Emotional Artificial Intelligence, Emotional Surveillance, and the Right to Freedom of Thought

Robbie Scarff

April 24, 2021

Emotional Artificial Intelligence, Emotional Surveillance, and the Right to Freedom of Thought

Robbie Scarff

(PhD Candidate, University of Edinburgh, School of Law)

1.0 – Introduction

In today's increasingly digitised societies, new devices and methods of analysis are making ever more aspects of people's lives machine readable, while organisations make decisions, from the trivial to the significant, based on data. Recent developments in computer vision and machine learning have facilitated a surge in interest in digitising people's *emotions*. This paper makes a novel contribution by analysing such emotional surveillance in terms of its impact on the human right to freedom of thought, suggesting ways in which this right may be violated. Furthermore, it applies theoretical insights from surveillance studies to develop a deeper understanding of emotional surveillance, including what motivates it, how it operates, and what its consequences are.

Surveillance studies scholars have been critiquing the 'surveillance society' for a long time, raising many concerns about the harms surveillance can cause.[1] As the range of technologies and their associated affordances expands, this practice of 'questioning surveillance' continues.[2] Generally, such work raises concerns about increasingly detailed 'digital personas' that facilitate surveillance,[3] leading to associated harmful practices like discriminatory decision-making and micro-targeting of content.[4] Pertinent to this discussion, surveillance is also said to, 'affect the way individuals … think',[5] and violate their intellectual privacy.[6] While scholars argue that surveillance *can* lead to a wide array of individual and societal harms, the exact nature and degree of harm differs depending on specific factors relating to the particular instance of surveillance activity.[7] Thus each new form of surveillance, such as emotional surveillance, must be investigated thoroughly to determine its particular set of effects. To this end, scholars have long been formulating lists of questions to critique surveillance practices in an effort to elucidate their nature and determine their legitimacy.[8] A key theme in this questioning approach is asking whether the surveillance system causes *harm*.[9]

[1] Gary T. Marx, 'The surveillance society: The threat of 1984-style techniques' [1985] The Futurist, 6; David H. Flaherty, 'The emergence of surveillance societies in the western world: Toward the year 2000' (1988) 5(4) Government Information Quarterly < https://bit.ly/3bctMw1 > accessed 17 January 2021; Oscar H. Gandy, 'The surveillance society: Information technology and bureaucratic social control' (1989) 39(3) Journal of Communication < https://bit.ly/2Nx2rwC > accessed 17 January 2021

[2] David Wright, Rowena Rodrigues, Charles Raab, Richard Jones, Ivan Szekely, Kirstie Ball, Rocco Bellanova and Stine Bergersen, 'Questioning Surveillance' (2015) 31(2) Computer Law and Security Review < https://bit.ly/3dmyQko > accessed 17 January 2021

[3] Roger Clarke, 'The digital persona and its application to data surveillance' (1994) 10(2) The Information Society < https://bit.ly/3auDF96 > accessed 17 January 2021. For a discussion of the process by which various methods are used to collect data about individuals, see also Roger Clarke, 'Risks inherent in the digital surveillance economy: A research agenda' (2019) 34(1) Journal of Information Technology < https://bit.ly/3biw9O9 > accessed 17 January 2021 63 – 66

[4] Clarke, 2019 (n 3) 67

[5] Wright et al., (n 2) 282

[6] Neil M. Richards, 'The Dangers of Surveillance' (2013) 126(7) Harvard Law Review < https://bit.ly/3dk2kPq > accessed 29 January 2021

[7] Individual harms are said to include many forms of discriminatory decision-making, decision-making using incorrect data, chilling effects, and restrictions on self-determination, see Clarke, 2019 (n 3) 68. Societal harms are said to include influencing electoral results and suppressing scientific and economic progress, as well as degrading trust and social cohesion, see Clarke, 2019 (n 3) 68 – 69 and Wright et al., (n 2) 282.

[8] Gary T. Marx, 'Ethics for the new surveillance' (1998) 14(3) The Information Society < https://bit.ly/3biHNIR > accessed 17 January 2021; Wright et al., (n 2) 283 – 287; David Wright and Charles Raab, 'Constructing a surveillance impact assessment' (2012) 28(6) Computer Law and Security Review < https://bit.ly/3avksEB > accessed 17 January 2021. See also, Surveillance Studies Network, 'A Report on the Surveillance Society' (2006) Office of the Information Commissioner

[9] Marx, 1998 (n 1) 174; Wright et al., (n 2) 6

Although detailed lists of questions are a valuable contribution, they provide little guidance as to the most appropriate strategy for *answering* such questions.

The urgency with which answers to such questions are required has increased in recent years. Surveillance capabilities have expanded due to the development of AI systems, prompting increased concern about the harms such systems pose to human rights,[10] as well as renewed effort on the foregoing question of how exactly to determine the harms of algorithmic surveillance systems. Against the background of, and perhaps due to, a range of different methodologies for identifying and mitigating the potential harms of technology, such as the ethical, legal, and social issues (ELSI) or responsible research and innovation (RRI) approaches,[11] McGregor et al. argue that there is no agreed upon *method* for determining the harm caused by AI systems, let alone agreement about the harm caused, meaning diverse actors can use their own understanding of harm that suits their needs and which may not fully account for international human rights law (IHRL) or provide effective remedies.[12] They therefore propose that IHRL itself provides a useful framework for assessing algorithmic systems as it provides broadly accepted methods for defining harms, dictates the obligations and expectations of States and businesses respectively, and can address all stages of algorithmic development and deployment.[13]

In summary, technological development and the process of digitising people's lives manifests in many ways, one of which is efforts to make people's emotions machine readable. The harms caused by surveillance practices are complex and unique to particular types of surveillance, and while surveillance studies asks many questions of surveillance practices, how to answer those questions remains open. Alongside concerns about AI-enabled surveillance systems, there has been recent interest in the impact of AI and data processing on the right to freedom of thought. This paper therefore brings these related concerns about AI, surveillance and freedom of thought together by focussing on emotional surveillance, applying the aforementioned IHRL approach as a framework for analysis. To clarify, "emotional surveillance" here refers specifically to *automated* emotional surveillance *by* technology.

The paper begins by explaining how emotional artificial intelligence (EAI) works and flagging the highly contested assumptions it relies upon. Section 2 then provides a number of examples of emotional surveillance. Section 3 draws on surveillance studies literature to analyse emotional surveillance, identifying its causes, courses, and consequences. Section 4 focusses on the impact of emotional surveillance on the right to freedom of thought. It begins by setting out the content of the right, consulting relevant case law and soft law, before looking in detail at the interaction between emotional surveillance and three key elements of the right: the right not to reveal one's thoughts, the right not to have one's thoughts manipulated, and the right not to be penalised for one's thoughts. Section 4 closes with a critique of the right to freedom of thought in IHRL.

[10] Council of Europe Committee of Experts on Internet Intermediaries (MSI-NET), 'Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' (2018) Study DGI(2017)12; Raso et al., 'Artificial Intelligence & Human Rights: Opportunities & Risks' (2018) Berkman Klein Centre for Internet & Society Research Publication; Susie Alegre, 'Rethinking Freedom of Thought for the 21st Century' (2017) 3 European Human Rights Law Review < https://bit.ly/2QqH7ds > accessed 3 February 2021

[11] Bernd Carsten Stahl, 'Responsible research and innovation in information systems' (2012) 21(3) European Journal of Information Systems < https://bit.ly/3adANNC > accessed 9 February 2021

[12] Lorna McGregor, Darragh Murray and Vivian Ng, 'International human rights law as a framework for algorithmic accountability' (2019) 68(2) The International and comparative law quarterly < https://bit.ly/3b9Q0Pk > accessed 17 January 2021 323 – 324

[13] Ibid, 324 – 325

1.1 – Emotional Artificial Intelligence: Theory and Controversy

EAI describes technology which attempts to interpret the emotional state of human beings. EAI is rooted in the field of affective computing, that is, "computing that relates to, arises from, or influences emotions", and relies on recent advances in machine learning (ML). EAI uses ML to identify patterns in training data then utilises this learning to interpret new data wherever the technology is deployed.[14] Stated simply, EAI takes input data such as speech, text, images, or heart rate, analyses that data using ML, and produces output data such as classified emotional state (e.g., happy, sad, angry, disgusted, scared, surprised, distressed), valence (e.g., pleasant or unpleasant), arousal level (e.g., calm or agitated) and confidence score.[15] This process describes EAI that is explicitly trying to identify emotion, however, emotional surveillance can also occur when ML is used to detect patterns of correlation between behaviours and prior signals, which are then subsequently interpreted as relating to emotion. For example, a pattern of liking particular content on Facebook could retrospectively be identified as revealing anger at that content.

