Universiteit Leiden
The Netherlands

Emotional privacy: a philosophical investigation of emotion recognition technology

Prégent, A.P.

Citation

Prégent, A. P. (2026, September 2). *Emotional privacy: a philosophical investigation of emotion recognition technology*. Retrieved from https://hdl.handle.net/1887/4309114

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4309114

Note: To cite this publication please use the final published version (if applicable).

Part One:

The Technology

2. The prospect of 'fully functional' ERTs: A social, technological, and political overview[3]

2.1. Introduction

On 13 May 2024, at an OpenAI event, ChatGPT was asked to read aloud a bedtime story, which it did, even using different emotional tones. A few minutes later, an OpenAI researcher asked ChatGPT to assess his emotions through his facial expressions. The artificial intelligence (AI) model responded in a female voice: "Happy and cheerful with a big smile and maybe even a touch of excitement. Whatever's going on, it seems like you're in a great mood. Care to share the source of those good vibes?" (Robins-Early 2024).

The following year, Neurologyca announced the launch of its Kopernica platform. It stated that Kopernica was "designed to interpret [...] a broad spectrum of human emotional and cognitive states with scientific precision, and to do it in real time, at the edge, and in the moment." (Neurologyca 2025). This goes "far beyond surface-level sentiment analysis or basic emotion detection; Kopernica creates a layered, evolving understanding of the user – their current emotional state, their typical patterns, and their likely reactions" (Udinmwen 2025).

ChatGPT and Kopernica are only two recent examples that illustrate well the rapid expansion and central position of the field of affective computing in AI research, giving rise to sophisticated tools that may enrich and complexify the sociotechnical realm of possibilities for cohabitation with (emotionally) intelligent machines. Researchers, but also companies and governments, have come to recognise emotions as a core element of human-computer interaction (HCI), understanding that they are the key to enabling computer systems not only to understand human conduct but also to know how to act in the human world (i.e. 'emotionally aware AI'). Since Rosalind Picard first coined the term in 1997, the field of affective computing has exponentially expanded and is now a mainstream topic in AI. Affective computing studies the development of systems which recognise, interpret, process, and simulate human affects (Picard 1997: 77). It is thus surprising to learn that AI researchers were uninterested in finding ways to integrate emotions during the early days of AI.

While there were continual efforts to find new ways of representing and simulating human thoughts and reasoning processes (e.g. the Turing test and the ELIZA project), these efforts

[3] Parts of this Chapter are published in Prégent (2025a).

mostly followed the cognitivist approach developed by prominent philosophers and psychologists at the time (Boden 2018). Philosophers long claimed that the 'rational' part of the mind, believed to be responsible for the most important reasoning capacities such as problem-solving, decision-making, and overall 'intelligence' (Dukes et al. 2021), did not include emotions. This view strongly influenced the development of the classical cognitivist approach, which also excluded emotions from the capacity to reason (Simon 1967; Solomon 1976), and, in turn, shaped the development of AI. As Boden notes (2018, chapter 3), this is potentially the main reason why, for decades, AI researchers did not consider including emotions in their efforts to simulate the human mind.

In his book *What Computers Still Can't Do: A Critique of Artificial Reason*, Hubert Dreyfus proposed another reason, simply stating: "[t]his is again a case of not being able to see what one would not know how to program" (Dreyfus 1992: 276). Dreyfus thus argues that AI researchers may have deliberately ignored the potential role of affects for the simple reason that they did not understand them well enough to include them in their algorithms. This would be consistent with the way researchers, especially those in modern medicine and the social sciences more broadly, have approached affects: by simply ignoring them, deemed "too complex" or "irrelevant" for yielding compelling results.[4]

But it is worth noting that some researchers did acknowledge the hidden potential of emotions for AI research, including Marvin Minsky, who said in 1986: "[t]he question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions" (1986: 163). This opposed the radical rationalist approach, already challenged by various studies linking 'intelligence', 'decision-making', and 'rationality' to emotions. And it was further shaken by Rosalind Picard's work on affective computing (Picard 1997). Picard's work was ground-breaking not only as a novel approach to AI, but also because affective computing, as a distinct branch of AI research, represented a strong stance against a long-standing era in the scientific community. In her work, Picard brings affective life to the centre of AI research: a stepping stone for follow-up work on HCI, decision support systems, and other AI-driven managerial processes.

> Designers of future computing can continue with the development of computers that ignore emotion, that cannot sense what is important to people, and that make decisions according to brittle rules and laws. Or, they can take the risk of making machines that recognize emotions, communicate

[4] Modern medicine and AI research have both deliberately dismissed phenomena traditionally associated with femininity – such as emotions – due to misogynistic biases in the fields. This exclusion has ultimately backfired, stalling progress in both fields and harming humanity as a whole.

> them, and "have" them, at least in the ways in which emotions aid in intelligent interaction and decision making (Picard 1997: 251).

Picard's work has changed the way AI systems are developed. Following in her footsteps, many affective computing researchers have taken up the challenge of creating such machines, as proven by the development of (affective) robots like Nao, Harmony, or more recently, Ameca.

In this thesis, however, the focus is on emotion recognition technologies (ERTs), a comparatively modest subtype within the broader field of affective computing. Affective computing has diversified into a set of sometimes overlapping subdomains, including ERTs, sentiment analysis, mood detection, emotion generation, and affective forecasting. ERTs are commonly defined as systems that detect and classify individuals' emotional states from diverse input data, such as facial movements, bodily gestures, heartbeat, respiration, vocal characteristics, eye movements, skin conductance, and, more recently, brain activity. Some ERTs specialise in detecting and classifying emotions along dimensions such as arousal and valence. For example, negative emotions (e.g. anger, fear, anxiety) or positive emotions (e.g. excitement, joy, satisfaction). Others monitor a broader range of (prototypical) emotional states (e.g. joy, fear, anger, surprise, confusion). Unlike more ambitious strands of affective computing, which aim to simulate or endow machines with affective states, ERTs are designed only to recognise such states in humans; it is this specific capacity, and its normative implications, that I examine in this thesis.

As the thesis is anticipatory of a technology that does not yet exist, the purpose of this chapter is to provide an overview of the social, technical, and regulatory state of the art of current ERTs so that the anticipatory claim is properly grounded and justified by existing developments. Contingently, it also prepares the subsequent analysis of emotions and emotional expressions by motivating the need for an understanding of the potential impact of fully functional ERTs on the self, social relationships, and society.

The chapter thus surveys past, present, and anticipated developments in ERTs, discussing known biases, current technical limitations, and international regulatory responses. **Section 2.1** traces the historical emergence of ERTs and clarifies their conceptual foundations, identifying assumptions that still shape contemporary research. **Section 2.2** examines the technological modalities through which ERTs operate and evaluate their current accuracy and technical limitations. **Section 2.3** maps the expanding market for ERTs, including commercial platforms, healthcare, education, and policing, providing an overview of the various types of ERTs available, the goals driving their deployment, and the main sectors financing them. **Section 2.4** reviews public and scholarly criticisms of already deployed ERTs and shows the conceptual and normative ground covered to date in terms of risks and associated harms linked to bias, discrimination, and technical errors. **Section 2.5** analyses the emerging international regulatory landscape on ERTs, with particular attention to the EU AI Act, and identifies normative gaps it currently leaves with respect to fully functional ERTs, thereby motivating the ethical analysis that follows in subsequent chapters.

As more regulatory frameworks are adopting a risk-based approach, regulators will need reasons to justify or prohibit the use of ERTs. At present, the AI Act frames those reasons chiefly in terms of harms arising from technical unreliability and from data use and handling. This dissertation argues that, in the case of fully functional ERTs, a close analysis of ERTs, the phenomena they prey on, and the normative disruption they cause in the privacy realm makes it possible to assert that additional harms are foreseeable, and that these should weigh in the balance when democratic societies confront the inevitable trade-offs between innovation and the status quo.

2.2. ERTs' historical development

2.2.1. Face-based ERTs

ERTs first gained widespread recognition through the development of face-based models. Building on progress made in face recognition, early face-based ERTs relied on identifying relevant patterns between facial expressions and emotional states. From the early days of ERTs, a recurring concern among engineers was that humans could "conceal their true feelings" (Zhang et al. 2020: 104), which would compromise technological attempts at "recognition". This concern partly shaped the initial development of ERTs, and theories that promised to circumvent this difficulty were therefore widely preferred.

Ekman's basic emotion model and its variants were for this particular reason widely favoured by the emotion recognition community and quickly became the mainstream approach (Y. Wang et al. 2022: 23). Ekman's model postulated that emotional states could be inferred from facial expressions *alone*, and that basic emotions appeared on the face in *unique* yet *universal* patterns (i.e. each basic emotion has a unique facial pattern, and these patterns are found universally across different populations). Drawing on the work of Charles Darwin ([1872]/ 1965), Silvan S. Tomkins (1962), Carroll Izard (1969, 1971), and others, Ekman argued for the existence of seven universal emotions, which he termed the "basic emotions" (Ekman & Friesen 1971; Ekman 1992, 1993, 1994). From this research, he later developed the FACS system, standing for Facial Action Coding System.[5]

FACS, still available today, employs key points on the face to identify specific patterns in emotional facial expressions, which are then correlated with particular emotions. The identification of a given emotion is based on calculations of the variation in distance between the neutral face and the expressive face at those key points. Several image databases were subsequently created using this technique to train new ERT models.

