



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).

Emotional Privacy

A Philosophical Investigation of Emotion Recognition Technology

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof. dr. S. de Rijcke,

volgens besluit van het college voor promoties

te verdedigen op woensdag 2 september 2026

klokke 11.30 uur

door

Alexandra Prigent

geboren te Longueuil, Canada

in 1996

Promotor: Prof. dr. Dorota Mokrosinska
Copromotor: Prof. dr. James McAllister

Promotiecommissie:
Prof. dr. Annemie Halsema
Prof. dr. Sabine Roeser (Technische Universiteit Delft)
Dr. Jan Sleutels
Dr. Marjolein Lanzing (Universiteit van Amsterdam)

This dissertation is funded by The Social Sciences and Humanities Research Council of
Canada (SSHRC), grant #752-2021-1281.

Publication details

Emotional Privacy

A Philosophical Investigation of Emotion Recognition Technology

Alexandra Prigent

This dissertation was prepared at the Institute of Philosophy, Leiden University.

This dissertation was funded by the Social Sciences and Humanities Research Council of Canada (SSHRC), grant no. 752-2021-1281.

© 2026 Alexandra Prigent

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means without prior written permission of the author. Parts of this dissertation have previously been published in revised form as indicated in the respective chapters. Copyright in those contributions remains with the author and/or the original publishers, as applicable

Acknowledgements

For a long time, I treated the acknowledgement section as something to skip. Like many impatient readers, I wanted to go straight to the *vif du sujet*. This PhD changed that. Now, whenever I start a book, I find myself reading the acknowledgement section first. If nothing else, it gives a glimpse of the people, places, and moments that made the work possible. So here comes mine.

First, I want to express my deepest gratitude to my supervisors, Professor Dorota Mokrosinska and Professor James McAllister, for supporting me all these years and guiding me through the many preliminary and often confused thoughts, ideas, and drafts that I presented to you over the years. I am particularly grateful, and forever in debt, to Professor Mokrosinska, for the countless hours she spent reflecting on, commenting on, and revising my work. But a supervisor's work, of course, is much more than that. While the length of the thesis itself would probably not be long enough to recount every instance that marked this journey, one in particular had a remarkable impact on me. At the very beginning of my PhD, when I shared my desire to leave Canada and move to the Netherlands, she both encouraged me and warned me about the challenges and hard times that were inevitably to arise in my academic and personal life. And they indeed came. But whenever these hard times arrived, I would think about that conversation we had, and it helped me push through the hiccups of life and moments of self-doubts that are part of such journeys.

I am also immensely thankful for the support I received from my family. Merci maman et papa, pour le soutien sans faille que vous m'avez apporté toutes ces années, du cégep au doctorat, de Saint-Jean-sur-Richelieu à Leyde. Merci de m'avoir soutenu dans les hauts comme dans les bas et à travers vos propres doutes et craintes sur mon futur. Je n'y serais certainement jamais arrivée sans vous. Une mention aussi à mon frère, Guillaume, qui a toujours été pour moi une source de force et d'encouragement, et dont l'épanouissement des dernières années me remplit de fierté.

But this thesis would not have been the one it is without my partner, Jakob. Words cannot express my gratitude for his constant presence and reassurance over the past few years. If someone had told me that I would meet a person in Oxford who would change my life so drastically, I probably would have not believed them. None of it would have been the same without you by my side and I am privileged to walk this earth with you. Danke für alles, mein Schatz.

Lastly, because I think any project to which one devotes years of one's life, especially in such an isolating writing process, inevitably leads one to form special bonds with things that are not human, I have to give a special mention to the various places that shaped my academic journey by welcoming me, my computer, and my constant frowning. These include the Leiden Hortus Botanicus and the Van der Werffpark with its magnificent weeping willows, UT Austin's Perry-Castañeda Library and Taco Joint restaurant, and the balcony of my apartment facing southwest in Groningen.

Table of Content

<i>Emotional Privacy: A Philosophical Investigation of Emotion Recognition Technology</i>	ii
Acknowledgements	v
List of Figures.....	x
List of Abbreviations	x
1. General introduction	1
1.1. Preamble	1
1.2. Assumptions, approaches, and methodological choices	2
1.2.1. The reflexive process behind the dissertation's structure	2
1.2.2. Methodological choices across the three general parts	5
1.2.3. Transdisciplinary terminology and conceptual trade-offs	9
1.3. Overview of the chapters	9
<i>Part One: The Technology</i>	14
2. The prospect of 'fully functional' ERTs: A social, technological, and political overview	15
2.1. Introduction	15
2.2. ERTs' historical development.....	18
2.2.1. Face-based ERTs	18
2.2.2. Technological progress: From unimodal to multimodal ERTs	22
2.2.3. Technical challenges: Noise, indeterminacy, and the absence of ground truth	26
2.3. ERTs on the market: From technical artefacts to embodied emotional AI	29
2.3.1. Standard ERTs.....	30
2.3.2. Self-tracking apps	31
2.3.3. Aggression detector.....	32
2.3.4. (Semi-)embodied affective AI: From chatbots to robot companions	32
2.3.5. Border control & interest for lie detection	35
2.4. Biases, criticisms, and anti-sentiment towards ERTs on the market	38
2.5. A look at the current international regulation of ERTs	42
2.5.1. ERTs in the EU AI Act.....	42
2.5.2. Forecasting problems in the EU AI Act for ERT regulation	43
2.6. Conclusion	45
<i>Part Two: The Emotional Life</i>	47
3. Emotions: On their nature, role, functions, and relation to expressions	48
3.1. Introduction	48
3.2. Emotions	49
3.3. Emotions and morality	50
3.3.1. Emotions as a threat to morality: Moral rationalism.....	51
3.3.2. Emotions as essential to morality: Virtue ethics and sentimentalism	54
3.3.3. Contemporary accounts of emotions in moral reasoning.....	57
3.4. Integrative perspective on emotions	59
3.4.1. The biological/evolutionary perspective	59
3.4.2. The cognitivist perspective	62

3.4.3. The social constructionist perspective	65
3.5. Adoption of a function-first integrative perspective on emotions.....	66
3.5.1. Transdisciplinary agreements on their nature, role, and functions	67
3.5.2. Emotions' action tendencies	70
3.5.3. Emotions and physiological changes.....	72
3.5.4. Emotions as performative acts	75
3.6. Conclusion	76
4. Emotional expressions in technosocial contexts: Rethinking functions and modes of recognition.....	79
4.1. Introduction	79
4.2. Theoretical grounding of emotional expressions	80
4.2.1. Core properties of emotional expressions	80
4.2.2. Causes of emotional expressions	83
4.2.3. Spectrum of emotional expressions	85
4.2.4. Debates and ontological commitments.....	86
4.3. Affective communication channels	91
4.3.1. Emotional expressions' functions in affective communication channels.....	91
4.3.2. Meaningfulness and reliability of emotional expressions.....	96
4.4. A revised design-based account of emotional expressions.....	99
4.4.1. Emotional expressions' epistemic functions for humans and technology	100
4.4.2. Potential challenges for the proposed view.....	105
4.5. Conclusion	106
<i>Part Three: The (Ethical) Analysis</i>	109
5. Privacy and emotional life: Re-engineering the concept of emotional privacy	110
5.1. Introduction	110
5.2. Privacy.....	111
5.2.1. Privacy and emotional life	111
5.2.2. Privacy and technology.....	113
5.2.3. (Informational) privacy scholarship.....	114
5.3. Emotional privacy	121
5.3.1. The case of mental privacy as a precedent to emotional privacy.....	123
5.3.2. Anchoring emotional privacy in the social perspective	126
5.3.3. Defining emotional privacy.....	130
5.4. Emotional privacy in affective communication channels	133
5.4.1. Expressive conventions and affective leakages	134
5.4.2. Privacy mechanisms.....	135
5.4.3. Disclosure privacy mechanisms	137
5.4.4. Concealment privacy mechanisms.....	137
5.5. The functional value of emotional privacy	140
5.5.1. Preserving self-presentation	140
5.5.2. Managing social role plurality and ensuring conformity.....	142
5.5.3. Enabling social regulation and social cohesion	144
5.6. Conclusion	145
6. Emotional privacy as a moral right? Normative foundations and technological implications	147
6.1. Introduction	147

6.2. Formulating a normative account of emotional privacy	148
6.2.1. Human vs technology-based emotion recognition	150
6.3. Illustrating emotional privacy losses through technomoral scenarios	156
6.3.1. Scenario I: Alfred's self-monitoring.....	158
6.3.2. Scenario II: Marie's hospital policy.....	159
6.3.3. Scenario III: The minister's public apology	160
6.4. Moral right to emotional privacy protects a distinct kind of interests.....	161
6.4.1. Variants of Control-based accounts	162
6.4.2. Variants of Access-based accounts	166
6.4.3. Contextual Integrity	171
6.5. Assessing emotional privacy's moral weight in future societal trade-offs.....	174
6.5.1. Substantiating the CI moral part of the heuristic.....	176
6.5.2. Evaluating the moral weight of emotional privacy on the HLEG ALTAI framework.....	178
6.5.3. Emotional privacy and the question of consent	182
6.6. Conclusion	185
7. General conclusion	188
Bibliography	193
English Summary	215
Curriculum Vitae	219

List of Figures

Figure 1. Simplified Face-Based ERT Process	21
Figure 2. Spectrum of Nonverbal EEs	86
Figure 3. Standard Affective Communication Channel	92

List of Abbreviations

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
CI	Contextual Integrity
CMU-MOSEI	Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity
DL	Deep Learning
DSS	Decision Support System
DSV	Deep Self View
EASI	Emotion as Social Information
ECG	Electrocardiogram
EE	Emotional Expression
EEG	Electroencephalogram
ERT	Emotion Recognition Technology
FACS	Facial Action Coding System
GDPR	General Data Protection Regulation
HLEG	High-Level Expert Group
HR	Human Resources
IEMOCAP	Interactive Emotional Dyadic Motion Capture

LGPD	Brazilian General Data Protection Law (Lei Geral de Proteção de Dados Pessoais)
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PIPEDA	Personal Information Protection and Electronic Documents Act
RBS	Rule-Based Expert System
TAP	Theory of Affective Pragmatics
VR	Virtual Reality

1. General introduction

This opening chapter prepares the conceptual and methodological ground for the dissertation. After a preamble that frames the central philosophical problem (§1.1), I discuss the general assumptions underlying this work, the approaches chosen, the methodological choices (§1.2), and I conclude with an overview of the chapters of the dissertation (§1.3).

1.1. Preamble

Suppose governments, private companies, strangers or close relatives could have access to your inner emotions via a particular technology. Suppose law enforcement could use it for surveillance, policymaking, and to tailor public information campaigns; private companies could use it for hiring, insuring, or marketing purposes; strangers and close relatives could use it as a way to understand you better. Dystopia? Science fiction? Not exactly.

In our world, where technosocial contexts seem to be our new *zeitgeist*, peering into people's inner emotional life is something that an emergent technology has been experimenting with. Emotion recognition technology (ERT) uses AI algorithms to identify, collect, and process data extracted from the human face, body, movements, voice, gait, brain activity, and more, in the hope of making relevant inferences about emotional reactions, actions, motivations, and intentions. ERTs and related affective-computing systems have already been implemented across diverse contexts, including targeted advertising that analyses faces for negative and positive expressions (Affectiva; iMotions); emotional contagion on diverse social platforms (Kramer, Guillory, & Hancock 2014); wearables that track moods (Empatica); cars that detect emotions in drivers (Smart Eye; Affectiva Automotive AI); and glasses that detect people's emotions throughout a conversation (Emteq Labs). There is keen interest in using emotion-tracking software in policymaking, commerce, healthcare, and robotics. The field is moving swiftly towards real-time, multimodal, context-sensitive systems capable of making increasingly fine-grained inferences about affective and motivational states.