Before proceeding, a quick overview of the highly contested nature of the science of emotions is provided. There are two opposing theories that attempt to explain emotions: the classical view and the theory of constructed emotions (TCE). The classical view has been around the longest, is widely held, and, crucially for this paper, *underpins* EAI. In contrast, the TCE is more recent, not as well known, and challenges the classical views assumptions.

On the one hand, the classical view asserts that each emotion category has a particular 'fingerprint' or physiological process that occurs in the brain and can be identified each time an emotion occurs. Furthermore, words like happiness or fear are used to describe both the emotion *and* what happens inside the brain.[16] This positivist view of emotions is informed by essentialism, a position which holds that categories like happiness and fear not only exist but have a central *essence* which makes them what they are.[17] An influential version of the classical theory, basic emotion theory, assumes that stimuli activate the essences of the six basic emotions of happiness, sadness, fear, surprise, anger and disgust.[18] It is basic emotion theory which underpins the Facial Action Coding System used in face based EAI to categorise facial micro-expressions.[19] While variations exist, such as classical appraisal theory,[20] *all* classical theories of emotion are premised on emotions having distinct essences.[21] It is for this reason that the classical view underpins not just face based EAI but the very rationale of the full panoply of EAI applications.

On the other hand, the TCE paints a radically different picture of what emotions are. The TCE argues that emotions do *not* have unique 'fingerprints', pointing to four meta-analyses of physiology studies which found no evidence for such emotion mechanisms.[22] Rather than being *triggered* by external

[14] Andrew McStay and Lachlan Urquhart, ''This time with feeling?': Assessing EU data governance implications of out of home appraisal based emotional AI' (2019) 24(10) First Monday < https://bit.ly/3tgwB78 > 2

[15] Ibid

[16] Lisa Feldman-Barrett, *How Emotions Are Made. The Secret Life of the Brain* (first published 2017, Pan Books/Macmillan 2018) 34

[17] Ibid 157

[18] Ibid 158

[19] Paul Ekman and Wallace V. Friesen, *Facial Action Coding Systems* (Consulting Psychologists Press 1978)

[20] Klaus R. Scherer, Angela Schorr and Tom Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research* (Oxford University Press 2001) 3-20

[21] Feldman-Barrett, 2018 (n 16) 158

[22] John T. Cacioppo et al., 'The Psychophysiology of Emotion' In Michael Lewis and Jeannette M. Haviland-Jones (eds) *Handbook of Emotions* (Guilford Press 2nd edition 2000); Gerhard Stemmler, 'Physiological processes during emotion' in Pierre Philippot and Robert S. (eds) *The Regulation of Emotion* (Lawrence Erlbaum Associates Publishers 2004); Heather C. Lench, Sarah A. Flores and Shane W. Bench, 'Discrete emotions predict changes in cognition, judgment, experience,

stimuli, the TCE holds that instances of emotions are *constructed* using concepts based on previous experiences.[23] Rather than being *distinct* 'fingerprints' found in *particular* brain regions, emotions are highly variable, with the entire brain involved in constructing a particular instance of emotion.[24] Importantly for this paper, the TCE suggests that, 'emotions are not, in principle, distinct from cognitions and perceptions'.[25]

A key element of the classical view is that emotions are universal. Research findings supporting this view began with a seminal study by Ekman and Friesen in New Guinea, in which they showed Fore tribe members pictures of facial expressions in the form of the six basic emotions and from the tribe members' interpretation thereof concluded they had found evidence for the universality of facial expressions of emotions.[26] A surge of similar research followed, with similar findings, thus building the impression of a sound evidence base for the universality of the basic emotion theory.[27] However, a new wave of psychology and neuroscience research casts considerable doubt on the validity of the previous body of research, with a recent meta-analysis concluding that the ways people express emotions, 'varies substantially across cultures, situations, and even across people within a single situation', and that similar facial expressions can convey multiple emotion categories.[28] Such findings have massive implications for face based EAI, and, I suggest, the classical view basis of EAI. The authors state that, 'It is not possible to confidently infer happiness from a smile, anger from a scowl, or sadness from a frown, as much of current technology tries to do'.[29] Despite a robust and convincing body of evidence challenging the assumptions and theoretical basis of EAI,[30] many companies offer emotion recognition services.[31] McStay identified twenty-one different sectors using EAI in 2018, including advertising, policing, social media, healthcare, workplace surveillance, and education.[32]

2.0 - Examples of Emotional Surveillance

This section draws on a number of examples of EAI to illustrate the multitude of ways in which this technology can be used for surveillance purposes. Space constraints preclude a detailed examination of every case, however the series of snapshots hopefully provide enough of a picture for the reader to comprehend the diverse surveillance capabilities of EAI. It must be noted that some examples are at very early stages of development, with some more speculative than examples of fully developed

behavior, and physiology: a meta-analysis of experimental emotion elicitations' (2011) 137(5) Psychological Bulletin < https://bit.ly/3sjRTzE > accessed 1 April 2021; Erika H. Siegel et al., 'Emotion Fingerprints or Emotion Populations? A Meta-Analytic Investigation of Autonomic Features of Emotion Categories' (2018) 144(4) Psychological Bulletin < https://bit.ly/3afJdDY > accessed 1 April 2021

[23] Feldman-Barrett, 2018 (n 16) 31

[24] Lisa Feldman-Barrett and Ajay Satpute, 'Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain' (2013) 23(3) Current Opinion in Neurobiology < https://bit.ly/3gplrtp > accessed 3 April 2021

[25] Feldman-Barrett, 2018 (n 16) 34

[26] Paul Ekman and Wallace V. Friesen, 'Constants across cultures in the face and emotion' (1971) 17(2) Journal of Personality and Social Psychology < https://bit.ly/3adrU6C > accessed 5 April 2021

[27] Hillary Anger Elfenbein and Nalini Ambady, 'On the universality and cultural specificity of emotion recognition: a meta-analysis' (2002) 128(2) Psychological Bulletin < https://bit.ly/3e5Jtqd > accessed 5 April 2021

[28] Lisa Feldman-Barrett et al., 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements' (2019) 20(1) Psychological Science in the Public Interest < https://bit.ly/3skZGNp > accessed 5 April 2021

[29] Ibid 46

[30] Ibid; See also James A. Russell, 'Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies' (1994) 115(1) Psychological Bulletin < https://bit.ly/3djWI7R > accessed 6 April 2021

[31] Andrew McStay, *Emotional AI: The Rise of Empathic Media* (SAGE Publications 2018) 67

[32] Ibid; See also, Andrew McStay, 'Report on the right to privacy in the age of emotional AI' for the office of the united nations high commissioner for human rights (2018) < https://bit.ly/3mOiVOk > accessed 14 March 2021

applications. The examples are expanded upon to facilitate discussion of *potential* similar uses and *possible* future uses. It is likely that some uses of EAI will be developed and some will not, but the point of this section is to use the examples to think through the possible implications of their development.