Accordingly, most ERTs were initially trained on datasets such as CK+, EmotionNet, and AffectNet,[6] which provided static images of facial emotional expressions (Kopalidis et al. 2024). Ekman's basic emotion model proved so influential that, for over a decade, most companies grounded the scientific legitimacy of their technologies in his work (Crawford 2021; McStay 2016; Barrett et al. 2019; Tcherkassof & Dupré 2021). This was highly problematic, as it meant that ERTs were trained on *static* and *prototypical* facial expressions presented without context (Barrett 2017).

As pointed out by Barrett (2017) and many others, this meant that ERTs trained on these datasets exhibited several important weaknesses. A first major issue was that many of the individuals used as "models" for the datasets images were instructed to *pretend* to express a given emotion, with the result that the datasets often contained no images of *genuine* emotional expressions. Another issue stemmed from the fact that these ERT systems were trained on static close-up facial images, meaning that the datasets provided no contextual information and did not allow perceivers (or annotators) to take individual variability in emotional expression into account. As human perceivers, we often require contextual information to recognise another person's emotional expression appropriately. Inferring that someone is happy merely because they are smiling in a static image is closer to a guess than to a science, just as it is to infer that someone is angry solely because they are frowning. Everyday experience provides ample

[5] The Facial Action Coding System (FACS) is a comprehensive, anatomically based system for describing all visually discernible facial movement. It breaks down facial expressions into individual components of muscle movement, called "action units". See Paul Ekman Group (2024).
[6] Examples of other comparable datasets are Aff-Wild2, RAF-DB, and FERG-3D-DB.

evidence that people, for various reasons, do not always display their emotions through prototypical (and often Westernised) facial expressions. Barrett (2017) illustrated this with a close-up image that appeared to depict an individual in terror: her mouth was wide open and her eyes were tightly shut (chap. 3). The image was misleading, however. It did not show an expression of terror but rather Serena Williams after winning the 2008 US Open, expressing the intense joy associated with becoming a champion in her sport.

In their extensive review on the possibility of a one-to-one mapping between facial expressions and emotions, Barrett et al. (2019) concluded that, to date, there is no such evidence: "When facial movements do express emotional states, they are considerably more variable and dependent on context than the [basic emotion] view allows" (2019: 46). This push against the one-to-one map hypothesis was further confirmed by Tcherkassof's work (2018), who showed that in addition to context sensitivity, facial emotional expressions are also gender- and culturally sensitive.

In accordance with Barrett (2017, 2019) and Tcherkassof (2018)'s work, accounting for affects through face-based patterns of (micro-) muscles movements appears to be an arduous and almost inevitably ad hoc task, making it virtually impossible to programme into clear rules, as emotions do not appear to follow any specific and universal "rules" or "laws" when it comes to facial emotional expressions.

Yet this was precisely the ambition pursued in affective computing with the first face-based ERTs. Inspired by Ekman's work and relying on tools such as the FAC system (which proposed ways to encode these supposedly universal patterns) and available datasets (such as CK+, EmotionNet, and AffectNet), researchers in affective computing relied on simplified versions of the basic emotion theory,[7] oftentimes disregarding important challenges that were raised by other scientists, such as Barrett and Tcherkassof.[8]

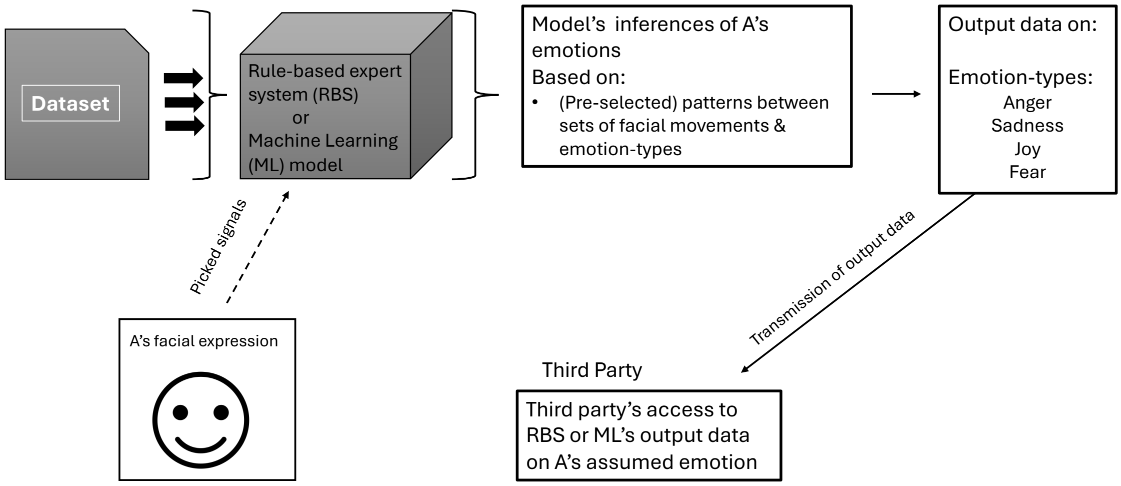
[7] Although Ekman defended the idea of basic emotions and a one-to-one mapping between these emotions and specific facial expression patterns, his overall theory is more nuanced, allowing space for display rules and other social factors to alter or challenge this mapping.

[8] As McStay notes, echoing Dreyfus, in retrospect, it should come as no surprise that research on emotion recognition has been heavily influenced by the work of Paul Ekman and his various studies on the existence of universal basic emotions, and most importantly, on the hypothesised one-to-one mapping with specific facial expressions. This is so because his work provided all the conceptual components that were previously lacking in order to bridge the gap between the tasks of capturing the affective life and programming it (McStay 2016).

Using different (but nonetheless similar) models and techniques, many tech companies have developed their own ERTs. Amazon was one of the first companies to launch its cloud-based software, Rekognition, in 2016, allegedly recognising seven emotions from facial features (happiness, anger, sadness, disgust, calm, confusion, surprise, and fear) (Amazon Web Services). Microsoft also launched its Azure Emotion Recognition service in 2016 (Lewis 2022), also known as API Face, which allegedly detected five emotions from facial features (fear, anger, happiness, surprise, and sadness). Microsoft claimed that the emotions detected were universally expressed across cultures (Crawford 2021: 155), thereby explicitly aligning with Ekman's theory of basic emotions. The startup Emotient, acquired by Apple in 2016, made similar promises to those of its competitors, allegedly using an ERT to detect individuals' emotions through facial expressions (Metz 2016).

Figure 1 illustrates how face-based ERTs typically try to infer an individual's inner emotional state based on the alleged unique patterns of facial emotional expressions.

Figure 1. Simplified Face-Based ERT Process[9]



An important difference between the ERTs developed for (most) research studies and the ones developed by major companies such as the ones mentioned above is that for such companies there is no need to restrict themselves to small datasets, as they are able to scrape their input data from massive social media platforms such as Instagram, Pinterest, TikTok, and Flickr to

[9] While some face-based ERTs could, in theory, only rely on rule-based expert systems, such as the ones directly derived from Ekman's model, others have also relied on simple forms of machine learning models. See Pantic and Rothkrantz (2000) for an example relying on an RBS approach and Stöckli et al. (2018) for an example of a study assessing the performance of two ML models, AFFDEX and FACET.

train their models, thereby accessing billions of images and increasing the presumption of better accuracy (Crawford 2021: 152-153).

With their development led by many of the world's most influential tech companies, face-based ERTs have reached almost all spheres of social life – including public spaces, retail stores, airports, schools, hospitals, HR processes, court hearings, law enforcement (e.g. border control), self-care apps, and social platforms. What should perhaps increase ethical worries is that these face-based ERTs, while widely endorsed despite their troubling shortcomings, only represent the beginning stage of the development of emotion recognition technology, now pointing towards more complex, multimodal systems.

2.2.2. Technological progress: From unimodal to multimodal ERTs

The strong focus on facial features, considered for a long time to be relevant (and sufficient) indicators of emotional states, has overshadowed other, potentially more promising avenues: multimodal systems using real-time feedback (Guo et al. 2024; Q. Wang et al. 2022). These systems depart from unimodal, face-based models in several important ways.