However, despite their growing deployment, ERTs remain highly controversial, with critics raising concerns about their lack of scientific legitimacy, ethical implications, and potential to reinforce discriminatory practices (Barrett et al. 2019; Tcherkassof & Dupré 2021; Crawford et al. 2019).

Yet while most attention is directed at the problems created by their current limitations, a crucial problem remains underexplored, namely how our social and normative landscapes would change if ERTs actually worked as intended. To my knowledge, no anticipatory ethical assessments have focused specifically on the case of fully functional ERTs. This dissertation aims to fill this critical gap. Rather than assessing current ERTs as they stand, this dissertation examines the social implications that would arise if ERTs were to be fully functional. Thus, this thesis investigates a central and timely question:

If ERTs were to function as intended, would this affect social practices, relations, and normative expectations, and if so, to what extent?

1.2. Assumptions, approaches, and methodological choices

This thesis contributes to the broader endeavour in the philosophy of technology that examines how technological and digital innovations shape human experience, social structures, and our understanding of the world. Its central aim is to anticipate the societal and ethical implications of a new technology, namely fully functional ERTs, in order to clarify what is at stake and what kinds of risks and normative challenges their use may introduce. The thesis is anchored in philosophy of technology given its central question, but the resources it mobilises bring together philosophy of technology, philosophy of emotion, and privacy theory.

The structure of the thesis is articulated in three distinct parts, corresponding to each of the broad topics of (i) technology (where I focus on ERTs), (ii) emotional life (where I focus on emotions and emotional expressions), and (iii) the ethical analysis (where I focus on emotional privacy). I dedicate one chapter to the topic of technology, two chapters to the topic of emotional life, and two chapters to the ethical analysis, centred on privacy. I explain the methodology undertaken in each of these parts in §1.2.2, but before I do that, I will say a few words about the way this structure came about within my personal reflection.

1.2.1. The reflexive process behind the dissertation's structure

One of the great things about philosophy is that it allows one to escape the methodological iron grip in which other disciplines are often trapped. Yet the danger associated with that freedom is to veer too far and lose touch with reality. While some, like Hermann Hesse, would probably argue that there is nothing wrong with such detachment, I have always been motivated by the

idea that philosophical work can make a genuine impact in society. Thus, one of the central aims of this thesis is to exploit the methodological freedom of philosophy to move beyond debates that focus narrowly on the present technical capacities of ERTs, while remaining firmly grounded in real-world concerns.

I began the thesis by mapping the technical landscape of ERTs. I had already examined existing (and thus inevitably flawed) systems in my master's thesis (Prégent 2021), but for this dissertation I conducted a second, extensive literature review. This revealed a striking series of recent breakthroughs in affective computing, which in turn seem to fuel the continued eagerness of both private and governmental actors to develop and deploy ERTs. While trying to understand the most likely path that research would take towards this development, I realised that the field of affective computing, despite its apparently strict methodologies, was drawing on anything and everything it could find in the study of emotions and emotional expressions to ground their own research. This meant often taking work out of context, overgeneralising claims made, and even relying on work that had been heavily criticised or disproven. This led me to question the true object of ERTs, which we often refer to as “emotions” or “inner affective states”. This is the moment I turned to the second part of this thesis.

I wanted to arrive at an understanding of emotions and emotional expressions that would be both sufficiently general and adequately tuned to the ambitions of ERTs. During my investigation, I realised that many theories which appeared, at least a priori, to be in radical opposition to one another (mostly due to their claims regarding the nature of emotions and emotional expressions) were in fact agreeing on many function-related aspects of emotions and emotional expressions. This motivated the integrative commitment I ultimately adopted, which I approached from a function-first angle. After spending more hours than I want to admit on this topic, I began reflecting on the fact that most research on ERTs focuses directly on the phenomenon that ERTs aspire to reach (viz. emotional states), but very few pause midway to consider the phenomenon that they *actually* reach (viz. emotional expressions). This made me realise that although emotional expressions are thought to be the expressive *component* of a given emotion, in social life, emotional expressions are much more than that and their causes do not limit themselves to emotional states but also include physical states and, more importantly for this research, social motives. As many researchers on ERTs were integrating under “emotion recognition” data related to stress, anxiety, but also dissatisfaction, and even things like “politeness”, I pursued the integrative approach and looked at transdisciplinary work on

emotional expressions. My aim was to find a way to define emotional expressions in a way that would allow both humans and technologies to draw their own inferences about an emoter's emotional states. To that end, I hypothesised and pursued the idea of a bifunctionality of emotional expressions. This bifunctionality offers an explanation on how 'meaningful' information can be captured through two different epistemic routes. From there, I began reflecting on the potential disruption that fully functional ERTs would pose to the functions that emotional expressions serve in social life. This was the turning point where I started reflecting on the third part of the dissertation.

Given these different epistemic routes, I wondered whether the information that was accessed through the technological route could put pressure on, or even severely disrupt, the current social structure of life, which relies heavily on the epistemic boundaries inherent to the "human route" of affective recognition and interpretation. This would imply that ERTs could provide information that were previously not available to ordinary perceivers under similar circumstances. Because emotional expressions are a type of information that is inherently social, I turned to both informational and social privacy, which significantly overlap in contemporary work. Using the methodological tools of conceptual engineering, I re-engineered the concept of emotional privacy, treated here as a subtype of informational privacy. It is a subtype of informational privacy because its object remains a kind of information, namely affective information. Emotional privacy is thus introduced as a conceptual addition to the privacy cluster, one that allows us to grasp more fully the normative (privacy-related) stakes of technologies such as ERTs. I argue that ERTs have a particularly disruptive potential with respect to the functions that emotional expressions serve in social life. This disruption deserves special attention not only because ERTs target emotional states that we would often prefer to keep private, but also because detecting such states risks undermining the core functions of voluntarily produced, and often conventionally crafted, emotional expressions on which social life relies.

As a result, another central aim of this thesis is to encourage future discussions on affective AI and privacy to adopt a broader conception of "emotional life" and to recognise the disruptive potential of ERTs for the structure of social life.

I now briefly sketch the main methodological choices underlying each of the three parts.

1.2.2. Methodological choices across the three general parts

Part one: The technology

Ethical reflection on emerging technologies faces a familiar methodological challenge known as the Collingridge dilemma (Collingridge 1980). It posits that, with regard to new and emergent technologies, we must choose between two evaluative stances: we must either (i) evaluate a technology in its early stage, or (ii) evaluate it later in its development. If we choose the former, we presumably still have reasonable control over its design but likely lack sufficient knowledge of its social and moral implications. If we choose the latter, and evaluate it later in its development, these social and moral implications have likely become clearer through widespread adoption, yet at this stage the technology is often so embedded and entangled in the social, economic, or political fabric that redirecting or constraining it becomes extremely difficult.

Sociotechnical experimentation and responsible innovation, for instance, focus on this latter stage. They aim to monitor technologies and adjust governance practices once systems are already in place (van de Poel 2013; Owen, Bessant, & Heintz 2013). However, such approaches inevitably remain backward-looking, responding to harms or unintended consequences only after the technology has already reshaped social practices. As a result, they are ill-suited to prevent harms that a more forward-looking evaluative approach, such as anticipatory ethics, might have identified long before actual deployment.

Anticipatory ethics focuses on the former stage and seeks to examine the possible future impacts of emerging technologies while intervention is still possible, bringing ethical analysis to the forefront of technological design and development (Swierstra & Rip 2007; Brey 2012; Vallor 2016).¹ However, working this early in a technology's life cycle means that anticipatory approaches risk drifting into purely speculative scenarios that bear little relation to real technological trajectories (Kudina and Verbeek 2019; Nordmann 2007). The main concern is that these analyses risk developing unrealistic scenarios that lead them to focus on theoretical questions that have little to no practical significance.² A responsible anticipatory methodology

¹ It might be worth noting that a third approach to the Collingridge dilemma has recently been proposed by Kudina and Verbeek (2019). Kudina and Verbeek propose the “ethics from within” approach, a type of technological mediation that focuses on the problem of value dynamism in new and emergent technologies.

² In other words, to borrow a popular expression from Québec, they risk ending up “pelleter des nuages” (shovelling clouds).

thus requires that its analysis be empirically grounded whenever possible, so as to understand and not depart from a technology's trajectory.

I believe that neither of these approaches is sufficient in itself and that both types of approaches are important for responsible innovation. Having said that, this thesis adopts and contributes to the former stage of the development of a new technology, following anticipatory ethics. Consequently, in the remainder of this paragraph I now briefly explain how I have sought to mitigate the risks that come with anticipatory approaches, particularly the risk of drifting away from real-world issues.

While the ultimate aim of this thesis is to contribute to mapping the ethical risks associated with the possibility that ERTs work as intended, I pre-empt some of the main methodological pitfalls of anticipatory ethics by drawing the prospective scenario of fully functional ERTs on how present developments appear to be moving. This entails a detailed reconstruction of their *current* sociotechnical trajectory across technical, social, economic, and regulatory contexts. In this way, I aim to ensure that the framing of the ethical problem is empirically anchored rather than completely detached from actual developments. This provides a plausible hypothetical scenario of fully functional ERTs grounded in existing research, commercial applications, and regulatory debates. Additionally, I was careful to not ground the ethical analysis too tightly on the empirical predictions regarding how, exactly, such technology would work. Even though I do make some assumptions throughout the thesis, the ethical analysis and its normative conclusions are only grounded on the high-level assumption that these AI systems work.

Part two: The emotional life

Part two turns to the object of ERTs, namely the emotional life. The methodological stance in these chapters is transdisciplinary and function-first. I bring together work from the philosophy of emotion, moral and social psychology, neuroscience, anthropology, and affective science in order to develop a joint account of emotions and emotional expressions. Rather than treating these disciplines as competing sources of authority, I use them as complementary lenses that illuminate different dimensions of affective life.

Chapter 3 adopts what I call an integrative perspective on emotions. This is also a function-first methodological lens that explains emotions in terms of the roles they play in perception, practical reasoning, social coordination, and normative life. It integrates insights from biological

and evolutionary approaches, cognitivist and evaluative approaches, and social constructionist approaches to make sense of how emotions structure attention, confer salience, guide action, and sustain interpersonal relations. The chapter also employs historical and conceptual analysis of major traditions in moral philosophy (e.g. moral rationalism, virtue ethics, sentimentalism, and contemporary hybrid accounts) in order to situate emotions within debates about moral judgement and motivation. The aim is not to resolve foundational debates about the nature of emotion, but to extract a multi-level functional picture that can be used in Chapter 4 and the later ethical assessment of ERTs.

Chapter 4 continues this function-first approach by focusing on emotional expressions and affective communication. Methodologically, it develops a conceptual overview of emotional expressions, analysing their properties, causes, and varieties, and it proposes to see emotional expressions along a spectrum, which includes natural, semi-natural, semi-conventional, and conventional expressions. It then endorses a probabilistic and pragmatic view of the informational role of emotional expressions. On this view, expressions are treated as a reliable though defeasible source of information in virtue of probabilistic regularities that are shaped by context, norms, and background knowledge. Building on this, the chapter advances a modified design-type account of emotional expressions. This allows me to distinguish the functional roles emotional expressions play within human and ERT-based recognition while preserving a unified account of emotional expressions.

Across both chapters, the analysis is structured by a distinction between intrapersonal and interpersonal levels. Chapter 3 focuses on the intrapersonal dimension of emotions as states that shape deliberation and action readiness within the individual. Chapter 4 focuses on the interpersonal dimension of emotional expressions as elements of affective communication that coordinate behaviour, regulate social expectations, and support mutual understanding. Taken together, these methodological choices yield an integrative and functionalist account of emotional life that prepares the conceptual ground for the subsequent ethical analysis of ERTs from a privacy perspective.

Part three: The (ethical) analysis

Once the technological and affective foundations are in place, the third part of the thesis turns to the ethical analysis proper. ERTs raise many ethical concerns, from creating new possibilities for manipulation and discrimination to posing threats to privacy and democratic practices. In

this third part, however, I intentionally restrict the ethical analysis to the informational privacy dimension. I do so because it is within the informational privacy paradigm that fully functional ERTs appear to introduce a new and structurally distinct kind of harm, namely the defeat of expressive discretion.