Another form of EAI is sentiment analysis of people's emotions and opinions as expressed in the form of written words, usually on social media.[33] During security operations for the 2012 London Olympics, UK security services monitored approximately 56,000 social media platforms, conducting sentiment analysis on their content to gauge levels of, and changes in, "social emotion".[34] The goal was to prevent terrorism and monitor protest groups, which David Omand, ex-director of GCHQ argues can help in deciding whether to use riot squads or other, less confrontational, methods.[35]

EAI is also being used in classrooms to monitor student's expressions for signs of focus, distractedness, and engagement with content.[36] A recent report found that in China, where many companies have introduced EAI to at least 600 classrooms, the technology is used to 'typologise' children based on their performance in class.[37] Data on students' views of EAI are rare, but the limited evidence suggests the use of EAI in classrooms causes feelings of fear and anxiety among students who worry about the future implications of their data being shared, for example, with potential universities.[38] In his analysis of social-emotional learning (SEL), Williamson identifies EAI as just one part of a, 'large-scale infrastructure for the definition and measurement of SEL' which is motivated by, 'the social, political and economic value to be derived from measurement and prediction of individuals' psychological characteristics, behavioural habits, and personality traits'.[39] If this policy trend of focussing on SEL continues, there is a reasonable chance that EAI will be used as one of the key ways of measuring outcomes. This raises serious concerns for McStay, who upon analysing the ethical and legal implications of EAI in classrooms, argues that such technology raises concerns about validity, value tensions, and violating the principle of data minimisation, concluding that, 'mining the emotional lives of children is normatively wrong'.[40]

In the city of Lucknow, India, there are proposals for EAI-enabled CCTV cameras with the stated goal of spotting distress on the faces of women being harassed. The system then alerts police officers on the ground who can intervene. While such goals are laudable, such a system could be repurposed for questionable ends, such as monitoring social gatherings for signs of anger in order to provide a justification for disrupting protests.

The use of EAI at border crossings has been tested as part of the 'iBorderCtrl' research project. One of the aims of the project was to test a system which would assess travellers' facial micro-expressions during an interview with an avatar in order to automatically detect signs of deception.[41] If travellers were flagged by the automated system, they would be questioned further by a human border security operative. Here, the use of deception detection has been applied to border crossings

[33] Bing Liu, *Sentiment Analysis: Mining Opinions, Sentiments and Emotions* (Cambridge University Press 2015) 1-15
[34] McStay, 2018 (n 31) 39
[35] Ibid
[36] Vidushi Marda and Shazeda Ahmed, Report from Article 19, 'Emotional Entanglement: China's emotion recognition market and its implications for human rights' (2021) < https://bit.ly/3tnKSip > accessed 25 March 2021 25-26
[37] Ibid 28
[38] Ibid 30
[39] Ben Williamson, 'Psychodata: disassembling the psychological, economic, and statistical infrastructure of 'social-emotional learning'' (2019) 36(1) Journal of Education Policy < https://bit.ly/3uSqRRv > accessed 9 April 2021 146-147
[40] Andrew McStay, 'Emotional AI and EdTech: serving the public good?' (2020) 45(3) Learning, Media and Technology < https://bit.ly/3mOjYOc > accessed 9 April 2021 281
[41] Keeley Crockett et al., 'Intelligent Deception Detection through Machine Based Interviewing' (2018) 2018 International Joint Conference on Neural Networks < https://bit.ly/32k1ID3 > accessed 19 March 2021

but could easily be similarly deployed in other situations where actors have an interest in detecting deception, such as during asylum claims or in court proceedings. For example, MediaRebel offers services to analyse witness deposition videos, though it is difficult to determine the extent to which this service has been deployed.

EAI is also used in the assessment of job candidates. Companies such as HireVue offer other companies services which typically involve candidates submitting responses to a series of questions in the form of a video. HireVue then analyse the candidate's video, looking at factors such as tone of voice and facial expression, use these and other factors to infer "underlying personality", predict future job performance, and assign each candidate an overall score.[42] Such tools are often used to screen out candidates. Perhaps tellingly, HireVue recently discontinued the facial analysis part of their hiring assessment tool. Here, EAI is used for assessing candidates' suitability for a job, but this could be repurposed, for example, to determine the suitability of loan applicants, benefits applicants, or university applicants. A similar use of EAI is in the surveillance of employees.[43] In this case, various attributes and behaviours of employees are monitored and analysed constantly. Such data may be used as a means of tracking and improving employee wellbeing, though could just as easily be used to inform performance reviews, thereby affecting promotions, demotions, and dismissals.

In São Paulo, EAI has been used in a digital interactive door system on the metro. The system displays advertisements like any other digital advertising space, but it also uses cameras to scan passengers faces, detecting their age, gender and emotions. The goal of the system is to understand passengers' emotional responses to the messages on display, allowing refinement of the messages. The system has come under scrutiny by civil society groups and is currently subject to a public civil action submitted by the Brazilian Institute of Consumer Protection, with the case ongoing at time of writing.[44] Here EAI has been used to tailor and refine an advertising message, but such a system could be used to tailor and refine political messaging. Furthermore, if combined with facial recognition systems, it may be possible to assess individual's responses to political messaging.

An area where EAI may be put to a range of uses is in personal healthcare and wellbeing. For example, Spire make a small, clip-on unit which measures activity levels, pulse rate, and respiratory effort. Spire claim their product allows wearers to track their "emotional and cognitive states", "focus", and whether they are in a period of "mental exertion".[45] The technology is primarily aimed at assisting doctors by allowing them to remotely monitor their patients, with patient data shared with a care team via a remote dashboard. In another case, Emotiv produce a range of EEG headsets to detect electrical signals from the brain which they say can "assess stress, focus and more" and offer solutions for improving "workplace wellness" and gaining "consumer insights".[46] Here, such personal devices as headsets and small wearables are intended for personal use only, but one could imagine such products being used in schools, where other forms of EAI are already being trialled.[47] One could also imagine their use being required in new insurance programs. For instance, just as

[42] HireVue Resource Library, Franziska Leutner and Clemens Aichholzer, 'Digital Video and Game Based Assessments: Psychometrics and Machine Learning' (2020) < https://bit.ly/2Q61ZXx > accessed 8 April 2021 14

[43] Sharon Richardson, 'Affective computing in the modern workplace' (2020) 37(2) Business Information Review < https://bit.ly/2Q8gSsm > accessed 1 March 2021

[44] Access Now, 'Expert Opinion on Facial Categorization in Brazil' (2020) < https://bit.ly/32dhLT8 > accessed 14 March 2021

[45] SpireHealth, Webpage < https://bit.ly/2PXrMS0 > accessed 15 March 2021

[46] Emotiv, Webpage < https://bit.ly/32fnKqi > accessed 15 March 2021

[47] McStay, 2020 (n 40)

drivers are offered lower insurance rates for installing a black box to monitor their driving,[48] so people could be offered lower health insurance rates for using a range of wearables to track emotional states and "mental exertion".

Social media companies are another group of actors who are increasingly interested in using EAI to infer the emotions of their users. For example, Facebook data has been used to predict the onset of post-partum depression.[49] Another notorious example from Facebook is the 'emotional contagion' experiment where people's news feeds were manipulated to investigate if such changes had an impact on their emotions. While equating number of "positive" and "negative" posts with emotion is problematic to say the least, this example at least demonstrates an interest in *manipulating* users' emotions. There is also interest in using social media to predict mental health status,[50] with specific examples including identifying signs of depression from Instagram pictures,[51] and trying to prevent suicide through natural language processing.[52]

So called "smart cities" are another area where EAI may be used. One example is Dubai, where there is explicit interest in using such data to improve the happiness of residents and tourists. Using smart city sensors for monitoring and improving happiness may be a beneficent endeavour, but smart city sensors could easily be used to track other emotions, such as anger or sadness. Scholars have noted the variety of data which can be 'biosensed' or collected *remotely* by different sensors, including body temperature, eye movement, heart rate, facial, and emotional recognition.[53] One can imagine a complex sentiment analysis system in which people's bodies are monitored in real time as they move throughout a city.