One major point of departure stems from the reliance on multiple types of input data, rather than on a single input type: they combine two or more different types of input data – such as facial expressions, body movements, tone of voice, eye movements, electrocardiogram (ECG) signals, and electroencephalogram (EEG) signals (Guo et al. 2024). Moreover, these multiple input data are now also extracted from combined input sources (such as video clips, audio clips, text), departing considerably from static images on which early face-based data models trained on. The datasets are thus rapidly improving. While face-based ERTs typically relied on datasets with over one million static images (e.g. AffectNet), in which these images mostly consisted of intentionally produced, prototypical emotional expressions, an increasing number of datasets are now being built using what expert annotators describe as "spontaneous" nonverbal emotional expressions: those produced "in the wild" and captured through video footage and audio streams, thereby moving beyond the limitations of static imagery. This helps mitigate problems associated with static images and prototypical facial expressions that were widely reported in earlier face-based ERTs.

Another major point of departure is the fact that the new datasets are now labelling emotions using multiple affective dimensions, including not only discrete emotion categories (like early face-based models), but also valence, arousal, and dominance (i.e. CMU-MOSEI dataset and IEMOCAP dataset)[10] (Yoon 2022).

It is worth noting that most improvements in multimodal ERTs over early face-based systems have only been made possible by recent and significant advances in AI along three closely intertwined dimensions: information-processing capacities, computing power, and shifts in modelling approaches from rule-based expert systems (RBS) and machine learning (ML) to deep learning (DL). Progress in information-processing capacities and computing power has made multimodal models viable, enabled real-time analysis and feedback, and supported increasingly fine-grained forms of emotional inference. At the same time, changes in modelling approaches have transformed how ERTs are built.

RBS relies on explicit, human-authored rules; in such systems, every rule used to classify emotional states from emotional expressions must be manually specified. For example, rules such as "If heart rate > 120 AND eyes wide AND open mouth = then classify as 'surprise'" have to be explicitly formulated for the system to correlate particular expressions with particular emotional states. ML methods, by contrast, learn statistical patterns from labelled datasets[11] based on pre-selected features (such as facial landmark distances, eyebrow angles, or mouth movements), which can then be used to classify inputs (e.g. images of people) as "happy", "sad", or "angry".

DL can be understood as a subset of ML in which deep neural networks "learn" to recognise both features and patterns directly from large amounts of data, typically requiring minimal human intervention in feature design or data labelling. For instance, a deep learning model trained on a large dataset (e.g. AffectNet or CMU-MOSEI) may "learn" which features and patterns are salient for recognising an emotional state such as happiness. A preference for DL can be observed in current ERT research (Zhang et al., 2020), primarily because DL, unlike classical ML, requires much less feature engineering and excels with unstructured data, such as images, audio, or text. Since unstructured data are the dominant forms of input in multimodal ERTs, DL has become the preferred method. However, this approach comes with a significant downside, commonly referred to as the "black box" problem.

The black-box problem refers to the opacity of many AI models, particularly deep neural networks, whose internal decision-making processes, while often stronger and more accurate in

[10] Respectively referring to the Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset, and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset.
[11] This typically necessitates that all data are properly labelled (see §2.2.3 for a discussion on the risks this entails)

their predictions than classic ML models, are not easily accessible or interpretable (Hassija et al. 2024). While the inputs and outputs are visible, the steps linking them often are not. This raises general concerns, particularly when the resulting classifications affect individuals. In the context of ERTs, the issue becomes especially salient. Because ERTs operate in domains involving interpersonal uptake and sometimes institutional decision-making, the opacity of their inferences makes it difficult to understand, challenge, or justify the conclusions drawn about an individual's emotional state. As a result, although DL methods likely improve overall performance, the black-box structure heightens the ethical risks associated with affective monitoring.

Yet the black-box problem, like the problems inherent to the first generation of face-based ERT models, does not halt the development or deployment of multimodal ERTs. On the contrary, because multimodal systems usually outperform unimodal ones, research continues to push forward. In their systematic review of affective computing, Wang et al. (2022) postulate that "by fusing physical information [e.g. facial, bodily, vocal] and physiological signals [pupil dilation, skin conductivity, heartbeat], useful features of emotional states can be obtained to enhance the performance of affective computing models" (Y. Wang et al. 2022: 19). Recent advances in multimodal emotion recognition based on ML and DL techniques, and ongoing developments in new methods for dealing with multi-channel emotion recognition are thus all contributing to the overall improvements of ERTs (Y. Wang et al. 2022: 23; Shoumy et al. 2020; Y. Jiang et al. 2020).

Today, an increasing number of ERTs on the market use not only multiple, but also combined input sources to detect individuals' emotional states (e.g. Noldus Hub, Smarteye, iMotions, Affectiva). ERTs can be embedded in a variety of ways, from wearable biosensors that measure respiration, heart rate, and skin conductivity, to voice and speech processors, video systems that track facial expressions, eye and body movements, and even headsets that monitor brain activity. Some have also expanded the conception of "emotional expressions" to include, for instance, interface-related behaviours (e.g. increased or decreased activity on an app, mouse movements, keystrokes, etc.) and broader online behaviours.[12]

[12] See Cognovi Labs for instance, which allegedly predicts emotional behaviours based on online textual data from social media posts or even online customer service interactions.

As for their predictive capacities, recent advancements suggest a significant increase in performance levels compared to unimodal models. It is possible that affinities between multimodal data sources (e.g. voice + face + text) may enable new ERTs to enhance both the quantity and quality of their output, potentially also increasing their predictive accuracy.

Moreover, like many other AI systems, ERTs are likely to improve in accuracy through training on larger and higher-quality datasets over time. This opens the door to new systems that may be personalised (i.e. tailored to individual users by learning idiosyncratic emotional patterns), adaptive (i.e. capable of learning after deployment), and context-aware (i.e. able to respond with a degree of situational sensitivity).

Current efforts to determine which types of input are most accurately associated with specific types of emotions suggest that emerging systems may rely on multimodal techniques to find new relevant patterns between the two. In addition, because physiological signals are generally understood to be responses generated by the central and autonomic nervous systems (and considered largely involuntary, therefore circumventing earlier concern around "humans concealing their true feelings"), they have become an attractive input type. Since such signals cannot easily be controlled, they help to avoid potential false positives caused by intentional attempts at emotional deception.[13]

This is not to mention the growing interest in ERTs that combine advances in brain-activity monitoring with recent progress in physiological sensing. By integrating multiple forms of physical input (e.g. images, video, audio) with physiological signals (e.g. ECG signals, galvanic skin response, photoplethysmography)[14] and now EEG signals, current and future ERTs may achieve more reliable levels of accuracy.

In light of this undeniable progress, it is necessary to consider the hypothesis that ERTs may soon reach a level of accuracy and robustness deemed acceptable for the tasks they are intended to perform. The possibility of a near future in which 'fully functional' ERTs become operational is what motivates this thesis, as it raises ethical, normative, and societal questions regarding their use – questions that go beyond those raised by current technical limitations.

Nevertheless, significant technical challenges remain on the path to achieving fully functional ERTs, and it is these challenges that the next section will address.

---

[13] See §2.3.5 for a discussion on technology-based deception detection.
[14] Photoplethysmography estimates the user's pulse.

2.2.3. Technical challenges: Noise, indeterminacy, and the absence of ground truth

ERTs face many technical challenges in their attempt to infer inner affective states from emotional expressions. Here, I discuss three major technical issues that recent ERTs have been grappling with: (1) noise, (2) indeterminacy, and (3) the absence of ground truth.

A key challenge for any emotion recognition technology is the presence of noise (Ahmad & Khan 2022). In information channels, "noise" typically refers to any external or irrelevant signals that interfere with transmission between sender and receiver. It is generally accepted that, for communication to succeed, the noise input must not overpower the intended signal input (Shannon 1948). In the case of a loud sound obscuring someone's voice, for instance, the signal-to-noise ratio may become unbalanced. In that case the noise is stronger than (or comparable to) the signal, preventing the recipient from distinguishing the intended signal from the noise.

In the case of ERTs, however, noise is not merely an environmental but a multi-dimensional problem. I identify three different types of noise, which are not mutually exclusive, and three key stages at which noise can impact the final output data: training-type noise, signal-type noise, and inferential-type noise.

Training-type noise refers to any noise that arises during the training stage of the model, including during dataset creation. This type of noise may include labelling errors, lack of precision, or biased annotation practices for instance (see Northcutt, Jiang, & Chuang 2021). These do not corrupt the quality of signal reception per se, but they undermine the quality of the learning process that enables that reception.[15]

Signal-type noise can be either environmental or technological. In both cases, it degrades the quality of signal reception. *Environmental noise* refers to any distortion, interference, or uncertainty that disrupts the accurate transmission of a signal from sender to receiver (e.g. deafening sounds or sudden flashes of light). *Technological noise*, by contrast, involves interferences with the original signal caused by the technology intended to capture it, such as low video resolution or microphone distortion. Both forms diminish clarity at the point of input.

[15] This noise-type although threatening, may be overcome without the need for innovation first.