The guiding question for this part is not simply whether ERTs are permissible, but whether existing informational privacy concepts and corresponding moral rights are suitably equipped to register and protect against the kinds of harms that arise when affective life becomes fully legible to technological systems. The ethical analysis deliberately brackets issues such as bias and unreliability so as to be able to isolate the privacy-related harms that stem from the very possibility of continuous affective legibility. Methodologically, I divide this task into a descriptive-conceptual stage (Chapter 5) and a normative stage (Chapter 6). In Chapter 5, I adopt a problem-driven, cluster-based form of conceptual engineering. Rather than treating privacy as a fixed concept, I treat it as a conceptual cluster that has historically been adapted in response to new technologies. Using this cluster-based approach, I identify a tech-induced conceptual gap. I thus proceed by asking what conceptual resources would be needed to fill this gap. I take inspiration from the recent introduction of mental privacy in response to neurotechnologies and treat it as a useful case study for this kind of conceptual adaptation. I then draw on it in developing emotional privacy. Emotional privacy is introduced as an additional subconcept to the informational privacy cluster. Once this concept is in place, I turn to the social and moral stakes raised by technologies that target emotional expressions and affective communication and the normative task of determining whether, and in what sense, emotional privacy should warrant normative protection.

To answer that latter question, the normative stage relies on technomoral scenarios, structured across individual, interpersonal, and societal levels of interaction, to examine how the defeat of expressive discretion manifests in different social roles and institutional settings. These scenarios then feed into Contextual Integrity and ALTAI-style ethical assessments, which are used not to derive a complete ethical verdict on ERTs, but to test whether recognising a moral right to emotional privacy can genuinely change the overall verdict of an ethical assessment on fully functional ERTs.

1.2.3. Transdisciplinary terminology and conceptual trade-offs

While being firmly anchored within the philosophical field, this thesis also relies on transdisciplinary work from psychology, anthropology, affective sciences, neuroscience, and computer science. This has enabled me to develop a rich, detailed, and nuanced understanding of the concepts and phenomena at stake from diverse points of view, but it has also brought specific challenges.

The main drawback of relying on such diverse literatures is that it has forced some terminological “trade-offs”. Philosophically meaning-laden words, such as “truth”, “knowledge”, “emotions”, and “happiness”, have, at times, had to be relaxed or partially bracketed in favour of some of their counterparts in other disciplines. Without this adjustment, it would have been impossible to engage seriously with these fields’ own operative understandings of the concepts they invoke. I therefore engage with these disciplines’ understandings of our shared concepts, subjecting their interpretations and the theoretical commitments they embody to philosophical analysis and, where needed, criticise them (for example, the tendency in affective computing and computer science to over-reduce, or to include too wide a range of phenomena among those that count as an “emotion”).

1.3. Overview of the chapters

Below I offer an overview of the five chapters that make up the body of this dissertation.

Chapter 2 provides a comprehensive overview of ERTs in their social, technological, and political contexts, establishing the groundwork for the thesis’s normative inquiry centred around the possibility of “fully functional ERTs”. It first traces the historical development of affective computing, from its roots in cognitivism to the emergence of multimodal ERTs, showing how technological progress has moved beyond static, face-based systems towards increasingly context-sensitive and adaptive models. It then surveys the current ERT market covering embodied agents, self-tracking apps, passive surveillance tools, border-control lie detectors, and aggression detection systems, highlighting their multiple functions across commercial, healthcare, educational, and security domains. It also examines technical limitations and pinpoints key challenges, such as noise, indeterminacy, and the absence of ground truth, while acknowledging ongoing advancements that could soon yield “fully functional” ERTs capable of informing high-stakes decisions. The chapter also reviews critical scholarship and political pushback, focusing on biases, the contested scientific foundations of ERTs, and concerns about

their discriminatory potential. Finally, it analyses current international regulation, focusing on the EU AI Act's treatment of ERTs, identifying two key shortcomings: its narrow focus on technical unreliability and its failure to provide a normative justification for restricting even highly accurate future systems. The chapter concludes by positioning the thesis as an anticipatory ethical analysis aimed at addressing this gap.

Chapter 3 turns to the phenomenon of emotion itself, examining its nature, its role in (moral) decision-making, judgement, and action, and its functioning at the intrapersonal level. It begins by distinguishing emotions from neighbouring intentional states such as beliefs, desires, and judgements, and emphasises their multifaceted nature as evaluative, motivational, and phenomenological phenomena embedded in social life. The chapter then examines the historical debate on emotions in morality: while moral rationalists (Plato, Kant) sought to exclude or control emotion, their indispensability for moral reasoning emerged in contrast; virtue ethics (Aristotle) and sentimentalism (Hume) instead foregrounded their constitutive role in moral perception and motivation. Contemporary theories (Nussbaum, Tappolet, Greenspan, Ben-Ze'ev, and Blum) further affirm emotions as epistemically significant and morally salient.

Building on this, the chapter surveys three major perspectives, namely, biological/evolutionary, cognitivist, and social constructionist, treating these perspectives as complementary lenses rather than rival theories. From their points of convergence, it extracts a picture of emotions as embodied, evaluative, culturally shaped, and socially embedded processes that are adaptive, normative, and dynamic. Yet the chapter also stresses that to speak of “the functions of emotions” without further specification can be misleading as different emotions fulfil different roles, and even a single emotion type can serve distinct functions depending on context, background knowledge, and internal tensions. On this basis, the chapter adopts an integrative perspective on emotions that understands emotions through the roles they play in structuring perception, orienting attention, guiding action, and mediating interpersonal relations. This approach underpins the thesis's later normative evaluation of ERTs, highlighting how tracking affective life also partly tracks one's decisions, values, and action tendencies. If emotions can be measured and understood with sufficient precision, their instrumentalization becomes a practical possibility. Yet a full account of emotional life also requires examining its outward dimension, which is the focus of the next chapter.

Chapter 4 shifts the analysis from the intrapersonal to the interpersonal domain by focusing on emotional expressions (EEs) as the primary medium through which affective states become socially legible. It argues that for ERTs to pose a substantive normative challenge, two conditions must hold: first, nonverbal affective communication must play a central role in everyday interaction; second, EEs must function as reliable sources of information on which emoters and perceivers depend to sustain other vital social functions. This chapter shows that these two conditions are met as EEs are primary elements of the affective communication channels on which interactions depend, sustaining coordination, managing expectations, and regulating social life. The chapter begins by providing a theoretical grounding of EEs, identifying their core properties, causal sources (including bodily states, emotions, and social motives), and mapping a spectrum from natural through semi-natural and semi-conventional to fully conventional EEs.

It then situates EEs within affective communication channels, adopting a probabilistic-pragmatic approach according to which emotional expressions convey meaningful and reliable affective information even though perceivers need to partly rely on social norms and background knowledge to make sense of it. Building on this, the chapter develops a functionalist account of EEs that positions itself between both reductive naturalism and pure perceiver-dependent accounts. In technosocial contexts, this framework underwrites a bifunctionality of EEs: for humans, EEs operate as communicative acts that structure and regulate social life; for ERTs, they become diagnostic data points for inferring inner states. In doing so, the chapter clarifies how EEs can, through two different epistemic routes, simultaneously sustain human social practices and feed into automated recognition systems, setting the stage for the later privacy-focused ethical analysis.

Chapter 5 develops the conceptual and theoretical foundations for the concept of emotional privacy, positioning it as a constitutive condition of social life rather than a mere shield from it. Building on the functionalist account of emotions and emotional expressions developed in Chapters 3 and 4, it reconceptualises privacy in affective life as the condition under which emotional expressions can fulfil their social roles. It first traces the historical evolution of privacy from individualistic to relational and social accounts and surveys contemporary informational privacy scholarship in order to show how privacy concepts have repeatedly been adapted in response to technological change. Using the recent emergence of ‘mental privacy’ in response to neurotechnologies as a precedent, the chapter argues that fully functional ERTs generate a

similar tech-induced conceptual gap in the affective domain that the informational privacy cluster currently cannot fill.

On this basis, Chapter 5 introduces and defines emotional privacy as a subconcept within the informational privacy cluster. It distinguishes between *actual emotional privacy*, which obtains when an agent enjoys expressive control in particular situations, and *standing emotional privacy*, which obtains when an agent retains the expressive ability to manage affective legibility over time. This two-level structure captures the idea that when it comes to emotional life, privacy concerns not only particular episodes of access but the background ability to modulate what one's expressions count as in context. The chapter thus argues that emotional privacy aligns with the Control-based perspective of informational privacy, even though it departs from standard Control-based accounts by showing that the obtention of standing emotional privacy, and not actual emotional privacy, is what matters for expressive discretion.

The remainder of the chapter examines how emotional privacy functions within affective communication channels. It first analyses expressive conventions and affective leakages, and then turns to privacy mechanisms, distinguishing between mechanisms of selective disclosure and mechanisms of concealment. These mechanisms operate by modulating the degree of affective opacity in given interactions, across audiences and contexts. The functional value of these mechanisms is then traced across three levels of social life: at the individual level, emotional privacy underwrites self-presentation; at the interpersonal level, it enables the management of role plurality and adherence to role-bound norms of expressiveness; and at the societal level, it supports social regulation and cohesion. The chapter concludes that (standing) emotional privacy is an integral part of the infrastructure of social life, and that its loss would disrupt not only individual interests but also the social functioning of affective communication. This grounds the normative inquiry undertaken in Chapter 6, which asks whether emotional privacy should warrant normative protection, and if yes, how recognising such a moral right would reshape the ethical assessment of ERTs.

Chapter 6 makes the normative turn of the thesis by developing and defending a moral right to emotional privacy and exploring its implications for the ethical assessment of fully functional ERTs. Building on the descriptive account in Chapter 5, it first articulates a normative account of emotional privacy and formulates conditions under which this right is realised or violated. On this view, emotional privacy is violated when technologies defeat individuals' expressive

discretion. The chapter then argues that this right cannot be reduced to, or fully captured by, existing normative frameworks, including Control-based and Access-based moral rights to informational privacy, and Contextual Integrity. This distinctiveness is further clarified through an analysis of the differences between human and technology-based emotion recognition, showing that fully functional ERTs restructure the affective and expressive fabric of social life rather than merely extending ordinary mindreading. To make the argument concrete, the chapter employs three technomoral scenarios, staged at the individual, interpersonal, and societal levels, which illustrate how the continuous legibility of affective life undermines practices of (i) self-presentation, (ii) role plurality and adherence to role-bound norms of expressiveness, and (iii) social regulation and cohesion. Finally, the chapter tests the proposed moral right by embedding it in Contextual Integrity and ALTAI-style ethical assessments, showing how recognising emotional privacy may change the moral evaluation in favour of privacy, particularly in trade-offs involving ERTs and other similar technologies. In doing so, Chapter 6 demonstrates that emotional privacy is a helpful addition to the informational privacy cluster when evaluating current and future affective technologies that pose a distinctive threat to individuals' expressive discretion.

Together, these chapters render intelligible the central claim of the thesis: that biases and discrimination caused by flawed ERTs only represent the tip of the iceberg when it comes to understanding and mapping the social and ethical impact of ERTs. Tech-induced change is not inherently problematic, but forecasting potential risks and challenges is essential to guide the development and use of these new technologies, as part of a collective effort to remain aware, reflective, and normatively informed regarding the changes that lie ahead. By combining an empirically informed anticipatory analysis of fully functional ERTs (part one), a function-first integrative account of emotions and emotional expressions (part two), and developing the concept of emotional privacy (part three), the dissertation contributes to contemporary debates in the philosophy of technology, the philosophy of emotion, and privacy theory, and offers a normative framework for future philosophical, ethical, and regulatory work on affective AI systems.