3.0 - The *Causes*, *Courses* and *Consequences* of Emotional Surveillance

While the foregoing examples provide an insight into the various ways EAI is being deployed, it is useful to organise the key characteristics of these disparate instances in order to draw some general lessons from them. To do so, Lyon's concepts of surveillance *causes*, *courses*, and *consequences* will be used. *Causes* refers to 'what drives surveillance', understood here as the objectives those conducting emotional surveillance are trying to achieve. C*ourses* refers to the 'main ways in which surveillance operates'; and *consequences* to the 'effects of surveillance on individuals, groups and the overall structuring of social relationships'.[54] Surveillance is here understood as, 'the focused, systematic and routine attention to personal details for purposes of influence, management, protection or direction'.[55] Surveillance has been chosen as the lens with which to view EAI because

[48] Andrew McStay and Duncan Minty, Report for the Centre of Data Ethics and Innovation, 'Emotional AI and Insurance: Online Targeting and Bias in Algorithmic Decision Making' (2019) < https://bit.ly/2OPpOCq > accessed 15 March 2021

[49] Munmun De Choudhury et al., 'Characterizing and predicting postpartum depression from shared Facebook data' (2014) Proceedings of the 17th ACM conference on computer supported cooperative work & social computing < https://bit.ly/32dllN4 > accessed 16 March 2021

[50] Stevie Chancellor and Munmun D. Choudhury, 'Methods in predictive techniques for mental health status on social media: a critical review' (2020) 3(1) NPJ Digital Medicine < https://go.nature.com/3e9gxOp > accessed 16 March 2021

[51] Andrew G. Reece and Christopher M. Danforth, 'Instagram photos reveal predictive markers of depression' (2017) 6(1) EPJ data science < https://bit.ly/3uSY3s1 > accessed 16 March 2021

[52] Mukhtarkhanuly Daniyar and Alan Abishev, 'Suicidal Post Detection in Social Networks using NLP' (2018) 7(3) Advanced Engineering Technology and Application < https://bit.ly/3uTqEO0 > accessed 16 March 2021

[53] Elaine Sedenberg, Richmond Wong and John Chuang, 'A window into the soul: Biosensing in public' (2017) < https://bit.ly/3e6mj3e > accessed 20 March 2021

[54] David Lyon, *Surveillance Studies: An Overview* (Polity Press 2007) 47

[55] Ibid 14

of the degree to which it fits the above description, and, relatedly, because the surveillance studies literature offers a range of useful theories with which to critique emotional surveillance.

3.1 – The *Causes* of Emotional Surveillance

Those who conduct emotional surveillance are trying to achieve a wide range of objectives, which can be understood as the *causes* of why they are interested in such surveillance. In this section I apply Lyon's concept of surveillance causes by trying to identify and then categorise the various motivating factors for emotional surveillance. However, caveat is that particular examples often fall into multiple overlapping categories. Firstly, there are various *security* objectives, such as identifying ongoing assaults and directing police officers on the ground, making operational decisions like deploying riot squads, preventing terrorism, monitoring protest groups, and attempting to detect deception at international borders. Secondly, there are numerous objectives relating to the *workplace*, such as assessing and screening out candidates and monitoring employee wellness and performance. The education setting may fit here as it is obviously a place of work for teachers but could also be considered as a *kind of* place of work for children and young adults, with the objective of assessing performance being similar to adult workplaces. Thirdly, numerous *market* objectives exist, including tailoring and refining advertising messaging and gaining consumer insights. Fourthly, *health and wellbeing* objectives include identifying signs of depression, preventing suicide, and self-tracking mental states, and remotely monitoring patients. Fifthly, social media companies have diverse objectives, ranging from inferring emotional states in order to deliver tailored content, to monetising users' emotional responses. Finally, in the smart city context, *social* objectives include improving services and increasing levels of happiness. There are clearly a very wide range of reasons why various actors are interested in surveilling emotional states. While these may differ with regard to the extent they either benefit or harm surveillance subjects, the relationship between those surveilling and those surveilled is usually a complicated one, with emotional surveillance resulting in multiple harms *and* benefits (the degree to which these are truly benefits is discussed later).

3.2 – The *Courses* of Emotional Surveillance

To understand how emotional surveillance works (*courses*) I will draw out two themes from the above examples: what the EAI technology, artifact, or application that facilitates surveillance *is*, and what *feature* of the surveillance subject is collected or analysed. Technology is here understood as a group of similar techniques that share a common purpose and/or feature.[56]

Emotional surveillance can be facilitated by CCTV cameras, social media platforms, avatar-enabled automatic interviewing systems, job application videos, cameras positioned in digital advertising spaces, small wearables, EEG headsets, and natural language processing. From this array we can see that a complex assemblage of highly diverse technologies, artifacts, and applications facilitate emotional surveillance. Furthermore, such surveillance can be facilitated either by integrating EAI capabilities into older technologies (e.g., integrating cameras, computer vision, and ML into digital advertising spaces) or by creating new, purpose-built artifacts for detecting emotion (e.g., wearables and EEG headsets). As EAI can be embedded as an additional feature of other applications, its ability

[56] Phillip A. E. Brey, 'Anticipating Ethical Issues in Emerging IT' (2012) 14(4) Ethics and Information Technology < https://bit.ly/2Qpl6vN > accessed 28 March 2021 310

to scale is profound, increasing the speed with which it can proliferate without due consideration of its social acceptability or the requisite safeguards and laws to regulate its use.[57]

The above *means* of surveilling emotions collect and analyse a multitude of *features* of the surveillance subject, including: social media content (written text and pictures), facial expressions (including micro-expressions), tone of voice, heart rate, respiratory effort, electrical signals from the brain, body temperature, and eye movement. The features that are intentionally captured for the purposes of emotional surveillance are thus also characterised by their high degree of diversity. In particular, there is a wide range of *bodily* features which are collected, in line with McStay and Urquhart's prediction of a shift to appraisal based EAI that integrates assessment of physiological contexts.[58] That being said, features do not *have* to be bodily, with many features external to the body, such as written social media posts, also used for emotional surveillance purposes.

Emotional surveillance thus describes a vast range of new or repurposed technological artifacts and analytical techniques that are being used to monitor and assess an increasingly broad range of corporeal and non-corporeal features of people in order to conduct surveillance of their emotional states. A useful concept for understanding EAI and anticipating its possible and probable uses is function creep.[59] Function creep can occur in two ways with regard to EAI. Firstly, EAI can cause other technologies' or artifacts' function to change due to the incorporation of EAI capabilities. This is the case with, for example, CCTV systems, digital advertisement boards, and social media platforms. Secondly, the function of EAI artifacts created purposefully for emotional surveillance purpose can change. This is the case, for example, when wearables for self-tracking are then used to determine insurance rates, or when facial analysis software is then used in hiring decisions and employee performance reviews. Policymakers, regulators, and those developing EAI must be keenly aware of the highly adaptable nature of EAI which renders it prone to unintentional and intentional misuse.

3.3 – The *Consequences* of Emotional Surveillance

What, then, are the consequences of emotional surveillance for individuals, groups, and society? While the vast range of uses of emotional surveillance clearly warrants further discussion in many directions about their consequences, covering all of these is clearly out with the scope of this paper. The main focus of this paper is on the potential impact on subjects' human right to freedom of thought. Nonetheless, before getting to that, the present section highlights some general observations about emotional surveillance that draw upon various theoretical perspectives from surveillance studies in order to interpret the foregoing examples of emotional surveillance.

3.3.1 – Direct, Indirect, and Bi-directional Consequences of Emotional Surveillance

The first category of consequences of emotional surveillance is the *direct* consequences which are inherently and immediately related to the particular use case. Such direct consequences include not getting a job interview, being denied access to a country, one's social media content being the object of surveillance analysis, emotional surveillance at work affecting performance reviews, and

[57] McStay and Urquhart, 2019 (n 14) 4

[58] Ibid 10

[59] Bert-Jaap Koops, 'The Concept of Function Creep' (2021) 13(1) Law, Innovation and Technology < https://bit.ly/2Q3IsqR > accessed 9 April 2021; See also, Langdon Winner, *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought* (MIT Press 1077) 57

emotional data being sold by, for example, social media or wearables companies for the purposes of targeting individuals with particular content. Such direct consequences raise two problems. Firstly, data about emotions is being used as a new way to treat people differently, raising concerns of discrimination occurring. Secondly, and relatedly, important decisions affecting access to various services and opportunities such as paid work are being made either exclusively or partially on the basis of algorithmic analysis that is subject to serious critique and which may be fundamentally flawed.[60]

The second category of consequences of emotional surveillance are the *indirect* consequences that flow from the *direct* consequences, often in complex ways. Here it is worth highlighting a point made by McGregor et al. about the use of algorithms in a range of decision-making contexts potentially affecting various social rights.[61] While their analysis is necessarily broad, I suggest that EAI is a specific example of a technology that can be applied in different decision-making contexts as a means of blocking or granting access to goods that are considered economic, social, and cultural rights. Of particular salience are the rights to equal treatment of men and women, to work, to just and favourable conditions of work, especially equal opportunity of promotion, to social security, to mental health, and to education.[62] To be clear, this is not a new phenomenon; rather, EAI is yet another tool to add to a growing set of algorithmic decision-making systems that are placed firmly in-between people and the services and opportunities they require.