Inferential-type noise refers to noise that arises during the interpretation of input data by the system. It may be internal or external to the system's inferential process. For example, inferential-type noise may be internal when the system is confronted with cross-modal conflict between different input types (e.g. contradictions or inconsistencies between voice and posture). External inferential-type noise may stem from a lack of contextual information, in which case the ERT's interpretation is disrupted by the absence of relevant situational information (i.e. situational under-specification).

Another major issue is the indeterminacy problem, which here refers to the fact that singular emotional expressions (e.g. a smile) do not follow a deterministic model in which a given internal affective state consistently correlates with a single observable prototypical 'emotional expression' (i.e. smile ≠ happy) (Dudzik et al. 2024). One reason for this lies in the fact that emotional expressions may have multiple possible causes and serve diverse social functions (see Chapter 4 for more on this point), which severely undermines one-to-one inference models (Barrett et al. 2019).

Despite the impressive progress made so far, ERTs remain confronted with a wide array of complex social contexts that may cause, shape, or obscure the expressive behaviours of the individuals they monitor (Wieser & Brosch 2012; Hess & Hareli 2015). As a result, these systems operate under conditions of ongoing indeterminacy. This means that when contradictory signals, epistemic constraints (or gaps), or interpretative ambiguity arise, the system is often left without the necessary resources to generate an adequate or reliable output.

A final, and related, major challenge for ERTs is that they cannot be trained on *genuine* ground truth data. Without diving into the intricacies of this term, in computing science, *ground truth* refers to information that is *known* to be real or true,[16] and against which a model's predictions can be verified (Martinez, Yannakakis, & Hallam 2014). For example, if a model is trained on images containing dogs and images without dogs, and its task is to identify which images contain a dog, its predictions can be verified against the ground truth dataset, which is, for instance, constituted of carefully hand-labelled images of dogs. By contrast, since emotions are not directly accessible phenomena, no *genuine* ground truth may be used. Instead, emotion recognition's ground truth datasets are typically created using: (1) expert annotators who provide

[16] Note that the terms "real" and "true" are used loosely in this context to mirror the way it is used in computing science, *not* in philosophy. See Chapter 1, §1.2.3 for an explanation of this choice.

their *own interpretation* of the *assumed* emotion felt by a person based on their emotional expressions, (2) self-reports from participants, or (3) proxy metrics.[17] As a result, ERTs are often trained on *assumed* labels rather than verifiable "truths" (Truong, van Leeuwen, & Neerincx 2007; Northcutt, Jiang, & Chuang 2021).[18]

As we shall see in the next chapter, emotions are extremely complex, dynamic, and multifaceted phenomena. This renders the accuracy of any training dataset inherently limited. Moreover, interpretative biases, whether from expert annotators, participants, or proxy metrics are virtually unavoidable.[19]

However, it is important to note that ERTs are not the only type of AI technology affected by these challenges. While every AI system faces noise challenges, other AI systems also face indeterminacy, and label uncertainty (often due to absence of ground truth). Decision-support system (DSS) models in criminal justice or healthcare, for instance, also face these problems as counterfactual outcomes are unknowable. An example of such a case is a DSS that is used to assess the likelihood of a defendant *reoffending*. While it can potentially be trained on cases where parole or release *was granted*, it is not possible for cases in which parole or release *was denied*. In such instances, the model's predictions cannot be compared to any *genuine* ground truth, as there is no way to know what *would have* happened had the individual *been released*. A similar case-example can also be found in healthcare, where AI-based DSSs are used to predict whether a comatose patient has a sufficient chance of recovery (Schoeffer, de Arteaga, & Elmer 2025). These systems can be trained on data from patients *who did receive* the treatment and whose outcomes – positive or negative – are known. However, they do not have access to outcome data for patients who *were never given the treatment*, as there are simply no data for an event that never occurred.

[17] Instead of direct ground truth, such as whether a user "enjoyed" a movie, data scientists can monitor proxy metrics like user engagement or click-through rates for a recommender system to gauge this "enjoyment".
[18] For instance, Truong, van Leeuwen, and Neerincx (2007) define a "ground-truth emotion" as "an emotion that is perceived by people, agreed upon by most receivers and at the same time, the experienced true emotion as felt by the person her/himself".
[19] It should be noted that recent proposals have been made to adopt alternative annotation practices instead, such as perspectivist ground truthing (see Cabitza, Campagner, & Basile 2023). This approach, for instance, proposes to take into account all available annotations and other additional information about the people annotating, such as their confidence or uncertainty level.

2.3. ERTs on the market: From technical artefacts to embodied emotional AI

A compelling reason to support the hypothesis that *fully functional* ERTs may soon become available, besides the technological progress surrounding their development, is the global market's eagerness for such technologies to exist. In the aftermath of the information revolution, the ability to understand human emotions is increasingly viewed as a critical component of AI development – and, by extension, as essential to the infrastructure of so-called smart cities. As a result, the market for ERTs has expanded rapidly over the past decade.

In 2020, the ERT industry was estimated to be worth approximately $20 billion (Schwartz 2019), marking the first widespread global deployment of these systems in airports, schools, social media applications, HR processes, and law enforcement tools (Crawford et al. 2019). The industry was projected to surpass $50 billion by the end of 2024 (Miles 2024), and is now on track to exceed $130 billion by 2030 (Grand View Research 2023).

Emotion recognition technologies have emerged across a wide range of global markets, including retail (e.g. emotion-responsive customer analytics), the hospitality industry (Kairos, Smart Eye), the sex industry (Realbotix), education systems (iMotions, Find Solution AI), healthcare institutions (SoftBank Robotics), national security and border control (iBorderCTRL project, AVATAR system, Silent Talker), and politics (e.g. Cambridge Analytica; Mascellino 2022; Hall 2024).

Giving a compelling overview of the scope and range of ERTs in society is a hard task. For one because ERTs are multistable systems and polyfunctional technologies, meaning that they can be used and interpreted in different ways, while also serving multiple functions. For another, this is because it is sometimes difficult to separate them from broader affective computing purposes.

ERTs refer to technological artefacts designed and used as tools to achieve practical ends related to affective life. This includes technologies that extend or augment an individual's affective capacities, as well as those that scale such capacities at the collective level (e.g. in surveillance, commercial, healthcare, educational, or employment contexts). By contrast, emotional AI or affective AI agents typically refer to current or future interactive and embodied systems whose (ultimate) aim is akin to functioning as autonomous, emotionally intelligent agents. In the remainder of this section, I first provide a brief overview of "standard" ERTs, then move on to ERTs in self-tracking apps and to (semi-)embodied affective AI, before turning to lie detectors.

2.3.1. Standard ERTs

The multistability and polyfunctional use of ERTs may be easily acknowledged through a brief overview of some of their practical applications. Currently available ERTs may assist users in selecting their next song, help them understand others' emotions in real time, assess a person's "suitability" for a job, guide teachers in evaluating student engagement during class, or support doctors in the diagnosis of conditions such as Parkinson's disease, Alzheimer, dementia, or depression (de Sant'Anna, Morat & Rigaud 2012; Garcia-Garcia et al. 2022; Kim & Lee 2023). ERTs may also serve as preventive tools. They can stop someone from driving if they appear too tired, stressed, or drowsy (Macarol 2025), assign emotional scores in workplaces or schools (Fisher Philips 2024; Mantello et al. 2023), or even be used to remove individuals from sporting or cultural events if they are flagged as exhibiting suspicious or "pre-criminal" behaviours based on facial expressions, gait, or eye movements (Wright 2021).[20]

Accordingly, some may view these systems as oppressive while others may regard them as helpful. It is therefore crucial to recognise first and foremost that ERTs may serve multiple purposes across many – if not almost all – sectors of society. This breadth of function significantly affects the risk profile associated with their deployment, and should be taken into account in regulatory processes.

Companies such as Amazon, Smart Eye, iMotions, MorphCast, HireVue, Kairos, and Neurologyca specialise in developing ERTs that can be purchased for a variety of purposes, including HR processes, e-learning, digital marketing, and security.

Standard ERTs typically assign emotional labels – such as "happy" – to the data subject under observation. These labels may also take the form of dynamic scores that fluctuate in real time as the individual interacts with their environment (e.g. 5% sad, 5% angry, 10% surprised, 0% disgusted, 80% afraid). Such systems can provide fine-grained real-time reports and may also generate predictions about future emotional states. For instance, Kopernica, an AI platform designed as a plug-in for other AI systems, tracks over 790 body points using a combination of

[20] Potentially before any violent act has taken place (Wright 2021).

cameras, microphones, and data fusion techniques to 'understand' human emotions more effectively (Udinmwen 2025).