Part One:
The Technology

2. The prospect of ‘fully functional’ ERTs: A social, technological, and political overview³

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, Neurologica 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.” (Neurologica 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 Prigent (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 FER-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

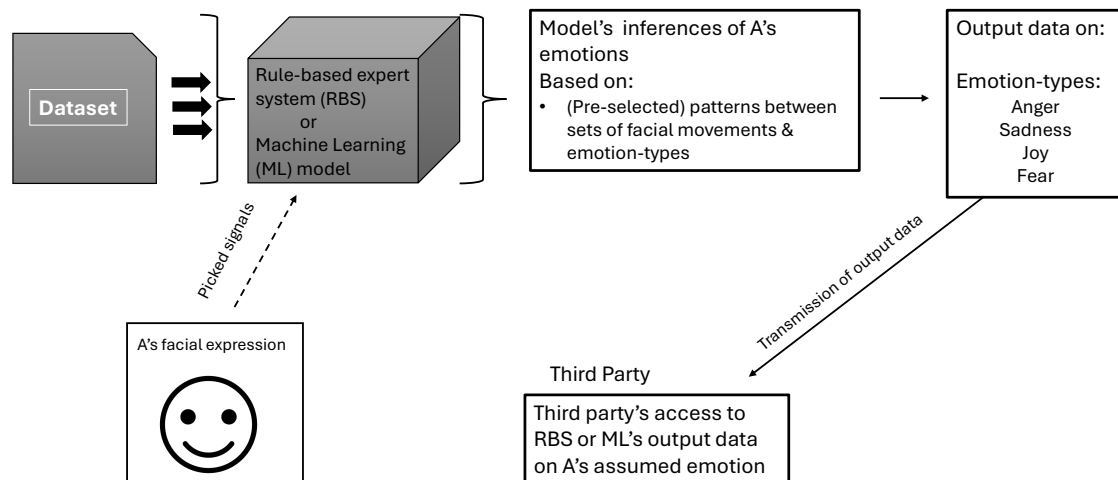
⁷ Although Ekman defended the idea of basic emotions and a one-to-one mapping between these emotions and specific facial expression patterns, his overall theory is more nuanced, allowing space for display rules and other social factors to alter or challenge this mapping.

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 Process⁹



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 datasets¹¹ 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 $\neq$ 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 Neurologica 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 systems²¹ 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 UK²⁹ 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 Robotics³² 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 (2016/679)); 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.

Part Two:
The Emotional Life

3. Emotions: On their nature, role, functions, and relation to expressions

3.1. Introduction

This thesis aims to identify the key ethical, social, and normative impacts involved in tracking affective life – moving beyond purely technical challenges. Given that affective life is a complex, multifaceted, and deeply social phenomenon, breaking down disciplinary silos in research on emotions and emotional expressions is essential. To gain a clearer understanding of their respective functions and the core elements of their interrelations, I draw on scholarship from multiple fields, including philosophy, psychology, affective computing, and affective neuroscience. This discussion establishes the necessary theoretical foundation regarding affective life, and lays the groundwork for the normative analysis presented in the last part of the thesis.

Accordingly, in this chapter and the next I examine the intricate relationships between emotion, emotional expression, and emotion recognition. This chapter focuses on the phenomenon of emotion, analysing its nature, its role in (moral) decision-making, judgements, and actions and its main functionalities at the intrapersonal level. The chapter seeks to combine normative insights from philosophy, especially work on human morality and the nature of emotions, with empirical findings from (social) psychology and affective neuroscience.

Section 3.2 briefly introduces the notion of emotions and disentangles the concept from other intentional mental states, namely desires, beliefs, and judgements. **Section 3.3** revisits canonical philosophical accounts of human reasoning and morality and analyses in particular their treatment of the relations between emotions and morality. This analysis underscores how the concepts of emotion, moral salience, and values are deeply intertwined and how emotions are an essential part of moral reasoning and moral actions – even within accounts that explicitly try to keep emotions separated from moral reasoning. **Section 3.4** expands the analysis by adopting an integrative perspective on emotions that include transdisciplinary insights structured along three broadly conceived perspectives: biological/evolutionary, cognitivist, and social constructionist. **Section 3.5** relies on key points of agreement between these perspectives, and then further turns to the relation between emotions and their expressions, focusing particularly on their action-tendencies, typically associated physiological changes, and performative potential to provide an integrative and function-first overview of emotions.

3.2. Emotions

In the effort to understand the nature, role, and functions of emotions, philosophy has, particularly over the past century, been increasingly joined by psychology, cognitive science, affective neuroscience, and, more recently, affective computing.⁴⁷ This has given rise to transdisciplinary contributions that have significantly expanded the scope of inquiry by shifting attention from the moral relevance of emotions to their broader significance in human life. This expanded lens allows for a richer appreciation of the complex roles emotions play not only in individual deliberation but also in social interaction, communication, and culture.

Emotions are highly complex phenomena that stand in close relation to other mental states, notably beliefs, judgements, and desires. Beliefs are typically recognised as cognitive states that represent the world as being a certain way and are evaluable for truth or falsity. Judgements are often taken as a subset of beliefs, marked by their normative commitment and evaluative character. Desires are conative states aimed at realising particular states of affairs and are often assessed in terms of prudential or moral coherence. Emotions, like these other states, possess intentional content – be it propositional, object-directed, or representational – and are typically evaluated in terms of their fittingness or appropriateness (i.e. whether the emotion suits its object in light of contextual and normative factors).

Emotions share significant features with beliefs, desires, and judgements, often blurring the boundaries between them. They are *belief-like* in their responsiveness to reasons and evidence, and in their potential revisability in light of new information (Greenspan 1988; Ben-Ze’ev 2000). They are *judgement-like* due to their evaluative character and their capacity to commit agents to certain stances or actions (Nussbaum 2001; Solomon 1976). Finally, they are *desire-like* in that they exert motivational force, prompting agents towards action (Goldie 2000; Hume [1739]2003). Some theorists go further, treating emotions as *subtypes* of beliefs, judgements, or desires (e.g. Hume [1739]2003), while others view them as hybrid states partly constituted by those elements (e.g. Greenspan 1988; Ben-Ze’ev 2000; Prinz 2004; Tappolet 2017).

Crucially, what helps disentangle emotions from their counterparts is that while they are all typically regarded as intentional states, emotions are characterised by a distinctive phenomenological dimension: there is something *it is like* to experience an emotion. This

⁴⁷ In this work, I will follow the terminology of Dukes et al. (2021) and refer to these various disciplines as ‘affective sciences’ when referring to them in the general sense.

experiential aspect distinguishes them from purely cognitive states and affirms their role in shaping our lived reality. Emotions are thus situated between affective, cognitive, and conative faculties. They are felt, evaluative, and motivational, capable of guiding perception, practical reasoning, and decision-making. Furthermore, emotions are profoundly interwoven with perceptual and social dimensions. They are central to interpersonal relationships and responsive to cultural and normative expectations, often shaped by socially embedded practices of expression, regulation, and interpretation. This theoretical diversity testifies to the conceptual richness of emotions and the tensions that arise in attempts to capture their nature.

3.3. Emotions and morality

Before turning to the transdisciplinary discussion on the complex topic that emotions represent, I first examine their relation to morality and the canonical works that first brought these concepts together. Much of the transdisciplinary research on emotions begins either by acknowledging the significant contributions and insights of these early accounts into emotional life or by critiquing their notable misconceptions of the phenomenon – both of which have profoundly shaped the ways in which emotions are conceptualised today. The nature and function of emotions, especially in relation to (moral) reasoning, have been central to philosophical inquiries into the human mind. Philosophers such as Plato, Aristotle, Hume, and Kant, among others, have laid the foundational work for exploring the role of emotions in morality. Primarily exploring this role within the broader context of developing ethical theories and conceptualising morality (Tappolet 2022), they interestingly took radically divergent positions on whether emotions are necessary for reasoning. While the work of Plato and Kant influenced later scholars – particularly within the ‘classical’ rationalist tradition – to view emotions as a threat to rationality and to reject them as potential core components of judgement (Solomon 1976), the work of Aristotle and Hume shaped competing traditions in virtue ethics and sentimentalism, emphasising the role of emotions in moral perception and as a key element of moral action.⁴⁸ Thus, engaging with both sides is necessary if we are to arrive at a more comprehensive understanding of the role and diverse functions of emotions in morality, and of how these traditions shaped later transdisciplinary studies.

⁴⁸ Interestingly, Plato’s tripartite division of the soul has been replicated in the early models of the mind in cognitive sciences, which continued to assume a strict division between cognition, conation, and affection (Hilgard 1980). It influenced early views in psychology and neurosciences, so much so that they rarely even questioned the division (Scherer 1984: 293). As noted by Frijda (1988), by Ellsworth and Scherer (2003), and by many other affective researchers, cognitive appraisal theories are also heavily influenced by philosophical reflections on emotions, particularly in the underlying role of pleasure and displeasure in different emotion-types such as happiness, anger, fear, and guilt.

3.3.1. Emotions as a threat to morality: Moral rationalism

Plato and Kant, who at least to some extent, appear to reject emotions from their conceptions of morality, particularly with regard to the possibility of making rational judgements, have at times conceded that emotions, or at least moral feelings, may play an instrumental role in moral reasoning. Their wariness of emotions seems to stem primarily from the belief that individuals have limited control over them, rendering emotions a potential obstacle to the attainment of rational judgement and moral action. Nonetheless, prominent interpretations of Plato's and Kant's theories have sought to reintegrate – at least partially – emotions into their respective frameworks of morality, arguing that their theories intrinsically require the endorsement of certain emotions to fulfil their aims. Precisely because of their (ultimately unsuccessful) ambitious attempt to demonstrate how motivations, moral actions, and moral reasoning might be achieved independently of emotions, their work reveals – perhaps more clearly than any other – the intrinsic interdependencies among emotions, motivations, actions, and (moral) reason.

Emotions in Plato's account of (moral) reasoning

Since the attainment of rational judgement in Plato's work depends heavily on the ability to control one's intellect, emotions were generally regarded as a threat to reason.⁴⁹ Yet, recent reinterpretations of Plato's work, such as those by Candiotto and Renaut (2020), Price (2009), Lorenz (2006), and Solomon (2008), argue that emotions may be integral to the philosophical journey and the development of one's moral reasoning.

Discussing “emotions” in Plato's work is always a delicate task, primarily because the term itself – of Latin origin – has no direct equivalent in ancient Greek. As Price notes, drawing on Liddell and Scott (1940), the Greeks instead used the term *pathos*, which broadly meant “that which happens to a person or thing” (Price 2009: 121). Consequently, scholars often prefer the terms “desires” or “motivations” when interpreting Plato's work, as these better capture the range of phenomena encompassed by *pathos*. This is particularly helpful, since Plato's work indeed covers experiences as diverse as thirst, fear, anger, and sexual arousal (Cooper 1999); phenomena that

⁴⁹ As noted by Solomon (1977) in Greece, *apatheia* (“apathy”) was a popular term that meant ‘freedom from the emotions’. *Apatheia* was believed to be a necessary step to attain *ataraxia*, a state where the individual achieves “peace of mind”. Moreover, it is interesting to note that this general idea, according to which it was preferable to distance oneself from one's own emotions, was also found in the non-Western views, as we can see in the Upanishads, Buddhists, Confucian, and Taoist theories of the mind (Solomon 1976).

contemporary affective researchers would not typically classify under a single (affective) category. Yet, setting aside these etymological and conceptual challenges, Plato's account nevertheless allows for a reading that anticipates some key elements of modern theories of emotion.