Another observation of emotional surveillance relates to its bi-directional nature. Staples states that, 'disciplinary power expands "bi-directionally", flowing from top to bottom and vice versa.[63] Elaborating on the examples he provides, this bi-directional disciplinary power is evident in the case of teachers, police officers, and hiring managers who use EAI to surveil and assess students, citizens, and job applicants respectively, while their employers can then use the same technology to surveil and evaluate *their* performance. Being mindful of these bi-directional power dynamics helps guard against the simplistic assumption that the individuals involved in the practical operationalisation of emotional surveillance are *always* in positions of *complete* power. Rather, the way EAI mediates relationships of power within society is complex and requires nuanced consideration of each person's particular circumstances.

3.3.2 – Body-objectifying Consequences of Emotional Surveillance

Through a discussion of drug testing, lie detectors, digital identification and a specific form of EAI, namely deception detection, Staples analyses modern surveillance by focussing on the surveillance subject's *body.* For instance, he observes that a primary objective of surveillance is to, 'circumvent the speaking subject', which is achieved by, 'deriving knowledge and evidence from the body'.[64] While Staples identifies the consequences of this approach for those conducting surveillance, namely reducing the need to rely on the subject telling the truth while simultaneously exercising 'disciplinary power' over the subject, more could be said of the consequences for the *subject* of surveillance. Here I suggest two such consequences in relation to EAI. Firstly, the surveillance subjects' *body* being treated as the source of truth, rather than their mind, rather than what they

[60] See earlier discussion on the academic debate on this issue, esp. Feldman-Barrett, 2019 (n 28)

[61] McGregor, Murray and Ng, 2019 (n 12) 310

[62] UN General Assembly, International Covenant on Economic, Social and Cultural Rights, 16 December 1966, United Nations, Treaty Series, vol. 993, 3 < https://bit.ly/3e8c0LM > accessed 27 February 2021 arts 3, 6(1), 7(c), 9, 12(1), and 13

[63] William G. Staples, *Everyday Surveillance: Vigilance and Visibility in Postmodern Life* (Rowman & Littlefield Publishers 2013) 202

[64] Ibid 118-119

say is the truth, diminishes their autonomy to determine their own identity by undermining and devaluing the importance and authority of their own interpretation of themselves and of events. EAI systems provide many situations where people's emotional "score" may not align with how they think of themselves. If emotional surveillance is used to make decisions about people, and if those decisions are informed by an analysis which contradicts people's understanding of their emotional selves, the effect is one of diminishing people's autonomy to shape their own identity. Illouz argues that, 'one's emotional attitudes and style, like one's cultural taste, define one's social identity'.[65] If this is true, then assigning emotion "scores", especially if they are accepted by others as true, has the potential to limit people's capacity to determine their own identity because people's identity becomes increasingly determined by what the EAI system says it is, rather than what the individual says it is. I would caution that Illouz's statement is perhaps too strongly worded, with emotion being just one of many aspects that comprise our identities, albeit an important one.

Secondly, emotional surveillance impinges upon human dignity, understood here as the *principle* which underpins and justifies human rights, rather than a right to dignity, in two related ways. Here I adopt O'Mahoney's framing of dignity as having both a descriptive (humans have dignity *because* of their very humanity) and a normative (because of their dignity, humans, 'should be afforded human rights on the basis of *equal treatment* and *respect*' (emphasis added)) aspect.[66] Emotional surveillance impinges upon human dignity, firstly, because it subjects people to an unvalidated process. By doing so, those deploying emotional surveillance are not treating people with the *respect* they deserve by virtue of their dignity. Such treatment is disrespectful because to treat someone using an algorithmic system that is not only unvalidated but subject to robust critique is both misleading and liable to producing inaccurate, incorrect results, with potentially harmful consequences for the surveillance subject. Secondly, emotional surveillance impinges upon human dignity because it is used to treat people as means rather than ends by extracting data from them which is used for self-interested purposes. People may argue against this second point by referring to the benefits emotional surveillance provides for either individuals, such as dealing with mental health problems, seeing more relevant content, or paying lower insurance rates, or for society, such as crime prevention, security, efficiency, or improved smart city services. However, I suggest the very framing of these outcomes as benefits is misguided because EAI is an unvalidated technology based on dubious assumptions, rendering the supposed benefits either empty, false, or outright dangerous.

Another useful way Staples frames various surveillance techniques is in their attempt to, 'evoke the legitimacy of science and technical objectivity', which I suggest is an inherent feature of EAI and one which makes it misleading, as discussed above. In terms of what this means for surveillance subjects, a key consequence of EAI gaining such legitimacy is that the grounds for contesting the decisions based upon such practices are limited. This may occur if enough relevant actors simply *perceive* EAI to be accurate and legitimate, regardless of whether EAI ever attains genuine scientific validation. By limiting the possible grounds for disputing EAI-enabled decisions, the likelihood of the subject having to accept such decisions is increased.

Staples also argues that surveillance practices can impose a 'disciplinary ritual' on subjects, the force of which may be enhanced, 'by creating the illusion that the truth can, in fact, be had'.[67] Three things may be said of EAI in light of such observations. The first regards the ritualistic aspect. The ability of

[65] Eva Illouz, *Cold Intimacies: The Making of Emotional Capitalism* (Polity Press 2007) 66
[66] Conor O'Mahony, 'There is no such thing as a right to dignity' (2012) 10(2) International Journal of Constitutional Law < https://bit.ly/32eBlyn > accessed 1 April 2021
[67] Staples, 2013 (n 63) 134

EAI to scale and spread rapidly throughout society, permeating many sectors, is evident from its ability to integrate with other technologies and systems. Such potential ubiquity may normalise people to the process of having their emotional states surveilled and used for decision-making purposes. Thus normalised, people are less likely to question the legitimacy of such practices, despite the scientific basis and assumptions of EAI systems being highly questionable. The second point regards discipline. Emotional surveillance can be viewed as a means of imposing a specific form of discipline, namely disciplining the emotional state of surveillance subjects, by dictating rewards or punishments for desirable or unwanted emotional reactions to specific stimuli. The third point regards, 'the illusion' Staples highlights above. EAI systems are a prime example of a technology which can create that which it seeks to find.[68] For those unacquainted with the theoretical assumptions and technical processes that underpin EAI systems, the results such systems produce are unlikely to be disputed. Rather, the manner in which such systems are presented and treated as objective 'truth machines' is likely to engender an attitude of acceptance and deference to the results of EAI systems.

Finally, Staples characterises modern surveillance practices' emphasis on deriving knowledge from subjects' bodies as the, 'pornography of the self'.[69] Although this metaphor may usefully highlight the exposing nature of emotional surveillance, I would argue that it is of limited use when discussing emotional surveillance because EAI systems are often, though certainly not always, *hidden* from the surveillance subjects' view. Staples' somewhat crude metaphor implies a certain level of awareness of the surveillance practice occurring, however such awareness is not always present during emotional surveillance. Perhaps a better metaphor for emotional surveillance is that of the "peeping Tom", where those doing the "peeping" have an extensive and elaborate set of technologies for such purposes.

3.3.3 – Predictive Consequences of Emotional Surveillance

A number of scholars have identified various characteristics of modern surveillance which are useful ways of considering the impact of emotional surveillance. Twenty years ago, Lyon observed that surveillance was focussing on the body, 'as a *source of data for prediction*' (emphasis added).[70] Similarly, van der Ploeg discusses how surveillance treats the body *as* information.[71] Lyon also proposes the influential idea that surveillance functions as a form of social sorting. These authors were writing before the recent rapid developments in AI, yet their ideas can combine to elucidate the consequences of emotional surveillance for subjects.