It is also worth mentioning emerging types of ERTs that rely on brain signals and neuroimaging to monitor and recognise individuals' emotional states (Chen et al. 2019; Suhaimi et al. 2020). I refer to these EEG-based systems[21] as "neuro-ERTs" because they bridge the gap between neurotechnology and standard ERTs. Using non-invasive methods, such as headsets, neuro-ERTs may monitor the neural networks underlying emotional processes. Although neuro-ERTs remain in their infancy, some progress has already been made – for instance, they already enable the monitoring of affective states during virtual reality (VR) immersion. This technique is already being experimented with in the military, healthcare, and educational contexts (Johnson-Glenberg 2018; McIntosh et al. 2019). Notably, the development of neuro-ERTs – but this is also the case for nearly all research on ERTs – begin from the premise of assisting individuals who, due to health conditions or other reasons, are unable either to produce or recognise emotional expressions (e.g. in cases of Parkinson's, dementia, depression, or autism). Neuro-ERTs are therefore often presented as a means of addressing inequalities caused by physical disabilities, mental disorders, or illnesses: by equipping individuals with intelligent tools that enhance their capacity to interact socially or to be more effectively diagnosed and cared for (Devlin 2025).

2.3.2. Self-tracking apps

Among the various types of ERTs, affective self-tracking apps represent a particularly popular category. Mostly composed of emotional health apps, well-being apps, and mental health apps, they constitute a significant portion of the global ERT market. Their common promise is to enhance "happiness"[22], reduce stress, anxiety, and negative thoughts on a daily basis.

Apps such as Happimeter, MindShift CBT, and the Amazon Halo wristband offer to track, learn from, and guide users through their daily emotions, moods, and stress levels. The Happimeter, connected to a smartwatch, collects input from various sensors and provides detailed reports with personalised suggestions (Budner, Eirich, & Gloor 2017). MindShift CBT, developed by the non-profit organisation Anxiety Canada, is a self-report app based on cognitive behavioural therapy, aimed at reducing anxiety and stress by interrupting negative emotional cycles. Amazon

[21] EEG signals are captured through electrodes placed on the scalp, and they detect and monitor electrical activity produced by neurons.
[22] Loosely conceptualised.

Halo detects emotional states through vocal tone analysis, with the stated aim of advising users on "how to interact more effectively with others" (Day 2019).

2.3.3. Aggression detector

The detection (and prediction) of aggressive behaviours has driven another significant portion of research on ERTs, highlighting the multiple and wide-ranging functions that such technologies may serve. Usually referred to as "aggression detectors", these systems are primarily designed for security purposes and are deployed in public spaces, hospitals, schools, prisons, banks, retail stores, public transport, and private company offices as a new mean of preventing violent acts, harassment, and potential threats. For instance, Amazon's Rekognition software is being implemented in UK train stations to monitor individuals in the hope of preventing unwanted and aggressive behaviour as well as increase safety across the public transport system (Burgess 2024).

Many other aggression detectors are voice-based, such as the systems developed by Sound Intelligence and Louroe Electronics (Stanley 2019). Sound Intelligence is one of the leaders in the field, with a system that allegedly detects fear, anger, duress, and verbal aggression. The company asserts that its aggression detector can "identif[y] aggression in a person's voice". It can function either as a stand-alone system or be integrated with video surveillance solutions. The analytic software sends an automatic notification to staff through a visual alert or by triggering an alarm.[23] This early warning allows security personnel to respond swiftly in order to prevent physical aggression. The company tested the model in 11 locations across the inner city of Groningen (Gillum & Kao 2019). The system was reportedly deemed a success and described as an indispensable tool by the Dutch police, railway companies, and prison authorities (Gillum & Kao 2019). The technology allegedly helped security staff at a hospital in Florida by issuing a warning "before the staff could push a panic button" (Gillum & Kao 2019).

2.3.4. (Semi-)embodied affective AI: From chatbots to robot companions

Up to this point, I have more or less implicitly treated ERTs as tools for assessing "present" emotional expressions and making inferences about future likely emotional reactions, actions, motivations, and intentions. Yet these "tools" exist not only as standalone devices but also within a wider ecosystem of AI systems whose ambitions extend beyond recognition, to the

[23] The aggression detector can either function as a stand-alone system or be integrated with video surveillance solutions (see Sound Intelligence).

simulation of emotional life. I refer to such system as (semi-)embodied affective AI (agents). While some possess a physical shell (e.g. sexbots), others are "embodied" only at the interface level, such as chatbots, therapy bots, or companion bots. I overview them below.

Semi-embodied affective AI systems

Screen-dependent companion bots, therapy bots, chatbots, and, more recently, so-called grief bots and postmortem avatars (AI-generated bots and avatars of deceased individuals) are what I refer to as semi-embodied affective AI systems.[24] Alexa and Siri, for instance, among the first chatbots to integrate basic emotion recognition functionalities, are now accompanied by a wide range of competitors from home and family bots to well-being and therapeutic bots. Regarding the latter, Woebot and Wysa are prime examples, as they specifically target mental health conditions, including depression and anxiety (Fitzpatrick et al. 2017; Inkster & Subramanian 2018).

Companion bots such as Xiaoice, launched in China in 2014, and Replika, a U.S.-based product deployed in 2017, offer users the option of configuring the bot to take on the role of a friend, a spouse, or even a coach. These two affective bots remain among the most popular screen-dependent companion systems today, providing an alternative to human–human companionship for individuals who feel isolated or in need of emotional connection.

The company Hume AI now also offers a range of chatbots (such as EVI and Octave) that respond to and adapt their interactions based on analysis of the user's current emotional state and predicted future emotions (Zeitchik 2022).

Noteworthy here is also the recent launched of ChatGPT. Although considered a general-purpose chatbot rather than an emotion-focused AI system, ChatGPT is worth noting here due to its performance in emotional awareness evaluations (Elyoseph et al. 2023). GPT-4 and GPT-4.5 demonstrated improvements in their 'EQ' (emotional quotient), particularly in text-based emotion recognition and their capacity to simulate expressions of genuine empathy. Discussions have already begun about whether individuals might feel more comfortable discussing personal struggles with AI "psychologists" than with human therapists. Several surveys suggest that users

[24] The principal aim of grief bots, such as the ones developed by Project December, is to recreate a deceased person as a chatbot in an attempt to help individuals cope with the loss of loved ones (Metzger & Kirchengast 2025). By contrast, postmortem avatars have been designed to replicate or extend a deceased person's identity, presence, or agency, regardless of whether the function is grief-related (Hollanek & Nowaczyk-Basińska 2024).

often regard interactions with ChatGPT and similar systems as truly "judgement-free" zones – a sentiment that has also been reported in educational contexts (Cross et al. 2024).[25] In any case, these screen-dependent bots may be regarded as positioned between ERTs and affective AI agents, since they interact with humans but are limited in their modes of embodiment.

Robots companions

In *Klara and the Sun* (2021), Kazuo Ishiguro imagines a possible future world in which emotional AI companions are widespread and widely embraced. Children may have their own affective bots to help them navigate life, better understand themselves, and relate to others. However, while utopian and dystopian literature often precedes technological reality, this is not the case here: robots similar to Klara were already being developed and tested more than a decade before the publication of Ishiguro's novel.

The healthcare sector and the sex industry were among the first to lead the development of emotional AI robots. As early as 2010, the ALIZ-E project, a European-funded initiative, was launched with the mission to design and build robots capable of sustaining long-term human-robot interaction. Researchers from the Netherlands,[26] Belgium,[27] Germany,[28] and the UK[29] collaborated with the aim of enabling these robots to forge constructive bonds with young patients at a hospital in Milan (Hornyak 2010). The robots were specifically tasked with showing empathy and engaging socially with children. One of the project's eight core technological goals was to analyse and synthesise emotions and affects in human–robot interaction.[30] ALIZ-E, however, was only one among many such initiatives.

QTrobots, developed and launched by LuxAI in 2021, are specifically designed to offer interactive affective learning for children on the autism spectrum.[31] Nao robots, created by the French company SoftBank Robotics[32] and now in their sixth generation, are also used in autism-related contexts, but their applications have expanded to include a wide range of uses in research, education, and healthcare institutions worldwide. Similar healthcare companion robots

---

25 Please note that this segment on ChatGPT is based on research done between 2023 and 2024 and consequently does not cover later developments.
26 Netherlands Organisation for Applied Scientific Research (http://aliz-e.org/.)
27 Vrije Universiteit Brussel (http://aliz-e.org/.)
28 Deutsches Forschungszentrum für Künstliche Intelligenz (http://aliz-e.org/.)
29 Plymouth University, Imperial College London, and University of Hertfordshire. (http://aliz-e.org/.)
30 See (http://aliz-e.org/).
31 See (https://luxai.com/press-release-april-2021-qtrobot-for-home-launch/).
32 Now owned by the Japanese company SoftBank Mobile.

for the elderly include the UK-based MiRo robot, developed by Sebastian Conran Associates in collaboration with the University of Sheffield (McMullan 2016), and the Japan-based Paro robot, developed by AIST (Tarantola 2017).

Alongside these companion robots are so-called humanoid robots, such as the sexbots Harmony from Realbotix and Ameca, the flagship figure of entertainment robotics.[33] All of these systems place strong emphasis on simulating "real" emotional expressions and affective states as part of their effort to establish long-term bonds with human users.