Candiotto and Renaut, for instance, argue for a reinterpretation of Plato's work in which some emotions should be recognised as having the potential to become "moral, just, and means for reason to rule over one's life" (Candiotto & Renaut 2020: 6). This concerns more particularly fear and love (*eros*). Plato's account attributes cognitive-like properties to certain desires and motivations that appear to arise from the lower parts of the soul, suggesting an early recognition of the role that affective states play in *shaping reasoning* and *guiding decision-making*. As Lorenz notes, "Plato [...] think[s] that the non-rational parts of the soul are able not only to give rise to fully formed motivating conditions, but also to be the bearers of expectations and of forward-looking emotions such as fear (Timaeus 69 C 5–D 4)" (Lorenz 2006: 58). Candiotto and Renaut support Lorenz's interpretation of the *Timaeus*, adding that Plato associates emotions such as fear with anticipatory judgements, implying that they can possess a cognitive dimension and influence reasoning about future events (Candiotto & Renaut 2020: 4-7).

By contrast, *eros* – the love of the Good – has long been seen as more than a mere perturbation arising from the lower parts of the soul. As Robert Solomon points out, "Plato's discussion of eros as the love of the Good in his dialogue *The Symposium* [demonstrates that] there are emotions involved in reason as well" (Solomon 2008: 4). *Eros* is not simply a disruptive force; it is integral to the philosophical journey, motivating individuals to aspire to higher ideals.

Viewed in this light, at least some emotions such as fear and love are not inherently antagonistic to reason, but can instead be essential to the pursuit of wisdom and moral virtue.

Emotions in Kant's account of moral reasoning

In his investigation of an *a priori* principle that guides rational agents' (moral) decisions and actions, Kant briefly considers, and then argues against, emotions as a potential foundation for providing an adequate basis for the objective laws of practical reason. Kant rejects the possibility of grounding practical reason in emotions on the basis that emotions are "empirical principles", with empirical weaknesses such as a lack of stability and apparent differences from one rational being to another (Tappolet 2022: 11). Emotions are neither seen as a good candidate to ground

morality nor do they seem to be compatible with the candidate that Kant chooses: the Categorical Imperative (Engstrom 2010). Moral actions are moral because they are done from duty to the moral laws. “From duty” is seen as one of the three possible *motivations* that guide human beings’ actions. As articulated by Korsgaard, you can perform an action from duty, from immediate inclinations, or “because you are “impelled to through another inclination,” that is, as a means to some further end (AK 4:397)”⁵⁰ (Korsgaard 1998: viii). From these three types of motivations, only actions performed from duty qualify as moral actions and are the instantiation of a free and autonomous will. Emotions (inclinations/desires), recognised as the second type of motivational states, can motivate an agent’s actions, but these actions *cannot* be moral. In addition to arguing that emotions cannot ground moral actions, Kant is clear that they *need not* participate in the decision-making process that leads to them: “an action from duty is to put aside entirely the influence of inclination and with it every object of the will ... there is left for the will nothing that could determine it except objectively the law and subjectively pure respect for this practical law” (G 400:01, in Guyer 2010: 130).⁵¹

Yet, Kant acknowledges the difficulty this requirement poses for moral agency and moral action, particularly through his concept of the “feeling of respect.” Although inclinations should be excluded from the reasoning process that leads to moral action, Kant repeatedly affirms the fundamental role of the subjective experience of a “pure respect” for the moral law (Herman 1993). What appears most troubling for Kant, therefore, is not the presence of emotions per se, but their potential *motivational force*. Barbara Herman, for instance, argues that moral emotions – such as respect – are not incidental but essential to Kantian moral agency (Herman 1993). This feeling of pure respect for the moral law leads Kant to recognise that, even if inclinations should play no role in the determination of duty and related actions, at the very least a feeling of respect remains necessary for the agent to grasp and *act* upon the demands of morality. Lawrence Blum (1982) also argues in favour of reconsidering the motivational power of some emotions in Kantian ethics. Specifically, altruistic emotions should be distinguished from the inclinations and changeable moods acknowledged by Kant, as they are “capable of the strength and reliability which the Kantian demands of moral motivation” (Blum 1982: 4) Blum argues for the recognition of the substantial role of sympathy, compassion, concern, and care in morality and moral motivation, claiming that these emotions possess the required capacity to counteract inclinations and personal feelings. In doing so, the original Kantian view of altruistic emotions

⁵⁰ By using “AK 4”, Korsgaard follows the German standard edition of Kant’s works, specifically referring to Kant’s *Grundlegung zur Metaphysik der Sitten* (Groundwork for the Metaphysics of Morals).

⁵¹ By using “G 400:01” Guyer uses a different abbreviation terminology but also refers to the Groundwork.

is significantly undermined and effectively replaced – though the formal structure of Kantian morality is retained (Blum 1982: 38-42).

This is evidence that even in some of the most rationalist philosophical psychologies of the mind, it has been extremely difficult, and, as I argue here, even impossible, to fully justify moral reasoning without sooner or later admitting to the intrinsic role of emotions in its fulfilment.

3.3.2. Emotions as essential to morality: Virtue ethics and sentimentalism

Plato's and Kant's wariness towards the potentially manipulative aspects of emotions has significantly constrained their accounts of emotions, particularly in terms of the nature, roles, and functions of emotions in morality. Thus, it is essential to turn to those ethical theories that have treated emotion as constitutive of moral understanding, in order to gain a richer and more comprehensive picture of the role and diverse functions of emotion in moral life.

Emotions in Aristotle's account of moral reasoning

As John M. Cooper (1999) noted, Aristotle's theory of virtue is a direct result of a coherent progression from Socratic rationalism – where emotions have little place – through the Platonist theory of mind – where emotions are discussed but mostly seen as non-rational and misleading, to Aristotle's account, where they are given a key role. Aristotle's theory of virtue is interesting because it embraces a view of emotions in which they are in an 'interfusion' with reason (Cooper 1999: 119). On the notion of control more particularly, Aristotle asserts that emotions are something that happens to rational agents, without any control or choice in the matter. Yet he argues that rational agents have control over their emotions once they have arisen. For him, the fact that one gets angry about x when x happens is indeed outside of one's control (Aristotle 1925: *NE* 1 105b28 and 1 06b9). However, one can *decide* whether one stays angry at x or not. In this sense, the emotions that are felt by the agent are as much of a choice as the actions performed in a situation. Thus, to be virtuous or not depends on the agent's actions and emotions. Feeling the right emotions and doing the right actions in a situation is what virtue is about (Stark 2001: 441). Exercising oneself to feel the right emotion, in its appropriate intensity and at the right moment, is essential to enable one to make the right decision. Appropriate emotions in a given situation guide the agent's decision by giving the appropriate degree of attention to the elements at play in that situation.

This is more complex than it might appear, as emotions can, and should in some contexts, be *multiple*. Hence, as noted by Stark, the courageous person who stands confident in battle where appropriate needs also to “fully and clearly appreciate the potential harm of proceeding into battle”, thus she needs to also feel fear (Stark 2001: 441). If the potential harm were staggering, the right decision would not be to stand and fight but to fear and flee, thus the appropriate virtue would be prudence or caution instead of courage (Stark 2001: 441).⁵²

Stark’s insistence on Aristotle’s recognition of the possibility – and the necessity – of experiencing a multiplicity of emotions in a given situation as a condition for appropriate decision-making highlights both the richness of affective life and the complexity of its relation to morality. Aristotle’s account treats emotions as subject to rational assessment; accordingly, emotions can be judged as appropriate or inappropriate, in terms of both their intensity and their object-directedness. The virtuous person can evaluate the appropriateness of her own emotions, and an external observer may likewise assess the fittingness of another’s emotional response.

Emotions in Hume’s account of moral reasoning

The classical rationalist view of emotions has been challenged by the antithetical view, which portrays emotions as dominating reason. In his work, Hume addresses the motivational foundations of moral action and consequently rejects the rationalist claim that morality is grounded in reason alone, famously asserting that “reason is, and ought only to be the slave of the passions” (Hume [1739–1740] 2000, in Solomon 1977: 41). Hume thus makes a significant move by inverting the traditional hierarchy of reason and emotion established by the rationalists.

Although he acknowledges that emotions can be dangerous and mislead one in one’s judgement of a situation, the key problem of the rationalist view is its failure to recognise the complexity of the emotional life. Hume posits a clear distinction between two types of emotions: “violent passions and calm desires and tendencies” (Mackie 1980: 46). In his view, rationalists have confused reason with the latter type (calm desires and tendencies), because these emotions, unlike the violent ones, produce almost no immediate sensations in the mind. Hence, apparently unaware of their existence, rationalists have mistakenly opposed reason alone to violent passions, when they should have opposed calm desires and tendencies (alongside reason) to violent passions (Mackie 1980: 46–47). In doing so, rationalists have also erroneously claimed

⁵² This is what distinguishes the courageous person from the rash one on Stark’s account (Stark 2001: 441).

that reason is the source of morality, when in fact its source is to be found in the emotions. The main reason why reason cannot be the source of morality is that reason alone cannot motivate action. For Hume, *reason* is a cognitive faculty concerned with the discovery of truth, the comparison of ideas, and the inference of causal relations (Hume [1739-1740] in Solomon 1977). Hume denies that reason has any motivational power of its own. It can direct us by informing us of the means to achieve our ends, but it cannot supply those ends or move us to act. The idea that emotions are necessary because reason alone cannot motivate the will to action is interestingly very close to the reinterpretation that Barbara Herman (1993) offers of Kantian morality, with the exception that in the latter ‘emotions’ are reduced to one specifically, namely, the feeling of respect for the moral law. But this nonetheless echoes a key point in the study of emotions, which is their motivational powers, which seem to ultimately be necessary for (moral) actions.

Following Hume’s account, emotions of the second type (calm desires and tendencies) possess *evaluative properties*. Their evaluative properties are what guide people’s actions and decisions by expressing approval or disapproval of a particular situation or individual. Approval and disapproval are kinds of pleasure and displeasure which one feels when assessing people’s behaviours and/or one’s own actions (Tappolet 2022). As Tappolet notes, these feelings are mediated by what Hume refers to as the mechanism of sympathy: “Given this mechanism, what happens when we consider a trait that tends to benefit others, such as benevolence, is that we tend to feel pleasure, thereby approving that trait” (Tappolet 2022: 10-11). Hume’s interest in sympathy as a human ability to *understand* the feelings and beliefs of others is not derived from pure reasoning but from a more phenomenological form of apprehension, in which the senses provide *information* unavailable through *a priori* reasoning.⁵³ In addition to being the mechanism by which we form moral judgements – and thus the basis of his moral theory – sympathy also enables socialisation and, to some extent, civilisation. Indeed, Hume roots justice in these natural feelings of approval (pleasure) and disapproval (displeasure), though he grants reason an important role in connecting them to “the social effects of the system of justice” (Mackie 1980: 85). Through the mechanism of sympathy, for instance, we can feel approval for the interests of society, and through reason we can understand that justice tends towards the common good. Emotions, therefore, are a significant source of knowledge, in that they can illuminate one’s preferred ends and goals (Tappolet 2022: 9). As Collier (2011) points out, Hume’s account of

⁵³ This also creates an interesting parallel with Blume’s modified account of Kant’s morality.

emotions (or feelings) significantly influenced today's contemporary theories of emotion rooted in sentimentalism and the psychological feeling tradition, including appraisal theory.

3.3.3. Contemporary accounts of emotions in moral reasoning

Contemporary accounts of moral reasoning rarely deny the role of emotions, though some assign them a more implicit instrumental role while others regard them as intrinsic to moral judgement and moral actions. The general idea that emotions provide moral salience⁵⁴ has been endorsed in different philosophical traditions, including sentimentalism (Greenspan 1988; D'Arms & Jacobson 2006), cognitivism (de Sousa 1987; Nussbaum 2001), phenomenology (Scheler [1923]1970; Stein [1917]1964), virtue ethics (Hursthouse 1999; Vallor 2016) and hermeneutics (Ricoeur 1990). Hence, considerable changes have been observed in philosophical work, requiring us to draw a broader and more complex picture of the full range of emotions that appear to underpin our most rational judgements and guide us in the pursuit of long-term goals. I will only say a few words about them here, as I will discuss them further in the next section.