Emotional surveillance as a process can be viewed as a manifestation of these three theoretical insights. Taking the concept of the body *as* information first, with the exception of sentiment analysis, an important feature of emotional surveillance is treating the body – faces, voices, heart rates – *as data*. A combination of old and new techniques allows these bodily features to be captured and used for various purposes, raising issues of consent if data is collected without the subject's knowledge.

Secondly, a fundamental purpose of emotional data is using it to make predictions about people, of which there are various categories. The first category is predicting *future emotional states*, as in the

[68] Staples, 2013 (n 63) 135

[69] Ibid 119

[70] David Lyon, *Surveillance Society: Monitoring everyday life* (Open University Press 2001) 72

[71] Irma van der Ploeg, 'Biometrics and the body as information: normative issues of the socio-technical coding of the body' in David Lyon (eds) *Surveillance as Social Sorting* (Routledge 2002)

case of using emotional data to predict future mental health problems. The second category is predicting *future actions*. This is perhaps most clear in the security and policing context when actors use emotional surveillance to predict when those in a crowd may turn violent. However, it is also evident in the economic context when predicting the likelihood of someone buying a product, and the political context when predicting voting intentions. One can see how these first two categories blur in certain cases, for example trying to prevent suicides. The third category is predicting *future performance*, such as when emotional data is used to predict the performance of job candidates.

Finally, the main consequence for emotional surveillance subjects is being sorted into a multitude of categories on the basis of emotional data. One could list many such categories but drawing on the above examples is enough to illustrate the various ways that people may be categorised as high or low risk of mental health problems, into different insurance price categories, as high or low risk of being violent, as more or less likely to buy a product or vote for a candidate, or suitable or not for a job. Crampton offers a way of interpreting the nature of this sorting process. He argues that a key difference between facial recognition and emotion recognition is that whereas with facial recognition the target of surveillance is *seen*, whereas with emotion recognition the target of surveillance is *seen as*.[72] Applying this interpretation to emotional surveillance, we see that despite convincing evidence and arguments that similar facial expressions express many emotions, those subjected to face based emotional surveillance will be *seen as* happy, sad, angry, at risk, capable, dangerous etc. and thereby be *seen as* or categorised in a way which determines further action and a range of possible consequences for those so surveilled.[73] For example, someone deemed 'sad' could be targeted with certain online content which exacerbates their sadness and creates a vicious cycle, with potentially harmful consequences. Limited empirical evidence suggests that people are worried about EAI on social media negatively impacting their mental health by exacerbating existing issues.[74] Of course, such an accuracy-of-the-technology type critique based on Feldman-Barrett et al.'s work would only apply to emotional surveillance based on facial expressions, when there are many other forms of emotional surveillance. However, I suggest that Crampton's point could be applied to all forms of emotional surveillance, no matter the source of data, because the fundamental process remains taking data from *somewhere* and using it to attempt to determine the emotional state of the subject, with the goal of *seeing them as*, rather than simply *seeing them*.

While useful, some of the limitations of using this lens of predictive consequences to evaluate emotional surveillance are highlighted. Firstly, with regard to Lyon and van der Ploeg's focus on surveillance treating the body as a source of data, this does not hold up in all cases of emotional surveillance, such as when textual data on social media is analysed. Secondly, EAI is not always used for prediction, it can be used for live analysis, such as in advertising boards and in security cameras. Finally, EAI is not always used for the purposes of social sorting, such as when used in gaming. However, simply looking to where EAI is used reveals little about whether it is used for social sorting or not. EAI in schools and workplaces *could* be used to sort people into groups, determining their progress, but it *could also* be used purely for altruistic purposes in efforts to enhance their wellbeing. The key factor is how various actors use the technology.

This section closes by summarising the lessons learned from applying some analytical perspectives from surveillance studies to EAI. Firstly, concerning causes, a diverse set of actors are motivated to

[72] Jeremy W. Crampton, 'Platform biometrics' (2019) 17(1/2) Surveillance and Society < https://bit.ly/3mW9QmD > accessed 18 March 2021 60

[73] Feldman-Barrett, 2019 (n 28)

[74] Nazanin Andalibi and Justin Buss, 'The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks' (2020) Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems < https://bit.ly/3srnBLq > accessed 9 March 2021 8

conduct emotional surveillance by many widely varying objectives. Secondly, emotional surveillance is facilitated by a large number of different technologies and applications, which can either be upgrades of old technologies, or new, purpose-built applications. As it can so easily integrate into other technological systems, EAI has the potential to scale up rapidly. Regarding the features of people captured during emotional surveillance, these too are highly diverse, with many, though not all, being corporeal. The diversity and scalability of EAI means it has particularly high potential for function creep, both by altering other technologies and by purpose-built EAI applications being repurposed.

Third, there are a wide range of consequences of emotional surveillance. There are direct consequences which relate to the immediate results of the way emotional surveillance is used, as well as the indirect consequences which are more complex, relating to, for example, the impact emotional surveillance has on people's social rights. The consequences of emotional surveillance are bi-directional, in that it can exert disciplinary power both on surveillance subjects *and* on those conducting the surveillance. In terms of body-objectifying consequences, I have suggested that emotional surveillance may undermine people's autonomy to determine their own identity, as well as impinge upon human dignity by subjecting them to unvalidated technological processes and treating them as means rather than ends. Emotional surveillance can also be viewed as a means of exerting power by disciplining the emotional state of those subjected to it. The potential ubiquity of and lack of knowledge about EAI may lead to emotional surveillance being normalised and unquestioningly accepted, respectively. In terms of predictive consequences, four points are salient. Firstly, EAI often, but not always, treats the body *as* data, raising concerns around consent. Secondly, a key objective of emotional surveillance is to make predictions about people, about their future emotional state, future actions, and future performance. Thirdly, emotional surveillance is used to sort people into a multitude of categories based solely, or in part, on data about their emotions, with an equally diverse range of consequences flowing from such categorisation. Finally, emotional surveillance classifies people by 'seeing them *as'* something, rather than just 'seeing them'.

As attention now turns to the impact of emotional surveillance on the right to freedom of thought, some remarks are in order about how the particular nature forms of EAI may have implications for the freedom of thought. For example, the type of EAI deployed may influence the degree to which the surveillance is deemed invasive and an impingement of one's right to freedom of thought. To illustrate, measuring someone's heart rate, gaze, and galvanic skin response while driving or playing a computer game may be less invasive than measuring someone's facial micro-expressions and vocal tones while they work. A further point concerns the close interrelationship between emotional surveillance courses and consequences. In essence, the exact *form* emotional surveillance takes, as well as *where* it occurs, has implications for the effects it can have on people. Such factors also dictate the degree to which people can know about, resist, and/or contest emotional surveillance.

4.0 - Freedom of Thought and Emotional Surveillance

The right to freedom of thought is found in all international human rights law treaties. Article 18(1) of the international covenant on civil and political rights (ICCPR) states:

> Everyone shall have the right to freedom of thought, conscience and religion. This right shall include freedom to have or to adopt a religion or belief of his choice, and freedom, either individually or in community with others and in public or private, to manifest his religion or belief in worship, observance, practice and teaching.

Article 9(1) of the European Convention on Human Rights (ECHR) contains a nearly identical provision. Bublitz highlights that the right to freedom of thought comprises two distinct elements: an internal element, or '*forum internum*', referring to the thoughts inside one's head, and a *forum externum*, referring to the external manifestations of those internal thoughts. This distinction is emphasised by virtue of the fact that on the one hand internal thoughts are given absolute protection.[75] The absolute and non-derogable nature of the *forum internum* aspect of freedom of thought has been clarified by the Human Rights Committee (HRC) and the Council of Europe (CoE).[76] On the other hand, external manifestations of those thoughts are subject to certain limitations. To illustrate, article 18(3) of the ICCPR states:

> Freedom *to manifest* one's religion or beliefs may be subject only to such limitations as are prescribed by law and are necessary to protect public safety, order, health, or morals or the fundamental rights and freedoms of others. (emphasis added)

Similarly, article 9(2) of the ECHR contains an almost identical provision describing the permissible limitations on the *manifestations* of one's thoughts. At first glance, the right to freedom of thought provides an absolute right to *have* or to *change* one's thoughts (internal), and a right to *manifest* those thoughts (external), which is subject to certain limitations. However, this summary raises more questions than it answers, prompting closer inspection of the relevant case law and soft law to further elucidate elements of the right to freedom of thought.

