extended by automated empathy in EdTech in two ways.

First, because in the context of automated empathy in education, inferences about students' emotions are used to train the neural networks owned by EdTech providers for purposes outside of education. Consequently, aggregated data

about child emotion would be commodified to improve algorithmic services, create competitive difference (in terms of how many faces are analysed) and serve business and strategic contexts for which the student data was not intended (such as testing responses to ads, linking well to van der Hof et al., 2020).

Second is emphasis of §103 on *personal* data, rather than simply child data. This is significant because in many instances automated empathy vendors will argue that their systems only deal in aggregate impressions (such as overall levels of pupil attention and happiness), and that data collected cannot be linked back to an individual. There is good technical legal debate in that personal data must exist in these systems for a fragment of a second as the 'insight' is collected and aggregated (George et al., 2019), but in practice, this has not stopped use of this approach to emotion recognition in out-of-home advertising in Europe (McStay, 2020).

This essay recommends critical attention to aggregated as well as identifying practices, especially given the scope for chilling effects, self-censorship and surveillant experience of being 'always-on'. After all, this is the antithesis of the social and emotional learning that should take place in education. Moreover, the moral basis for recommended critical attention is not that aggregated data about students may conceivably be personal data due to the fraction of a second processing of personal data. Although it should be noted that EU and UK data protection does not prescribe a minimum amount of time personal data should exist within a data processing system for it to be governed by legal rights over personal data, the moral basis argued here is that privacy and related rights may be held by a group as well as individuals (Floridi, 2014; Wachter, 2020).

Conclusion

Between rights to freedom of thought, privacy and access to education, there is the glaring question of whether automated empathy can do what is claimed. However, even with improvements in methodology, automated empathy in education does not align with the need for mental and emotional reserve to ensure human flourishing. This essay concludes that automated empathy technologies are incommensurable with current and near future social values. The core methodological and normative problems are as follows:

- Serious questions about effectiveness, validity and social representativeness of training data
- Lack of alignment between financial incentives in automated empathy and the wellbeing of schoolchildren
- Moral problems in using aggregated inferences about children's emotions to train neural networks that will be deployed for other commercial purposes
- Mission creep, where in-class data may be used for other socially determining purposes (such as social scoring)
- Already demonstrated risk of self-surveillance and chilling effects in the classroom
- Data minimisation questions that ask whether automated empathy is necessary for successful education.

Andrejevic, M., & Selwyn, N. (2020). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology, 45*(2), 115-128

ARTICLE 19 (2021). *Emotional entanglement: China's emotion recognition market and its implications for human rights*

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1-68

Class. (2022). *Add teaching & learning tools to Zoom*. www.class.com

Clynes, M. (1977). *Sentics: The touch of the emotions*. Anchor Press/Doubleday

Council of Europe. (2021). *Consultative Committee of the Convention for the Protection of Individuals with regard to automatic processing of personal data. Convention 108: Guidelines on facial recognition*

Crampton, N. (2022). Microsoft's framework for building AI systems responsibly

Daniela, L. (2020). *New perspectives on virtual and augmented reality*. Taylor & Francis

D'Mello, S. K. (2017). Emotional learning analytics. In C. Lang, G. Siemens, W. Alyssa, & D. Gašević (Eds.), *Handbook of learning analytics and educational data mining* (pp. 115-127). Society for Learning Analytics Research.

D'Mello, S. K., & Graesser, A. (2012). AutoTutor and Affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems, 2*(4), 23, 1-39

Dror, O. E. (2001). Counting the affects: Discoursing in numbers. Social Indicators Research, 68(2), 357-378

EDPB (European Data Protection Board). (2021). *EDPB–EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. 18 June

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. 21 April

Floridi, L. (2014). Open data, data protection, and group privacy. *Philosophy & Technology*, 27, 1-3

Forbes-Riley, K., & Litman, D. J. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication, 53*(9-10), 1115-1136

George, D., Reutimann, K., & Tamò-Larrieux, A. (2019). GDPR bypass by design? Transient processing of data under the GDPR. *International Data Privacy Law, 9*(4), 285-298

Google Cloud. (2022). *Vision AI*

GOV.UK. (2021). *AI Barometer Part 5 – Education*

Intel Education. (2022). *Applying artificial intelligence to transform how we learn*

McDuff, D., & Czerwinski, M. (2018). Designing emotionally sentient agents. *Communications of the ACM, 61*(12), 74-83

McStay, A. (2019). Emotional AI and EdTech: Serving the public good. *Learning Media & Technology, 45*(3), 270-283

McStay, A. (2020). Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy, *Big Data & Society*

McStay, A. (2022: forthcoming). *Automating empathy: When technologies claim to feel-into everyday life*. Oxford University Press

McStay, A., & Rosner, G. (2021). Emotional Artificial Intelligence in children's toys and devices: Ethics, governance and practical remedies. *Big Data & Society, 8*(1)

McStay, A., & Urquhart, L. (2019). 'This time with feeling?' Assessing EU data governance implications of out of home appraisal based emotional AI. *First Monday, 24*(10)

McStay, A., & Urquhart, L. (2022). In cars (are we really safest of all?): Interior sensing and emotional opacity. *International Review of Law, Computers & Technology*

McStay, A., Miyashita, H., Rosner, G., & Urquhart, L. (2021). *Comment on children's rights in relation to emotional AI and the digital environment*

Microsoft Azure. (2018). Face API

Microsoft Education. (2022). Face detection and attributes

OECD (Organisation for Economic Co-operation and Development). (2015). *Skills for social progress: The power of social and emotional skills*. OECD Skills Studies, OECD Publishing

Picard, R., Cassell, J., Kort, B., Reilly, R., Bickmore, T., Kapoor, A., Mota, S., & Vaucelle, C. (2001). *Affective learning companion*

Rhue, L. (2018). Racial influence on automated perceptions of emotions. *SSRN*, 9 November

Roach, J. (2021). Mesh for Microsoft Teams aims to make collaboration in the 'metaverse' personal and fun

Stark, L., & Hutson, J. (2021). Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal, SSRN*. 20 September

UNCRC (United Nations Convention on the Rights of the Child). (2021). *General Comment No. 25 (2021) on children's rights in relation to the digital environment*

UNESCO. (2021). *Reimagining our futures together: A new social contract for education*

UN (United Nations) General Assembly. (2021). *Resolution adopted by the Human Rights Council on 7 October 2021. 48/4. Right to privacy in the digital age*

van der Hof, S., Lievens, E., Milkaite, I., Verdoodt, V., Hannema, T., & Liefaard, T. (2020). The child's right to protection against economic exploitation in the digital world. *The International Journal of Children's Rights, 28*(4), 833-859

Wachter, S. (2020). Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal, 35*(2)

Williamson, B. (2019) Psychodata, 7 October