Within the broad view that emotions contribute to morality, we can distinguish between holistic accounts – which typically investigate how moral emotions play an intrinsic role in the formation of moral judgement and action – and emotion-specific accounts, which concentrate on one or a few individual emotions. Emotions such as guilt, shame, pride, compassion, empathy, remorse, fear, and anger are among the most frequently discussed in moral theories of emotion, as their presence is often a key indicator of the presence of genuine moral dilemmas or normatively complex situations. Below I focus mostly on holistic accounts as I first want to draw a general view of emotions.

On the holistic side, Martha Nussbaum is one of the most prominent contemporary philosophers regarding the study of emotions in moral reasoning and judgement. Her hybrid account, between neo-Aristotelian and cognitivist appraisal theory, describes emotions as evaluative judgements, which “shape the landscape of our mental and social lives” (Nussbaum

⁵⁴ I understand “moral salience” to refer to an agent's ability to recognise which features of a situation are morally relevant. In philosophy of emotion, that ability is intertwined with the nature of emotions and hence it is seen as emotions capacity to render salient the moral features of our lives in given situations (e.g. feeling anger when seeing injustice). On the concept of moral salience *tout court*, see notably Barbara Herman (1993). On the concept of salience as a property of emotions that enable them to appraise relevant moral features of a situation, see notably Martha Nussbaum (2001).

2001: 1). She claims that emotions are judgements of value, and as such, are deeply intertwined with practical reasoning. But her account goes further by attributing an eudaimonistic direction to these judgements: “emotions appear to be eudaimonistic, that is, concerned with the person’s flourishing” (Nussbaum 2001: 31). Christine Tappolet (2017) also argues for the necessary role of emotions in reasoning, claiming that emotions are epistemic touchstones that allow us to perceive values, hence they help us track practical reasons. Without denying that they may be defeatable, she argues that they nonetheless “confer *prima facie* justification to evaluative beliefs” (Tappolet 2017: 168). Hence, an agent who meets Tappolet’s conditions – namely, possessing appropriate epistemic and practical dispositions – may act on the basis of her emotions and legitimately claim to be acting in accordance with reason (Tappolet 2017: 194). Drawing heavily on Aristotle’s virtue ethics, Rosalind Hursthouse identifies emotions as central to moral agency (Hursthouse 2000). On her account, emotions are moral responses to situations, which will shape what the agent takes to be relevant in a given context. Finally, Ben-Ze’ev (2000) and Blum (1982)⁵⁵ have both addressed the important question of the ‘partial’ character of emotions,⁵⁶ and defended the idea that this property of emotions is, in fact, *essential* to distinguish morally relevant features in specific contexts. In interpersonal situations, being partial instead of impartial is necessary to communicate but also to perceive other people’s moral values and concerns.

As we will see below, philosophical accounts of emotions continue to exert a substantial influence on contemporary theories in psychology and affective neuroscience. Sentimentalist traditions, which locate the roots of moral agency in emotional experience, find clear echoes in research demonstrating the centrality of affect in moral judgement, empathy, affiliation, and pro-social motivation (Haidt 2001; Damasio 1994). Cognitivist theories, which conceptualise emotions as evaluative judgements or appraisals (e.g. Nussbaum 2001; Solomon 1976), directly inform modern appraisal theories in psychology (Lazarus 1991; Scherer 2005) and have shaped neuroscientific research showing that emotional responses depend on higher-order interpretation rather than mere physiological reaction (Scherer 2009). These approaches contribute to the now-standard view that emotions track meaning and are modulated by goals, beliefs, and context. At the same time, virtue-theoretic perspectives, which treat emotions as dispositions central to moral character and practical wisdom, anticipate contemporary findings in developmental and social psychology on emotional learning, habituation, and regulation

⁵⁵ See Blum (1982) for a more emotion-specific account focusing on altruistic emotions in morality.

⁵⁶ The idea that there is no such thing as an ‘objective’ emotion.

(Gross 1998), as well as neurobiological accounts of emotional plasticity and (affective) social scaffolding (Singer & Klimecki 2014).

3.4. Integrative perspective on emotions

In this section, I analyse three general perspectives on emotions, which I consider to provide a broad overview of the work done on emotions, particularly in terms of the primary assumptions that have driven the field of affective computing.⁵⁷ These three general perspectives are 1) the biological/evolutionary view, 2) the cognitivist view, and 3) the social constructionist view.⁵⁸ I see these three general perspectives as three different ways to approach the same group of phenomena (viz. emotions). For the present work, the tensions among these perspectives are less interesting than the agreements they share. Each perspective brings to our attention a set of important features of emotions, which are not necessarily incompatible with the other perspectives if we do not endorse some of the fundamental assumptions of each individual perspectives. Perhaps more importantly, together, they provide us with an integrative understanding of the relations between emotions and their expressions. Below I present a brief summary of the core tenets of each perspective.

3.4.1. The biological/evolutionary perspective

The biological and evolutionary approach to emotion finds its classical articulation in Darwin's work *The Expression of the Emotions in Man and Animals* ([1872]1965).⁵⁹ On this view, emotions are understood as biological products shaped by millennia of evolutionary constraints, serving functions related to both survival and communication. Hence, emotions are biological responses to adaptive challenges across species.

Proponents of this perspective typically endorse the idea that emotions are independent of, and often precede, cognition, in contrast to the cognitive perspective (see Zajonc 1980, 1984). Accordingly, they posit a set of primary or basic emotions, conceived as automatic responses

⁵⁷ Other ways of grouping emotion accounts have been proposed. For example, the threefold division, which divides accounts into 1) feeling accounts, 2) appraisal accounts, and 3) motivational accounts, has received considerable support (see Solomon 1977; Scarantino 2020; Tappolet 2022). However, here I am following Cornelius (1996) division, as it is better adapted to also include accounts dedicated specifically to emotional expressions rather than only emotions, which will be useful for linking this section with the subsequent ones.

⁵⁸ Note that some of the authors mentioned in the different perspectives most of the time endorse a more complex and nuanced view than the general one defined here.

⁵⁹ On Darwin's view, emotional expressions have been finely tuned to meet our evolutionary needs for both expressing our inner affective states (see the *principle of serviceable habit*) and communicating with others (see the *principle of antithesis*) (Darwin [1872]1965).

elicited by situational changes (or absence thereof). This perspective received unparalleled popularity following Paul Ekman's claim (1992) that the basic emotions he identifies, namely fear, anger, sadness, joy, disgust, and surprise, are universally recognisable and innately expressed through distinctive physiological patterns and facial expressions (see Chapter 2, §2.2.1). Following the basic emotion theory, emotions are considered functionally linked to specific action tendencies (e.g. fear prepares for flight, anger for confrontation), and are often associated with characteristic patterns of autonomic activity such as increased heart rate, changes in skin conductance, or muscular tension, though with recognised variability across contexts and individuals (Ekman 1992; Izard 1971). These physiological changes are thus treated not as incidental by-products of emotions but as integral components of the emotional response system, preparing the body to act in ways that have been proven evolutionarily advantageous.

Proponents of this view differ in their commitments. Strong versions assert that emotions are innate mental states, universally expressed, and biologically predetermined. Within these strong variants, some theorists hold that physiological changes are inherent parts of the emotion itself (Ekman 1992), while others treat them as components of the emotional system rather than mere outward behavioural expressions (Izard 1971). Weaker variants maintain that while the underlying physiological mechanisms may be species-wide, their expressive manifestations are modulated by learning and context. Thus, the weaker accounts will typically state that, while emotions are grounded in biology, their expressions can be shaped by environmental and cultural variables in functionally adaptive ways (Griffiths 1997). To get around these social difficulties, stronger versions, such as Ekman's theory of emotion, will develop notions such as the concept of display rules, which are cultural norms that regulate the 'surface' of emotional expressions (Ekman & Friesen 1969). Display rules aim to account for the observation that emotional expressions are often suppressed, exaggerated, or altered depending on social context. Yet, Ekman (1992) maintains that one can always see the "genuine" emotion if one knows what type of micro-expressions to look for in a given prototypical emotion-type.

As mentioned in the preceding chapter, this biological and evolutionary framework has significantly influenced the field of affective computing, particularly in efforts to design systems that can recognise, model, or simulate human emotions. The FAC system, inspired by the work

of Carl-Herman Hjortsjö⁶⁰ and developed by Ekman and Friesen (1976, 1978), is still in use today in state-of-the-art scientific research in affective computing.⁶¹

In sum, the biological/evolutionary perspective often emphasises the innateness, universality, and (necessary) physiological embodiment of emotions. While its early formulations focused on observable expressions, contemporary variants integrate neurophysiological evidence to support the view that emotions are complex, biologically embedded states with deep roots in evolutionary history. Contemporary affective neuroscience has deepened our understanding of the biological basis of emotion by identifying specific brain structures, pathways, and bodily systems involved in emotional processing. LeDoux (1996), comparing the amygdala to the hub of a wheel, demonstrated its involvement in the appraisal of emotional meaning by showing how fear responses are coordinated by the amygdala and subcortical circuits. Panksepp (1998)⁶² identified primary affective systems in mammals that govern emotional behaviours across species, attempting to bridge the gap between the mammalian brain and the human mind.

But this perspective has been challenged, most notably by Barrett (2017) and Barrett and Satpute (2019) who, following a constructivist framework, argue that emotions are not discrete neural packages, but dynamic and context-sensitive events constructed by the brain through predictive processes, emotional regulation, and concept-based inference. This view shifts the emphasis of the biological/evolutionary perspective from hard-wired and fixed emotion circuits to flexible brain networks that support embodied inference.

Hence, despite sustained critiques of the basic emotion theory, the biological/evolutionary perspective as a whole, and particularly its contemporary developments, remains foundational, offering indispensable insights into the biological, embodied, and adaptive functions of emotion.

⁶⁰ See Wagner, H. (1997) for a detailed explanation of how the work of Ekman & Friesen took inspiration from Hjortsjö, C. H. (1969), including the development of FACS (Wagner 1997: 34-35). See also on this point Eibl-Eibesfeldt (1989).

⁶¹ See Chapter 2. iMotions, for instance, explicitly states its use of FACS. The assumption that emotional expressions have identifiable physiological patterns – such as facial configurations, vocal prosody, or biometric data – underpins many, if not most, contemporary computational approaches to emotion detection and human-machine interaction.

⁶² See also Montag & Panksepp (2016).

3.4.2. The cognitivist perspective

Cognitivist theories of emotion share the foundational view that emotions are intentional, meaning-laden responses to states of affairs in the world. Unlike biological/evolutionary perspectives, which conceive of emotions as hard-wired mechanisms (e.g. Ekman 1992; Panksepp 1998), cognitivist theories treat emotions as cognitive engagements with value – albeit in varied forms. Within the cognitivist perspective, scholars diverge on whether emotions are best understood as judgements, appraisals, perceptions, or narrative structures, but all agree that emotions are not reducible to physiological states or behavioural reflexes. Instead, they are intentional phenomena, responsive to reasons and assessable in terms of rationality or fittingness.

The judgement theories have played a major role in overturning the traditional philosophical idea that emotions were irrational (or arrational). Philosophers in the classical rationalist tradition were notably wary of the negative effects that emotions could have on reason, and generally looked for ways to suppress them from the identified components that led to rational judgement (Solomon 1976). The judgement theories may be conceived as the most classical (or radical) form of cognitivism, directly opposing the classical rationalist assumptions. They hold that emotions have evaluative properties through which it is possible to process certain types of information and yield value-laden judgements (Solomon 2003; Green 1992; Goldie 2000; Nussbaum 2001). According to Solomon (1976, 2003), an emotion such as anger is not simply a feeling but a complex evaluative stance: a judgement that one has been wronged and that this matters. Although not all judgements are emotions, emotions themselves are evaluative judgements; they enable individuals to interpret and navigate morally salient situations.

My anger *is* that set of judgements. Similarly, my embarrassment is my judgement to the effect that I am in an exceedingly awkward situation. My shame *is* my judgement to the effect that I am responsible for an untoward situation or incident. My sadness, my sorrow, and my grief *are* judgements of various severity to the effect that I have suffered a loss. An emotion is an evaluative (or a “normative”) judgement, a judgement about my situation and about myself and/or about all other people (Solomon 1976: 126, author’s original italics).