Unfortunately, there is very little case law on the right to freedom of thought, which is matched by a paucity of literature on the issue.[77] I concur with Alegre that the most likely explanation for the lack of case law is the widely held and understandable assumption that internal thoughts are inaccessible and therefore inviolable by default.[78] Nonetheless, the limited case law and commentary do distinguish some more detailed features of the right.

Firstly, regarding the scope of the right, the HRC state that, 'it encompasses freedom of thoughts on all matters'.[79] In one of very few cases where freedom of thought specifically was at issue, the right was interpreted broadly on the basis of, 'the comprehensiveness of the concept of thought'.[80] In a separate case, the European court of human rights (ECtHR) confirmed the right covered both religious and non-religious thoughts.[81] However, in more recent case law, the ECtHR appears to have narrowed its approach, stating that for thoughts to be protected under article 9 they must, 'attain a certain level of cogency, seriousness, cohesion and importance'.[82] Thus there is a divergence between the ECtHR and the HRC in how the right should be interpreted with regard to what thoughts fall under its protection. This distinction between levels of thought that do and do not deserve protection clearly suggests a category of thoughts which are *not* cogent, serious, cohesive, and important. One might argue that emotions fall into this category. However, Nussbaum argues that emotions are, 'suffused with intelligence and discernment', 'contain in themselves an

[75] Article 4(2) of the ICCPR states that there can be no derogations from Article 18.

[76] UN Human Rights Committee (HRC), *CCPR General Comment No. 22: Article 18 (Freedom of Thought, Conscience or Religion)*, 30 July 1993, CCPR/C/21/Rev.1/Add.4, < https://bit.ly/3dgBqrv > 17 March 2021 1; Council of Europe, Guide on Article 9 of the European Convention on Human Rights Freedom of thought, conscience and religion, Updated on 31 August 2020 < https://bit.ly/3mOGDKa > accessed 17 March 2021 11

[77] Alegre, 2017 (n 10) 1; Simon McCarthy-Jones, 'The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century' (2019) 2 < https://bit.ly/3e3E5Ej > accessed 18 February 2021 5

[78] Alegre, 2017 (n 10) 1

[79] General Comment no. 22, 1993 (n 75)

[80] *Salonen v. Finland* App no 27868/95 (ECtHR 2 July 1997)

[81] *Kokkinakis v. Greece* App no 14307/88 (ECtHR 25 May 1993) 31

[82] *Izzettin Do˘gan and others v. Turkey* App no 62649/10 (ECtHR 26 April 2016) 68; *S.A.S. v. France* App no 43835/11 (ECtHR 1 July 2014) 55

awareness of value or importance', and are, 'part and parcel of the system of ethical reasoning'.[83] Nussbaum's position therefore treats emotions as an important part of the thinking process, thereby challenging the position that would place emotions out with the protection of article 9.

Much of the discussion of the right to freedom of thought in case law, soft law, and academia focusses on the nature of the thought in question.[84] Is the thought religious or non-religious, trivial or serious, philosophical or farcical? The requirement of the ECtHR that thoughts reach a certain level of, 'cogency, seriousness, cohesion, and importance' typifies this approach of focussing on what the thought is *about*. Additionally, courts have only dealt with questions arising from *manifestations* of thought, not thoughts themselves.

However, this focus on *what* thoughts are protected, which furthermore have only been protected in practice if they are manifested, misses another important consideration, namely protecting the *conditions* in which thinking may occur freely. To this end, commentary on the right to freedom of thought distinguishes three key elements that, taken together, start to flesh out the details of the right to freedom of thought in the *forum internum*. Vermeulen's commentary identifies the right as comprising:

- The right not to reveal one's thoughts,
- The right not to have one's thoughts manipulated, and
- The right not to be penalised for one's thoughts.

Similarly, another right that concerns the *forum internum* is the right to hold opinions. Following in-depth analysis of treaty texts and case law, Aswad concludes that the right to hold opinions comprises three identical key elements.[85] If these three elements form the core of the right to freedom of thought in the *forum internum*, a latent area of enquiry, as Alegre points out, is determining, 'what interference with those absolute rights might look like in practice'.[86] The following three sections therefore discuss these three elements of the right to freedom of thought in light of emotional surveillance. The general argument is that emotional surveillance may amount to violations of the right to freedom of thought.

4.1 – The Right Not to Reveal One's Thoughts and Emotional Surveillance

Vermeulen's analysis suggests a right not to reveal one's thoughts, whereas the HRC state that, 'no one can be *compelled* to reveal his thoughts' (emphasis added).[87] This introduces an additional factor, that of being *compelled* to reveal one's thoughts, which would limit the circumstances in which the right may apply. A would-be complainant would have to show not only that their thoughts were revealed, but also that they were compelled to reveal them. The right not to reveal one's thoughts may be violated in the case of analysing emotional responses to public advertising boards, when social media platforms analyse their users' emotional state and responses to particular content, when smart city sensors evaluate people's emotions and, similarly, when CCTV systems

[83] Martha C. Nussbaum, *Upheavals of thought: The Intelligence of Emotions* (Cambridge University Press 2001) 1
[84] Sjors Ligthart, 'Freedom of thought in Europe: do advances in 'brain-reading' technology call for revision?' (2020) Journal of Law and the Biosciences < https://bit.ly/32gz3ib > accessed 17 March 2021; Jan Christoph Bublitz, 'Freedom of thought in the age of neuroscience: A Plea and a Proposal for the Renaissance of a Forgotten Fundamental Right' (2014) 100(1) Archives for Philosophy of Law and Social Philosophy < https://bit.ly/3wXJBRq > accessed 17 March 2021
[85] Evelyn Mary Aswad, 'Losing the freedom to be human' (2020) 52(1) Columbia Human Rights Law Review < https://bit.ly/3wZF68Q > accessed 1 April 2021 359-364
[86] Alegre, 2017 (n 10) 4
[87] General Comment no. 22, 1993 (n 75) 3

purposefully attempt to surveil emotions. The right not to be *compelled* to reveal one's thoughts may be violated in the iBorderCtrl, job application, employee monitoring, and classroom examples. Such instances involve situations in which either being able to access something (e.g., another country, paid employment) or continuing to work or learn in a particular environment is being made conditional upon accepting emotional surveillance, hence they are being compelled. There may be other situations where while it may not be the case that one is being explicitly and unequivocally *compelled* to reveal one's thoughts, there are strong motivating factors to do so, such as when insurance rates may be lower if one owns a wearable that tracks mood.

Such examples of emotional surveillance demonstrate potential violations of the right not to reveal one's thoughts. As to the question of whether Vermeulen's broader (right not to reveal) or the HRC's more limited (right not to be *compelled* to reveal) formulation of the right ought to be the right one, I suggest the broader interpretation is the correct one for two reasons. Firstly, it allows for a wider range of possible violations to be considered under the right to freedom of thought. Secondly, the narrower interpretation would likely render the right much weaker in practice as actors could find numerous ways to argue that their particular use of emotional surveillance did not *compel* people to reveal their thoughts.

4.2 – The Right Not to have One's Thoughts Manipulated and Emotional Surveillance

Concern over the capacity for algorithms to, 'influence emotions and thoughts' and manipulate economic, social, and political choices have been raised recently by the CoE.[88] While they pay particular attention to the impact on the right to form opinions, the parallels with freedom of thought are clear as both rights relate to the *forum internum*, as previously discussed. Various examples of emotional surveillance may violate the right not to have one's thoughts manipulated, such as when emotional responses to content on social media is used to target people with content with the goal of either reinforcing or changing particular views about that content. For example, supposing someone is deemed to react angrily to a particular political message, such information may be used to target the individual with similar or more extreme messages with the aim of radicalising their views. In another case, employees may be manipulated to conform to a particular institutional viewpoint. For example, those in a position of power within an organisation could create an environment where employees are acutely aware of emotional surveillance, which causes them to suppress or alter their feelings about operational or other decisions, thus manipulating thoughts to subdue dissent.