2.3.5. Border control & interest for lie detection

> "For it is the affect that accompanies the verbal claim – in effect, the paralinguistic cues – that is perceived as the true window to the person's essence."
> (Weisman 2014: 32)

Aside from the types of ERTs discussed above, another significant segment of the ERT market comprises technologies used for security and border control purposes. Emotions and emotional expressions have long played a central role in how individuals assess the character and intentions of others. Others' emotions inform some of our most consequential judgements – whether confronting a spouse about suspected infidelity, deciding whether to entrust a colleague with sensitive information, or serving on a jury tasked with determining whether someone should be sentenced to death. In the latter case, displays of remorse and guilt are often the most scrutinised emotional expressions juries look for in defendants (Sundby 1997; O'Hear 1996; Rocksheng et al. 2014). Legal systems around the world have historically taken defendants' emotional expressions (or the perceived absence thereof) into account in both civil and criminal cases (Zebrowitz 1997; Bandes 2014; Wistrich et al. 2015). In extreme criminal sentencing, such as capital punishment, perceptions of untrustworthiness based on facial appearance can aggravate the sentence – sometimes tipping the balance towards a death sentence rather than life imprisonment (Wilson & Rule 2015).[34] As US Supreme Court Justice Anthony Kennedy remarked in 1992, knowing whether a defendant feels guilt or remorse is central to "know[ing] the heart and mind of the offender" (Riggins v. Nevada 1992; Barrett 2017).

[33] Designed by Engineered Arts, UK. (https://engineeredarts.com/.)
[34] Wilson & Rule (2015) found that perceptions of untrustworthiness predicted death sentences (vs. life sentences) for convicted murderers in Florida.

But the idea that a person's emotional expressions reveal their *true* state of mind has always been controversial, based on the fact that individuals can sometimes *deceive* others in this regard – and that *some* people are particularly skilled at it. This follows a common belief that people's inner states are never fully accessible to others, and therefore certainty in this regard will never be possible. Parallel to this belief is the belief that affective states are correlated with physiological bodily symptoms, making them susceptible to detection. Thus, if the 'correct' correlations between bodily symptoms and internal affective states could be discovered, all attempts at deception could be circumvented and a person's 'true' affective states could be accessed regardless of whether the person is willing to share them or not. This latter belief was the driving force behind the development of the polygraph, often referred to as a "lie detector", "deception detector", or "truth detector", a forerunner to contemporary ERTs.

Already in the early years of the 20th century, William Marston (1917) proposed to detect the deceptiveness of suspected criminals by monitoring the person's heartbeat and blood pressure.[35] The assumption was that 'hidden thoughts' – thoughts that the subject is actively trying to conceal from the examiners – produce 'galvanic effects' that can be detected (Marston 1917: 117).[36]

Many studies have been conducted on lie detectors (Lykken 1974; Honts 2004; Vrij et al. 2011). These devices have been used primarily by law enforcement agencies, and their results have also been accepted as evidence in various courts of law (Honts, Thurber, & Handler 2021; Meijer & von Koppen 2008; Widacki 2007). Primarily focused on detecting deception, the various existing kinds of lie detectors were originally based on heartbeat frequency, perspiration rate, blood pressure, and breathing patterns (Maschke & Scalabrini 2000). Their results, however, have been *highly* contested, with field data indicating high false negative rates (Elaad, Ginton, & Jungman 1992; Honts, Thurber, & Handler 2021). Despite these findings, interest in deception detection has not diminished,[37] and the quest to identify relevant indicators from physiological states has continued (van der Zee et al. 2019), resulting in an increasing number of national and supranational projects linked to preventive and punitive purposes.

[35] Marston proposed extending the work of Cannon (1914), which found that six emotions could be correlated to the autonomic nervous system, causing bodily reactions that are not controlled by the subject (see Cannon 1914: 263).

[36] Interestingly, only in 2020 was John Augustus Larson recognised as having invented the first polygraph (Widacki 2020).

[37] Discern Science Inc. (DSI), for instance, an AI company, acquired an exclusive worldwide licence for a broad portfolio of deception-detection technologies in 2018. DSI initial "target" markets for these technologies include airports, government institutions, mass transit hubs, and sports stadiums.

Several projects have already been deployed and made public, including SPOT, AVATAR, and iBorderCtrl. In the years following the 9/11 attacks, the US Transportation Security Administration attempted to shift its security practices towards a risk-based approach (Halsey 2013). To this end, they developed and deployed the SPOT programme (Screening of Passengers by Observation Techniques) in several airports across the United States. Using over 94 criteria, SPOT (allegedly) automatically detected travellers' facial expressions, flagging signs of stress, fear, or deception (Crawford 2021: 170–171).

The United States followed the SPOT programme with the AVATAR project. AVATAR was likewise developed to enhance the US border control screening process, but unlike SPOT, it was specifically designed as a deception detector for use at international borders (Kendrick 2019). AVATAR, which stands for "Automated Virtual Agent for Truth Assessments in Real Time", is owned by Amazon and assists border control agents by directly questioning travellers and detecting physiological and behavioural changes during the border control verbal exchange (Javelosa 2017). The system monitors changes in the eyes, voice, gestures, and posture to determine potential risks, and "it can even tell when you're curling your toes" (Javelosa 2017). When suspicious answers or reactions are detected, they are classified into four predefined categories: green (low risk), yellow (medium risk), orange (medium-high risk), and red (high risk). Depending on the type of risk identified, customs officers may carry out further investigations. The technology was for the first time tested in 2012 at the Arizona-Mexico border (Greenemeier 2012), and subsequently by the European Union at Bucharest Airport in 2013. During the EU Hooligan experiment, AVATAR was reported to have an accuracy rate of 85% in identifying imposters, while human border officers detected only 67% of the total of imposters (Elkins et al. 2014). Moreover, humans exhibited a false positive rate of up to 80% for "innocents", whereas the AI system produced only 15% false positives. The EU-approved report described it positively, stating that it "has the potential to improve and extend existing automated border control (ABC) technology in the EU" (Elkins et al. 2014: 2). Still, the technology has not yet been officially deployed, as it remains under development.

Similar to AVATAR, the European Union designed iBorderCtrl, a deception detection system intended to assist border agents in identifying travellers' lies, which has also undergone extensive testing in recent years (Marda & Jakubowska 2023). Rooted in the Silent Talker technology (Kennedy 2014; Bittle 2020), the system features a virtual border agent that questions travellers while analysing their micro-expressions, which are regarded as sufficient and reliable biomarkers of deceit (Sánchez-Monedero & Dencik 2022). The virtual agent is tasked with questioning travellers about their personal background and their intentions for entering EU territory (Boffey 2018; Heaven 2020). These are only a few examples of so-called smart border control technologies, but similar projects are being developed globally, and lie detectors are increasingly used by government agencies and private companies for other tasks, such as hiring and fraud screening (Harris 2018; Bittle 2020).[38]

2.4. Biases, criticisms, and anti-sentiment towards ERTs on the market

Although emotion recognition technology has gained popularity in a wide array of sectors, it has also faced strong criticism, which is worth discussing here. Alongside academic researchers, political and advocacy groups such as AI NOW, Big Brother Watch, European Digital Rights (EDRi), and Article 19 have strongly opposed the use of ERTs. Some countries and supranational unions (e.g. the EU) have also issued explicit statements against their deployment (Information Commissioner's Office 2022; Artificial Intelligence Act (Regulation (EU) 2024/1689) (hereafter EU AI Act).[39]

Despite its decades-long presence, emotion recognition technology has been accompanied by controversy with nearly every new launch. Recurrent criticisms highlighted the lack of scientific justification for the core assumptions underpinning ERTs, which has also been directly linked to their high risk of discriminatory bias (MacAskill 2015; Boffey 2018; Schwartz 2019; Crawford et al. 2019; Miles 2024; Wang et al. 2023).

This pushback comes after several scandals, in which primarily face-based ERTs were found to be either biased or based on false claims. In 2018, for instance, researcher Lauren Rhue demonstrated that two face-based ERT models – Face++, and Microsoft AI, exhibited significant racial biases (Rhue 2018). Testing the two systems on over 400 images of NBA players, her research showed that Face++ *systematically* assigned more negative emotional scores

[38] See the case of the UK border Agency (Parliamentary Office of Science and Technology (2011), the use of Converus at the University of Florida Police Department (see Converus), and police departments in Salt Lake City and Columbus, Georgia. (see Bittle 2020).

[39] The Information Commissioner's Office (ICO) is "warning organisations to assess the public risks of using emotion analysis technologies, before implementing these systems. Organisations that do not act responsibly, posing risks to vulnerable people, or fail to meet ICO expectations will be investigated […] The inability of algorithms which are not sufficiently developed to detect emotional cues, means there's a risk of systemic bias, inaccuracy and even discrimination" (ICO 2022). Similarly, the EU AI Act has classified ERTs in the category of 'high-risk' technologies and explicitly banned their use in the workplace and educational institutions.

to black-skinned players on average than to white-skinned players. The Microsoft AI programme identified them as, on average, more 'contemptuous' than their white-skinned counterparts. In all cases, these scores were found to be incorrect due to racial biases.