Hence, for Solomon, the judgement constitutes the emotion itself. Similarly, Nussbaum (2001) argues that emotions are evaluative judgements that reveal the importance of external events to

one's flourishing. Thus, emotions are more precisely *judgements of value*, making them a type of (evaluative) belief.⁶³

Alongside judgement theories, we can find appraisal theories as another subtype of cognitivism, although slightly more moderate than their counterparts in their claim of emotions as evaluative phenomena. Appraisal theories do not hold that emotions may yield *explicit* judgement, but rather, that their nature is evaluative. Klaus Scherer's Component Process Model (1984, 2001), for example, formalised this idea by considering emotions as complex processes rather than events. Through a combination of cognitive appraisal and physiological responses, Scherer's model conceptualises emotions as "the interface between an organism and its environment mediating between constantly changing situations and events and the individual's behavioral responses" (Scherer 1984: 295). His theory emphasises the dynamic and interactive nature of emotional experience. Phoebe Ellsworth and Klaus Scherer (2003) have further refined the appraisal model by proposing to recognise *dimensional appraisal variables*, such as control, agency, and certainty, which differentiate emotions like fear, anger, and guilt. Under the appraisal model, appraisals may be considered as either antecedents of emotions or components, with the latter being retained as the most probable (Ellsworth & Scherer 2003: 575). Richard Lazarus (1991) proposes that emotions need not involve full-blown conceptual thought but rather involve appraisals of events in terms of their relevance to the individual's goals and coping capacities. One of Lazarus' notable contributions to the field is his distinction between primary appraisal and secondary appraisal. Emotions evaluate the significance of a situation for our well-being (primary appraisal), but they are also an evaluation of our potential to cope with the new situation (secondary appraisal). Thus, in the secondary appraisal, there is an element of anticipation, where the appraisal is actively projecting the individual into assessing the probability of potential outcomes. In his secondary appraisal component for compassion, Lazarus notes

If there is self-directed blame, then the probability of guilt is increased; if blame is directed at the victim, then the probability of anger is increased; if there is no blame, then compassion is likely, depending on the dispositional characteristics of the person. (Lazarus 1991: 290).

This not only shows the forward-looking component in secondary appraisal but also the interrelations between emotions and formal types of events, and between a person's dispositional characteristics and their emotions. These models integrate physiological, cognitive, and motivational components. While appraisal theorists share with judgement theorists the idea

⁶³ Even though it should be noted that in Nussbaum's account emotions also have an affective (feeling) component and thus are not *purely* a type of belief.

that emotions are about value, they are generally more open to non-conscious forms of evaluation, including embodied affective processing.

Within the cognitivist perspective, narrative and perceptual theories complete the picture by offering further refinements. On the narrative side, Keith Oatley (1992, 2004) suggests that emotions are simulated narratives that help individuals understand and navigate social interactions. Emotions, on this view, are like cognitive “mini-dramas” that reveal what matters to the agent in a given situation, guiding attention, inference, and behaviour. Ronald de Sousa (1987) offers a closely related model in his theory of “paradigm scenarios”. These socially and culturally learned scenarios shape emotional perception and response by linking types of situations with patterns of evaluation and action. For de Sousa, emotions are rational insofar as they are formed within coherent repertoires and respond fittingly to evaluative features of the world.⁶⁴ Emotions are thus embedded in the narrative structure of consciousness, linking past experience, present meaning, and future action.

A yet different group of theories is proposed by the perceptual account. Tappolet’s theory of emotions, for instance, postulates that emotions *represent* the world as being a certain way. However, unlike judgement or appraisal theories, perceptual theories argue that the content of emotions is not conceptual but *representational*: emotions do not evaluate or appraise events, people, or states of affairs, but rather perceive *value properties* in a way analogous to sense perception. Emotions, like perceptual experiences, are non-conceptual but representational: they present the world as dangerous, offensive, admirable, or shameful. Emotions, on this view, are not conclusions of reasoning but rather quasi-perceptual modes of awareness, capable of being accurate or inaccurate in how they track evaluative facts.

Despite their differences, they all defend the view that emotions are structured engagements with meaning. Whether as judgements, appraisals, narratives, or perceptual tools, emotions are treated as cognitively structured, rationally assessable, and responsive to normative considerations. On this perspective, emotions process information from situations or events and attribute meaning to it. This ascribed meaning is seen as essential for navigating daily situations, by guiding one’s decisions and actions towards goals that matter to them. Together,

⁶⁴ The “evaluative features of the world” are those aspects of a situation that *matter* to an agent, in terms of personal significance, practical relevance, or moral or social meaning.

these cognitivist theories provide a rich and flexible account of emotions that accommodates their intentionality, complexity, social embeddedness, and evaluative character.

3.4.3. The social constructionist perspective

The social constructionist perspective holds the idea that emotions are designed to serve social roles and purposes (Averill 1980, 1982; Armon-Jones 1985; Parkinson, Fischer & Manstead 2005). One of its central points is that emotions (and emotional expressions) are learned rather than innate (Harré 1986; Lively & Weed 2016). Catherine Lutz, for instance, argues that the very existence of emotions could be situated at the cultural level (Lutz 1988).⁶⁵ This claim stems from the observation that emotions are subject to fundamental cultural variability, which results in some emotion-types being present in some cultures and not in others (Averill 1982; Wierzbicka 1999; Tcherkassof 2018). Anthropological studies conducted among the Utkuhikhalingmiut in northern Canada have revealed that some communities lack a concept corresponding to one of the so-called “basic emotions” – namely, anger (Briggs 1971). As a result, they are allegedly unfamiliar with the emotion itself. This directly challenges the claims of the biological and evolutionary accounts of emotion. Moreover, the social constructionist perspective has assumed a powerful interplay between emotions, cultures, and languages. For instance, the English language does not have an equivalent for some emotions that are present in other cultures such as “*dépaysement*” in French or “*Schadenfreude*” in German (Russell 1991). Kundera ([1978]1987) supposes that the word (and corresponding emotion) “*Litost*” in Czech, has “no exact translation in any other language” (Kundera [1978]1987 in Russell 1991: 426).

Extending this perspective through a neurobiological foundation, Singer and Klimecki’s work (2014) on empathy and compassion has supported the view that social emotions are not only evolutionarily adaptive but also deeply embedded in brain systems that facilitate social cohesion and moral concern. In particular, their work brings empirical evidence that it is possible to transform social emotions through training: “This finding underlines the malleability of social emotions as it shows that a short-term compassion training of several days can foster positive feelings and related brain activations, even when persons are exposed to the distress of others.” (Singer & Klimecki 2014: 877). Hence, it seems that societies, through culture and social norms, at least partially shape the landscape of our emotions.

⁶⁵ Additional claims have been made in this view besides the central role of culture, one particularly interesting being gender variability (Fischer 2000).

Set out in this way, these views may appear irreconcilable.⁶⁶ However, if we relax some of their stronger claims, they need not be. In the next section, I adopt a function-first integrative perspective on emotions, grounded in core agreements between these different perspectives, to further investigate ERTs' potential reach into the affective life.

3.5. Adoption of a function-first integrative perspective on emotions

The preceding sections show that no single philosophical or scientific perspective can, on its own, account for the complexity of emotional life. Each, however, isolates a genuine dimension that ought to be retained. I therefore adopt an integrative and functional perspective on emotions: a methodological approach that explains emotions in terms of the functions they fulfil in lived practice (particularly in practical reasoning, social coordination, and normative life) while integrating transdisciplinary insights from their embodied, evaluative, and social dimensions. This approach aligns with contemporary function-first and hybrid models in philosophy and affective science (e.g. de Sousa 1987; Frijda 1986; Griffiths 1997; Prinz 2004; Scarantino 2017), which likewise combine biological, cognitive-evaluative, and social analyses to render emotions intelligible in terms of action and interaction.⁶⁷ In this dissertation, however, theories of emotion function as explanatory lenses, assessed by how well they illuminate what emotions do in practice. The function-first integrative perspective adopted here is thus selective in the following sense: it retains only those theoretical commitments needed to understand emotions as forces that orient attention, confer salience, guide action, and shape social meaning.

Although this perspective could serve as a general approach to emotions, it is adopted here for a more specific reason: to better grasp the distinctive ethical challenges raised by fully functional ERTs. Because ERTs aim to render affective life externally readable, interpretable, and classifiable, a proper ethical evaluation of such technologies requires an understanding of what

⁶⁶ As is the case almost everywhere else in research, specific research foci and ideologies inevitably lead to the rejection or deliberate neglect of certain facts or possibilities. As Dukes et al. (2021) rightly point out, affective sciences as a whole have been rejected (or deliberately ignored) for a long time because scientists “typically neglect what they cannot measure”. This is true of emotions but also of the different perspectives on emotions. Hence, we can observe that, typically, the biological/evolutionary views tend to ignore cognitive-related aspects and reciprocally, cognitivist views typically ignore the biological/evolutionary related components – which does not necessarily mean that they reject them but rather that they *overlook* them (Dukes et al. 2021). This may be seen from two different points of view, either that the result of this is that fundamental theories of emotions seem to have irreconcilable core tenets, or that, by focusing on things so drastically different, they may complement each other. I endorse the latter. (see Green (2007: 108-109), for a comparable stance.)

⁶⁷ As already discussed, other hybrid approaches may be found in the work of Solomon (1976), Nussbaum (2001), Goldie (2000), Ben-Ze’ev (2000), and Armon-Jones (1985), to name only a few.

is at stake when emotions are treated in this way. An integrative perspective on emotions provides us with such understanding.

3.5.1. Transdisciplinary agreements on their nature, role, and functions

Despite ongoing disagreements among biological, cognitivist, and social constructionist perspectives on both the definition and ontological status of emotions (Fehr & Russell 1984; Scarantino & de Sousa 2018), recent work across philosophy, psychology, affective neuroscience, and anthropology has generated many hybrid approaches that try to reconcile these different views. These hybrid approaches have been very helpful to pinpoint where major agreements may be found. These convergences, while framed within different methodological and theoretical commitments, offer a productive basis to better understand both the capacities and contexts in which emotions operate.

There have been many attempts at complete or partial reconciliation between some of the deepest existing tensions between those perspectives. Fridlund's work (1991) notably attempts to reconcile both the biological/evolutionary view with the social constructionist view. De Sousa's account (1987), despite being cognitively oriented, treats emotions as culturally shaped but biologically grounded. Frijda (1986), Griffiths (1997), and Barrett (2017) also developed hybrid approaches that bring together components of the biological/evolutionary perspective with the cognitivist perspective. Frijda's work (1986, 1988), on emotions as states of action readiness, blends both the cognitivist and biological perspective, while also accounting for some key aspects of the social constructionist perspective. Solomon's work, although classically cognitivist at first (1976), later embraces the social constructionist perspective (2003). And Ben-Ze'ev's work (2000) notably brings together aspects of the cognitivist view with the social constructionist one by recognising emotions as grounded in cognitive evaluations while also being deeply shaped by social, cultural, and interpersonal contexts.

Hence, I now turn to some of the main transdisciplinary agreements on the nature, role, and functionalities of emotions, with the function-first integrative perspective and the hybrid approaches in mind.

Agreements on the nature of emotions

Contemporary agreements can be noted across the different disciplines concerned with the study of emotions. Regarding the nature of emotions, most affective researchers agree that emotions are multifaceted, object-directed, and context-sensitive, and that a comprehensive theory of emotions needs to consider both their psychological and social dimensions.