4.3 – The Right Not to be Penalised for One's Thoughts and Emotional Surveillance

What it means to be 'penalised' for one's thoughts is unclear; however, it seems reasonable to assume that such penalties would exist on a spectrum, ranging from trivial inconveniences to serious impacts. The more trivial end of the spectrum may concern things like not seeing a particular advert or being deemed 'sad' by smart city sensors. Moving along the spectrum, more serious penalties may include things like paying higher insurance rates or being targeted for closer surveillance. Penalties at the serious end of the spectrum may include things like being denied freedom to travel to another country, being denied a job, being demoted or fired, or being denied access to key

[88] Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes (2019) < https://bit.ly/2Q6ee6r > accessed 3 April 2021 8

services and opportunities such as benefits, loans, or university places, all on the basis of emotional surveillance. The CoE recently raised concerns about the risks to social rights from the use of automated decision-making systems, and EAI, if used for example during assessments of suitability for social services, has the potential to cause such risks.[89] Where the impact of a particular use of emotional surveillance falls on this spectrum of severity also depends on individual circumstances. As an illustrative example, if person A's use of wearables to track mood results in them having to pay more for health insurance, but they can still afford it comfortably, then the effect is certainly not as serious as for person B, who's similar use of wearables results in them being unable to continue to afford health insurance. In summary, I suggest that one of the key features of emotional surveillance is creating situations in which one may be penalised, to a more or less serious degree, for one's thoughts.

4.4 – Critique of the Right to Freedom of Thought in IHRL

To reiterate, McGregor et al. argue that IHRL offers a framework for defining harms of AI systems and dictating the obligations and expectations of States and businesses, respectively. However, does this argument hold when assessing EAI systems impact on the right to freedom of thought? It must be noted that the authors argue that IHRL offers a, '*holistic* approach to accountability' (emphasis added), meaning my critique of the right to freedom of thought is not intended to, indeed could not, undermine the value of the overall approach.[90] It is nonetheless useful to consider the ways in which the particular part of IHRL that is the right to freedom of thought may not achieve the goals set for it by McGregor et al.

Regarding the objective of defining harms, I suggest that IHRL is only moderately successful. The foregoing discussion has been an attempt to sketch out the possible violations of the right to freedom of thought that emotional surveillance may cause. However, in doing so, I have had to rely heavily on expert commentaries on, and academic analysis of, the right to freedom of thought. There is a distinct lack of case law and soft law guidance on the right to freedom of thought specifically, and that which does exist sheds very little light on how the right should be interpreted in the face of emerging technologies such as EAI.

The very structure of the right to freedom of thought is another issue. Vermeulen states that, 'It is true that thoughts and views, as long as they have not been expressed, are *intangible*" (emphasis added).[91] However, one of the goals of EAI is precisely to render tangible what *to humans* is intangible by interpreting very subtle physiological signs from the body e.g., microexpressions, slight changes in vocal tone, or heart rate. This is part of a process of trying to infer what the person is thinking, without them voluntarily and/or consciously expressing it. The way the right to freedom of thought is structured, as well as the way it has been interpreted, emphasise a clear distinction between the *forum internum* and the *forum externum*. However, EAI, in some instances, appears to target what I would suggest to be the *forum limina*. The *forum limina* refers to things which sit in that liminal place between the unconscious and conscious. For example, consider facial microexpressions, tiny movements of the facial muscles, in response to some stimulus. People may not be consciously aware of "expressing" these movements, but they may nonetheless reveal

[89] Council of Europe, Declaration by the Committee of Ministers on the risks of computer-assisted or artificial-intelligence-enabled decision making in the field of the social safety net (2021) < https://bit.ly/3sgen4x > accessed 3 April 2021
[90] McGregor, Murray and Ng, 2019 (n 12) 329
[91] Ben Vermeulen, 'Freedom of Thought, Conscience and Religion (Article 9)' in Pieter van Dijk, Fried van Hoof, Arjen van Rijn and Leo Zwaak (eds) *Theory and Practice of the European Union Convention on Human Rights* (Intersentia 2006) 752

something about the agent's emotions or thoughts, though I must reiterate my scepticism as to the ability of EAI to *accurately* interpret emotional states. Importantly, EAI systems treat the things they detect, like facial movements, *as if they are* detecting emotions, and lead to consequences for the subject as previously discussed. As a result of this hard distinction between *forum internum* and *forum externum*, it may be hard for the right to freedom of thought to account for and define harms which arise from practices that *target* either the *forum externum* or the *forum limina*, but that are ultimately interested in revealing, manipulating, or punishing thoughts in the *forum internum*.

As regards dictating the obligations and expectations of States and businesses, I suggest that in the case of EAI and freedom of thought, IHRL as it stands offers very little guidance. Although McGregor et al. state that IHRL, 'provides established tests to assess when and how rights may have been violated', in the case of freedom of thought this is arguably not the case because there is next to no case law providing guidance or establishing such tests relating to how the right should be interpreted.[92] Lacking such guidance, States, businesses, and individuals are currently in a position of uncertainty regarding what uses of EAI are and are not permissible.

<u>5.0 – Conclusion</u>

This paper evaluated EAI in two key ways: from a surveillance studies perspective, and in terms of EAI's impact on the right to freedom of thought. It began by highlighting the methods and assumptions that underpin EAI, as well as the significant criticisms they currently face. It then gave an account of EAI that drew on various theories of surveillance studies, framing the surveillance EAI facilitates as 'emotional surveillance'. In this regard, some key takeaways from the paper are as follows. Emotional surveillance is characterised by a high degree of variability in the motivations for it, technology and applications that facilitate it, and the features of people that it surveills. EAI also has potential to scale quickly and is particularly liable to function creep. Emotional surveillance can result in a multitude of direct and indirect consequences for people, treats the body *as* data, makes predictions about people's future emotional state, actions, and performance, sorts people into various categories, and presents challenges to individual autonomy and dignity.

The paper then offered a novel critique of emotional surveillance by considering its impact on the human right to freedom of thought. In this regard, the main takeaway from the paper is that emotional surveillance may violate three key elements of the right to freedom of thought. Firstly, it may violate the right not to reveal one's thoughts by attempting to make visible and interpret people's emotions, which are herein understood as an important aspect of the thinking process. Secondly, EAI may violate the right not to have one's thoughts manipulated by exerting disciplinary power over the surveillance subject in the form of dictating what emotions are and are not acceptable in certain situations, and by allowing people to be targeted with particular content while experiencing a specific emotional state. Finally, EAI may violate the right not to be penalised for one's thoughts as it can be used as a means to grant or deny access to services, or as the basis for treating people less favourably than would otherwise be the case.

This paper also drew some conclusions as regards the utility of the IHRL approach to defining and addressing the harms AI systems may cause, though only insofar as EAI and the right to freedom of thought are concerned. In terms of the utility of the approach for *defining* harms, IHRL fails to offer established tests to determine if EAI violates the right to freedom of thought. However, Vermeulen's assessment of the right as having three key elements does provide scope to develop such a test in

[92] McGregor, Murray and Ng, 2019 (n 12) 326

the future. In terms of dictating the obligations and expectations of States and businesses as regards EAI and the right to freedom of thought, IHRL offers very little at the moment due to the lack of case law and soft law guidance. Finally, the structure of the right may be problematic as EAI *targets* the *forum externum* or the *forum limina*, which enjoy only limited protection, *as a means of* revealing, manipulating, or punishing thoughts in the *forum internum*, which enjoy absolute protection, making it difficult to determine whether absolute or limited protection should apply.

Finally, this paper is not claiming that emotional surveillance definitely does violate the right to freedom of thought, it simply suggests that it might, and that greater academic attention is required to investigate this issue further. Ultimately, rather than providing definitive answers, this paper hopes to raise further questions about the nature and impact of emotional surveillance on the right to freedom of thought.