Many other companies have also either admitted to, or been accused of, using unreliable ERTs. For instance, Sound Intelligence's CEO, Derek van der Vorst, has admitted that their aggression detector was "imperfect", registering "rougher" tones as aggressive and at times "misconstru[ing] innocent behaviour" (Gillum & Kao 2019). The EU-funded lie detector iBorderCtrl, trialled in Hungary, Greece, and Latvia, has been mainly criticised for relying on pseudoscience to support a publicly funded initiative aimed at reducing illegal immigration in the EU (Marda & Jakubowska 2023; Boffey 2018). Similarly, the US programme SPOT was widely criticised as severely biased against minorities and disadvantaged groups, reinforcing systemic injustices. By flagging signs of anxiety or stress, this $900 million programme was effectively reduced to a racial profiling tool, in which social groups that experienced on average greater stress and anxiety when interacting with the police or other authorities (e.g. security, customs, etc.) were disproportionately identified as potential threats (Halsey 2013). The programme itself was strongly criticised by the US Government Accountability Office and the Legal Associations for the Defence of Rights and Freedoms, whose criticisms focused primarily on the lack of evidential support and scientific methodology underpinning it (Crawford 2021: 170-171). Although signs of stress might have been correctly identified, they did not reliably correlate with imminent threats, leaving the risk-assessment programme evidentially unsupported. As a result, the programme contributed instead to reinforcing a vicious cycle of distrust and disproportionate scrutiny between historically oppressed and discriminated-against groups (i.e. Afro-descendants and Arab communities) and the authorities.

Additionally, a major concern already emerging from the use of ERTs is linked to a tendency among companies to use them as tools to exert greater control over their employees. In Japan, Aeon supermarkets have introduced *Mr. Smile*, a system that monitors employees according to the "length and sincerity of [employees'] smiles" (Fischer Phillips 2024). And *Mr. Smile* is only the tip of the iceberg. Snap Inc. recently obtained a patent for its *Emotion Recognition for Workforce Analytics* project, which proposes measuring employee performance through emotional metrics. The patent notes: "[e]mployers may want to introduce control processes to monitor the emotional status of their employees on a regular basis to ensure they stay positive and provide high-quality service" (Rubio-Licht 2024).

Affectiva's Affdex SDK, Microsoft's Seeing AI, MorphCast AI, BeEmotion, HireVue, Kairos, and many others have developed ERTs specifically for hiring, employee management, and assessment (McDuff et al. 2016; Escobar 2019). The underlying assumption is that emotional scores reveal more than just an individual's internal state at time *t*; they are treated as relevant indicators of broader personality traits or behavioural dispositions. These include predictive markers for diseases and illnesses, physical states such as fatigue, mental states such as anxiety and stress, and character traits such as trustworthiness, engagement, aggressiveness, impatience, and frustration.

Conversely, though, it is worth noting that other companies have adopted ERTs with the aim of assisting employees rather than monitoring them. SoftBank, for instance, plans to implement a new voice-altering artificial intelligence system that "aims to make the lives of call operators less stressful – by making irate customers sound more relaxed" (Hamzah 2024).[40] The CEO of Hume, Alan Cowen, stated that the Hume AI platform could potentially assist crisis call centre operators in detecting signs of depression in callers' vocal tones, possibly even preventing suicide attempts (Zeitchik 2022). Trained on both facial and vocal expressions, the Hume platform categorises over 20 emotions from multimodal inputs collected from the face, voice, and body. In 2024, Cowen claimed that Hume AI could "predict how a given speech utterance or sentence will evoke patterns of emotion" (Miles 2024).

Nevertheless, doubts about the ways in which major companies have started to use these technologies are only increasing. The company HireVue, whose clients include Intel, Goldman Sachs, and Unilever, has used ERTs in its AI-based assessment of job candidates, attempting to determine whether a potential hire is 'motivated', 'curious', and possesses 'grit' (Harwell 2019). Thus, in response to these claims, serious doubts about whether ERTs can ever accurately identify such affective 'qualities' have been publicly expressed by several affective researchers (Barrett et al. 2019; Tcherkassof & Dupré 2021; Mantello et al. 2023).

[40] Although the intended purpose of these tools may shift – from instruments of control to mechanisms of support – ethical questions, beyond what is covered by this dissertation, remain. In particular, concerns arise about the suppression of individuals' freedom to express disagreement, disapproval, or dissatisfaction through their vocal tone. Equally pressing is the question of whether individuals should be prevented from feeling unwell, sad, or even depressed. In a technocultural environment where emotions are commodified and negative affect is increasingly flagged as problematic, this issue becomes particularly contentious in relation to employers' duties, responsibilities, and the limits of their reach into employees' lives. While offering support upon request is one matter, imposing unsolicited emotional diagnoses on employees risks being oppressive and may carry broader societal consequences. Negative emotions, just as positive ones, are an integral part of the human experience and play a vital role in psychological resilience, emotional learning, and individual flourishing. They help us better understand ourselves, and their expression also enables others to regulate their responses to us and to discern our needs and expectations.

Political activists and scholars have severely criticised the use of *biased* ERTs. ERTs relying solely on facial expressions (i.e. face-based ERTs) have been the subject of the strongest criticisms. A substantial body of work has challenged the core assumptions of basic emotion theory on which many face-based ERTs rely. Barrett (2017) and Barrett et al. (2019) dispute claims of cross-cultural universality in facial expressions of emotion and reject the idea of a stable, one-to-one correspondence between discrete emotional states and specific facial configurations. Rather than functioning as universal, reliable readouts of internal states, patterns of facial expression are shown to be variable, context-sensitive, and multiply determined (Barrett 2017; Barrett et al. 2019). As Barrett et al. note, "similar configurations of facial movements variably express instances of more than one emotion category" (Barrett et al. 2019: 1). The central difficulty, therefore, is that face-based ERTs typically rely on one-to-one reverse inference, meaning that they treat a facial expression as evidence of a particular emotional state. This inferential model is scientifically unsupported and leads to significant risks of misinterpretation. Along similar lines, Cabitza, Campagner, and Mattioli (2022) argue that we cannot meaningfully speak of "accuracy" with face-based ERTs, since no reliable ground truth exists against which error rates could be computed (2022: 14).

Importantly, these concerns extend beyond face-based systems. Any ERT built on biased data, flawed assumptions, or faulty inferential models risks producing misleading conclusions about an individual's emotional state. Such errors are far from trivial. As critics have noted, they may contribute to wrongful arrests, disproportionate use of force, discriminatory surveillance, or the denial of fundamental rights such as access to travel, employment, loans, parental custody, parole, or even, in extreme cases, the right to life (Wilson & Rule 2015; Barrett 2017; Crawford et al. 2019; Crawford 2021; Mantello et al. 2023).

But as McStay and Urquhart (2019) warn, the current push for higher accuracy in ERTs risks incentivising more 'invasive' practices than those used by today's largely unimodal systems. The authors' warning is coherent with Paiva-Silva et al.'s observation that "[r]esearch that uses functional and structural MRI or EEG as methodological alternatives has increased in the last decade" (2016: 153)[41], signalling a broader shift towards more invasive approaches. Taken together, these critiques reinforce the broader aim of this thesis: to bring philosophical analysis to bear on the design, deployment, and legal regulation of such systems.

[41] "MRI" stands for magnetic resonance imaging.

2.5. A look at the current international regulation of ERTs

Internationally, regulatory frameworks for ERTs remain brittle. In recent years, however, many countries have begun to adopt similar rules governing the collection and use of consumer and citizen data, and have developed new frameworks for specific AI systems, including facial recognition technologies, biometric systems, autonomous vehicles, drones, and deepfakes (or synthetic media). Most of these initiatives also address broader classes of AI systems deemed potentially harmful or discriminatory, often described as "high-stakes", "high-risk", or "high-impact" AI systems. This terminology typically refers to the use of AI in predictive and decision-making contexts such as policing, hiring, lending, insurance, and education, as well as in algorithmic profiling.

ERTs are thus not entirely unregulated, as their deployment can fall under existing data-protection and consumer-protection instruments (e.g. the European General Data Protection Regulation (GDPR; Regulation (2024/679)); the UK General Data Protection Regulation; the Canadian Privacy Act and the Personal Information Protection and Electronic Documents Act (PIPEDA)[42]; and the Brazilian General Data Protection Law (LGPD)). They are also increasingly included under AI-specific regimes (e.g. the EU AI Act). Nevertheless, explicit ERT-specific rules are only slowly emerging, with the EU currently leading through its AI Act (2024).[43]

It is, however, likely that countries already working on new AI regulation, such as Canada, the UK, Brazil, and Australia, will follow the EU's risk-based approach and develop similar regulatory frameworks, including some guidance on what constitutes, for instance, "high-risk" AI systems.

2.5.1. ERTs in the EU AI Act

The EU AI Act is the first major law to explicitly define and regulate emotion recognition. It defines an "emotion recognition system" as an AI system "for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data," listing

[42] Canada will soon adopt Bill C-27, which will include the Artificial Intelligence and Data Act (AIDA) and the Consumer Privacy Protection Act (CPPA), and will replace the previous PIPEDA.

[43] To briefly sketch out the picture, in 2016, the word *emotion* did not appear once in the EU GDPR. By 2021, it appeared 10 times in the draft version of the AI Act and 20 times in its revised and official 2024 version, suggesting that ERTs are slowly gaining regulatory attention.

examples such as happiness, sadness, anger, surprise, disgust, embarrassment, excitement, shame, contempt, satisfaction, and amusement (Recital 18).

Adopting a risk-based approach, the Act acknowledges "serious concerns about the scientific basis" of such systems (Recital 44) and, on that ground, classifies AI systems intended to identify or infer emotions or intentions from biometric data as presumptively "high-risk AI systems", while also prohibiting their use in workplaces and educational institutions (Recital 44).

Beyond the prohibitions, the Act also notes that technical inaccuracies in biometric systems (which, under the Act, include ERTs) can lead to discriminatory outcomes, particularly concerning age, ethnicity, race, sex, or disabilities (Recital 32). While this addresses biometric *identification*, which is not necessarily the case with the use of ERTs, it articulates the general fundamental-rights concerns that also motivate the limits placed on emotion recognition in sensitive contexts.

By classifying ERTs as high-risk technologies, the European Commission acknowledges that such systems fail to meet the levels of accuracy, reliability, accountability, and trust required to participate in high-stakes decisions in a democratic and fair society. However, mentions of, say, privacy, are rather scarce and constrained to the notion of "data privacy".[44]

2.5.2. Forecasting problems in the EU AI Act for ERT regulation

While generally welcomed as the first major attempt at proactive AI regulation, the EU AI Act is not a flawless framework (Veale & Zuiderveen Borgesius 2021; Laux et al. 2024). With respect to ERTs in particular, the Act acknowledges their technical limitations and associated risks, yet a closer examination reveals several regulatory gaps.

First, it is intended that providers of ERTs must comply with the obligations applicable to high-risk AI, but only *if* their system does fall under "high-risk classification" (Article 6(2)). Although a first reading of the Act might lead one to believe that ERTs are either prohibited or high-risk, Article 6(3) introduces a conditional derogation on that apparent binary classification: "An AI system […] shall not be considered high-risk where it does not pose a significant risk of harm to the health, safety, or fundamental rights of natural persons." Consequently, even though

[44] See Häuselmann et al. (2023) for an argument supporting the idea that "emotion data" should be protected under the Act's data privacy arrangements.

ERTs are presumptively high-risk, a provider may avoid the high-risk status if they convincingly demonstrate, through a documentation process, that the system does not pose a significant risk, in its intended context of use, to health, safety, or fundamental rights.

In addition to this derogation path, an exemption may be applied to otherwise prohibited ERTs. ERTs, for which deployment would be intended in the workplace and/or educational institutions are prohibited *except* when such systems are deployed for safety or medical purposes, meaning that some of the most intrusive forms of ERTs may still be authorised under these labels.

A further complication arises from a regulatory distinction between ERTs whose final aim is to identify *expressions* of emotion and those designed to *infer* an individual's emotional state from those expressions. The prohibition applies only to the latter, meaning that the former viz., identifying expressions of emotion, is permitted.[45] As McStay points out:

> Under the law, a call centre manager using emotional AI for monitoring could arguably discipline an employee if the AI says they sound grumpy on calls, just so long as there's no inference that they are, in fact, grumpy. "Anyone frankly could still use that kind of technology without making an explicit inference as to a person's inner emotions and make decisions that could impact them," McStay says. (Carter Miles 2024)

From a risk-based perspective, the rationale behind this distinction is understandable, since at least prima facie, the risks of only identifying expressions of emotions appear lower than the risks of inferring emotional states from those expressions. In practice, however, this very same distinction now creates space for problematic uses and risks incentivising, among other things, some linguistic gymnastics on the providers' part in an attempt to get an exemption from the high-risk or prohibition categorisation.

Finally, the prohibition on explicitly inferring the emotional states of individuals from their emotional expressions is grounded primarily in current (valid) concerns about the scientific unreliability of ERTs, with the Act specifically mentioning the ways in which serious technical limitations risk causing severe harms. However, this justification is normatively fragile. In its current form, the Act protects individuals not from the *existence* (and use) of ERTs, but from

[45] In doing so, the prohibition does not include physical states, such as pain or fatigue. It does not refer for example to systems used in detecting the state of fatigue of professional pilots or drivers for the purpose of preventing accidents. It does also not include the mere detection of readily apparent expressions, gestures or movements, unless they are used for identifying or inferring emotions. These expressions can be basic facial expressions such as a frown or a smile, or gestures such as the movement of hands, arms or head, or characteristics of a person's voice, for example a raised voice or whispering.

inaccurate or *biased* ERTs (Prégent 2025a). As technological progress continues and the global market's eagerness remains strong, it is entirely plausible that future multimodal, real-time systems will overcome many present technical challenges. Should technical performance improve, the legal protection may weaken, while concerns relating to emotional privacy, autonomy, and affective surveillance, would remain the same or even intensify.

In short, although the AI Act has without any doubt introduced the first head-on regulatory path on ERTs in Europe, the protection it affords is partial and potentially temporary. Given that it (i) creates an official route to circumvent the "high-risk" classification and thus the obligations that come with it, (ii) has explicitly embedded exemptions on the prohibition (safety and medical reasons), which authorise some of the most intrusive and problematic types of ERTs to be used under the label of "safety" or "medical" reasons, and (iii) fails to convincingly regulate the future use of fully functional ERTs, it leaves the door open to legitimate concerns about the actual scope of protection it offers.

Given this state of affairs, it is necessary to ask what consequences, if any, the use of prospective *fully functional* ERTs would have for individuals and society beyond bias and discrimination-based risks. This dissertation takes up that question. Its ultimate aim is to identify (normative) reasons that would justify prohibiting, limiting, or, conversely, permitting the use of ERTs; reasons not contingent on current technological shortcomings. Accordingly, for the present purpose of this thesis, and from this point forward, I depart from current existing technology and by *fully functional* ERTs, refer to systems that are assumed to be technologically trustworthy and capable of providing robust affective assessments regarding individuals' inner affective states.[46]

2.6. Conclusion

This chapter has provided a multidisciplinary overview of ERTs and situated the anticipatory claim of this dissertation within the current technological, social, and regulatory landscape. Section 2.1 traced the emergence of ERTs from affective computing and identified the conceptual assumptions that still underly dominant research trajectories. Section 2.2 examined the technological affordances and limits of current systems, showing that even the most advanced multimodal models remain to this day technically limited, facing important challenges.

[46] To be clear, at present, *no* ERTs qualify as "fully functional" in that sense.

Section 2.3 mapped the rapidly expanding market and the diverse domains in which ERTs are now deployed, demonstrating their multistability and the broad spectrum of interests driving their adoption. Section 2.4 reviewed existing scholarly and activist critiques, highlighting how technical unreliability, bias, and flawed inferential models have already produced social and political harms in real-world use. Section 2.5 analysed emerging regulatory responses, focusing on the EU AI Act, and revealed that although regulation is developing, existing frameworks primarily target technical risks and other related data-harms, leaving normative questions about fully functional ERTs unresolved.

The chapter has showed that, while currently limited, it is unlikely that the challenges they face will halt the progress or deployment of ERTs. On the contrary, market incentives, technological advances, and political interests are converging towards increasingly intrusive and multimodal systems. At the same time, the regulatory landscape remains narrowly framed around technical reliability and data protection. This asymmetry reveals a deeper conceptual gap. Should near-future developments yield fully functional ERTs, there is no current answer regarding (i) the desirability and legitimacy of their use, (ii) what forms of social disruption they would generate, and (iii) whether their deployment is compatible with key democratic and relational values.

For these reasons, a philosophical account of the nature and functions of emotions and emotional expressions is needed to later assess the ethical stakes of preying on the affective life through technological means. While the thesis cannot, on its own, cover all the potential angles of future ethical risks, it aims to contribute to an in-depth investigation of the risks associated with privacy. The next two chapters lay the conceptual foundation for this task. Chapter 3 develops a philosophical analysis of emotions as intentional, socially embedded, and normatively significant phenomena. Chapter 4 offers an account of emotional expressions as interpersonal communicative signals that structure and sustain social life. These chapters provide the background needed for the ethical analysis developed in Chapters 5 and 6, which evaluates whether fully functional ERTs warrant the introduction of a new privacy-related concept and a corresponding moral right.