Emotions are complex, multifaceted phenomena that cannot be reduced to a single component or causal mechanism. Most affective researchers agree that emotions involve a constellation of physiological, cognitive, behavioural, and expressive elements. Key components are typically recognised, even though the role, importance, and relations of these components are still debated. In general, affective researchers recognise that emotions have a 1) physiological (e.g. increased heartbeat), 2) phenomenological (e.g. unpleasant feeling), 3) evaluative (e.g. appraising the bear coming your way), 4) expressive (e.g. upper eyelids raised, jaw dropped open, lips stretched horizontally), and 5) behavioural component (e.g. an impulse to flee). From an integrative perspective, emotions are thus best viewed as biologically *grounded*, cognitively *structured*, and socially intelligible modes of affective orientation, *shaped by* and shaping the (cultural) contexts in which they emerge.

Additionally, it is generally agreed that emotions are intentional, object-directed phenomena,⁶⁸ which makes them potentially assessable based on their fittingness and/or appropriateness. In contemporary research, “object-directed” is broadened to also allow emotions to be directed towards people, situations, events, or state of affairs. This object-directedness is often analysed in terms of both particular objects (e.g. the bear, the sister, the rug) and formal objects (e.g. danger, loss, slights) (Kenny 1963; Teroni 2007). Assessment of particular objects is dependent on their relation to a formal object; in other words, the appraisal necessitates some conception of the given object in the formal form to evaluate the emotion’s fittingness. For instance, the reason one feels fear when seeing a bear is because one appraises the bear as something dangerous (Scarantino 2020: 122; Naar 2021). Relatedly, if the fitting objects of fear are dangers, the fitting objects of anger are slights, and those of sadness are losses, and so on (Scarantino 2020: 122).

⁶⁸ This is one particularity that distinguishes emotions from other affective states, such as moods for instance, which have no particular object.

Parallel to this is the agreement on the idea that people may (and likely will) have multiple, simultaneous emotions at a given time (Greenspan 1980; Nussbaum 2001; Fischer, Pauw, & Manstead 2019). Within virtue ethics, emotions are often understood as multiple and co-occurrent in the agent, because this multiplicity is required for the full appraisal of complex situations. Some virtues are attained only through this navigation of the emotional life. Recall the courageous person, who following Stark's analysis, can only be courageous if she feels both the fear of the fearful and the confidence, hope, or feeling of duty (in battle) (Stark 2001: 441). Fischer, Pauw, and Manstead (2019) give another example, where someone receives a diagnosis of a terminal disease. In their scenario, the affected person feels sad about the situation but also angry and guilty, all for different reasons. After analysis, it appears that these emotions seem to be all appropriately directed, and in that sense, fitting, yet they are simultaneously occurring.⁶⁹ The co-occurrence of multiple emotions usually arises when the emoter is hesitant in the face of, or in recognition of, several valuable possible points of view. This complexity is also well aligned with folk intuitions about inner emotions, especially in complex contexts or social interactions.

Agreements on the role of emotions

Among the major agreements on their role is their central part in reasoning processes and decision-making. Contemporary affective researchers consider emotions as the *dominant driver* of most meaningful decisions in life (Frijda 1986; Lazarus 1991; Nussbaum 2001; Keltner et al. 2014), but they are also seen as playing an active role in the formation of judgements and choices on a daily basis (Kahneman & Tversky 1979; Lerner et al. 2015). Emotions are recognised as providing useful information to the self about what is valuable and what should be pursued (or avoided) (Solomon 2003; Frijda 1986; Nussbaum 2001). Hence, there is a broad consensus that emotions play a central role in human life, not merely as internal states but as processes that mediate between agents and their environments. Emotions are recognised as orienting devices: they direct attention, structure perception, and highlight what is salient in a given context. They inform practical reasoning, especially in conditions of uncertainty, where explicit deliberation may be insufficient to resolve value-laden dilemmas. In this respect, emotions function as both affective responses and evaluative guides, enabling agents to navigate complex social and moral landscapes (Frijda 1986; Nussbaum 2001; Tappolet 2017). Crucially, emotions also mediate social interaction. They signal evaluative stances to others, communicate expectations, and invite

⁶⁹ See Fischer, Pauw, & Manstead (2019: 9) for more details on the example.

particular forms of recognition, resistance, or alignment. They are thus implicated not only in individual inner experience but also in the regulation and negotiation of social life.

Agreements on the functions of emotions

In terms of their functions, emotions are increasingly understood as adaptive, epistemic, normative, and social. Studies in neurobiology tell us that, biologically, at least some emotions appear to (have evolved to) serve to regulate the body's responses to significant stimuli, coordinating actions such as fight, flight, or care, and to cope with new (social) situations. Studies in cognitive science tell us that emotions foreground what matters to the agent in a given situation (de Sousa 1987; Tappolet 2017). They also perform an essential normative function, reinforcing social expectations and enabling the affective transmission of values and norms. Guilt, shame, indignation, and pride, for instance, do not merely reflect internal states but operate as mechanisms for social cohesion and moral accountability (Armon-Jones 1985; Ben-Ze'ev 2000; Nussbaum 2001). In addition, emotions are culturally and socially sensitive (Averill 1982; Wierzbicka 1999; Parkinson, Fischer, & Manstead 2005), and have a plastic, developmental dimension: they can be shaped, refined, or redirected through cultural training, interpersonal interaction, and reflective practice (Singer & Klimecki 2014; Barrett 2017). This shows that emotions are not static properties of the mind but complex, multifaceted, dynamic, and to some degree malleable.

The biological, cognitive, and social dimensions of emotional experience are not competing explanations but interdependent lenses for understanding emotions. Taken together, their agreements support a conception of emotion that sees them not as self-contained episodes, or states, but as dynamic processes that mediate between the individual and their world.

3.5.2. Emotions' action tendencies

So far, I have focused on emotions themselves, emphasising agreements on their nature, role, and functions. I now briefly turn to the *relation* between emotions and their related expressions, focusing in particular on the following relations: 1) emotions and action tendencies, 2) emotions and physiological changes, and 3) emotions and *performative* acts (in the sense that they iterate something).

The claim that emotions are action-oriented is not a new one. It is already well known that, for instance, anger is typically linked with aggression, fear with avoidance, happiness with prosocial

behaviour, sadness with reflection or withdrawal, disgust with avoidance or purification, guilt with reparative actions or distancing, and love with nurturing or commitment. These action tendencies of emotions are highly valuable for any emotion recognition task that wants to expand the ‘recognition’ task into a ‘prediction’ task. Understanding that anger commonly entails a disposition towards confrontation or aggression is useful for purposes of both self-preservation and security. Likewise, understanding that sadness often tends to yield social withdrawal, and that love tends to manifest into nurturing or commitment-type action is also useful. Frijda’s theory of action-readiness (1986, 1988) already discusses the ways in which emotions are action-oriented. For him, it is clear that the “primary” or “basic” emotions typically identified by the biological/evolutionary views, namely fear, anger, joy, surprise, disgust, and sadness, yield clearer correlations between specific emotion-type and specific action-type, whereas ‘higher-order emotions’ (Frijda 1988: 351), such as remorse, guilt, or jealousy, may keep the door open for more complex and/or uncertain action-type. In the case of guilt, for instance, narrowing the action-type to “reparative actions” and “distanciation” still leaves open two widely different type of actions. Andrea Scarantino (2017, 2019) similarly makes the case that different emotion types correspond to different forms of action tendencies, however, on his account, these associations are not deterministic as they only represent *probabilistic patterns* of action readiness that are highly context-dependent, and subject to modulation by individual knowledge and other social factors. Barrett’s theory of constructed emotion (Barrett 2017) offers a more nuanced alternative, where emotions function as context-based anticipatory predictions. Emotions infer the meaning of a situation and make sense of it based on prior experience, and from these prepare the body for appropriate action. On this account, emotions proactively navigate uncertain situations.

As Scarantino and Barrett’s accounts highlight, human emotional life rarely conforms to a neat one-to-one mapping between stimulus and response, as context, individual’s social situation, and culture greatly complicate the picture.⁷⁰ Additionally, as the moral theories of emotions have pointed out, a defining feature of emotions in lived experience is their simultaneity and potential for conflict. Emotions are frequently multiple, co-occurring processes that thus generate *competing* action tendencies. For instance, one may feel both guilt and anger towards a loved one, or experience fear and compassion in the face of another’s suffering. Nussbaum (2001), de Sousa (1987), and Greenspan (1980) have also acknowledged this emotional multiplicity, arguing that

⁷⁰ This is to say that, for instance, you may know that you tend to smile when nervous, which would not follow a strict one-to-one “universal” mapping in which smiles correspond to joy (or happiness).

it reflects the layered, value-laden nature of our concerns. Greenspan, in particular, emphasises the idea of emotional ambivalence as both a challenge and a resource for practical reasoning. De Sousa's concept of paradigm scenarios shows how emotional training in culturally salient contexts can generate patterns of response that pull in different directions, especially when new or hybrid situations emerge. These conflicting tendencies reflect the moral complexity of emotional life, especially where multiple values or duties intersect. As a result, emotions often present different tensions, constraints, but also possibilities through which agents must navigate. Emotional multiplicity does not undermine their action-oriented function, but it certainly makes it more complex.

Because emotions are processes of anticipating social outcomes, they are inherently other-responsive and norm-guided: they contribute to the structuring of social life through their expressions. Ben-Ze'ev (2000) argues along those lines when stating that emotions arise from and feed social interactions. Similarly, Fischer, Pauw, and Manstead (2019) have shown how emotions like shame, guilt, and anger function as moral regulators in group dynamics. These emotions not only orient the agent towards a response (e.g. reparation, protest, withdrawal) but also signal value violations or commitments to others once expressed, thereby supporting (or hindering) moral expectations within communities. Hence, emotions and their expressions are viewed as reactions to changes in situations or state of affairs but also as the cause of these changes for others. This means that their related action tendencies should not be assessed in isolation but rather understood as constituting the social context that they help maintain or change.

3.5.3. Emotions and physiological changes

Knowing that emotions are closely tied to what individuals value in the world – and that they shape both every day and life-defining decisions – is crucial for assessing the potential impact of ERTs. If fully functional, such technologies do more than simply detect emotional states; they also reveal a person's underlying values and their sense of what is morally salient.

As previously discussed, emotions also indicate the most probable action tendencies of emoters, at least in some cases. However, such recognition is possible only if there are some ways of *accessing* this type of information. On the question of the relevant indicators that reveal this type of information, we find on one end of the spectrum evolutionary and biological theories which claim that emotions are directly embodied and therefore detectable through physiological

changes. According to this view, specific autonomic changes – such as elevated heart rate or pupil dilation – are not merely correlated with some emotions but constitute an essential part of the emotion itself. One cannot feel a particular emotion type without *exhibiting* the associated physiological changes.

At the other end of the spectrum, theorists like Barrett (2006, 2017) reject the idea of fixed “fingerprints” for emotion types. While she acknowledges that certain prototypical physiological changes may commonly occur with specific prototypical emotions, she emphasises that such correlations are context-sensitive and highly variable. This means that in opposition to essentialist claims, ‘anger’ does not always involve a heightened heart rate, ‘fear’ does not necessarily involve sweaty palms, and ‘embarrassment’ does not always involve blushing.

What does it mean that four meta-analyses, summarizing hundreds of experiments, revealed no consistent, specific fingerprints in the autonomic nervous system for different emotions? It doesn’t mean that emotions are an illusion, or that bodily responses are random. It means that on different occasions, in different contexts, in different studies, within the same individual and across different individuals, *the same emotion category involves different bodily responses*. Variation, not uniformity, is the norm. (Barrett 2017: 15, italics from original author).

Between these poles lies the cluster-based approach. This view maintains that while no single physiological marker maps uniquely onto a particular emotion type, some patterns of changes tend to cluster around specific emotions. Isolating one variable – say, increased heart rate – and interpreting it as a marker of a single emotion type is methodologically flawed, as Barrett’s critique makes clear. Yet researchers have not exhausted the possibility that identifiable clusters of physiological responses may track emotional experiences, simply because not every possible physiological response have been measured in every possible way at the same time. (Scherer 2005; Larsen et al. 2008; Mauss & Robinson 2009).

This hypothesis has motivated a range of empirical efforts to identify which combinations of physiological changes might form meaningful emotion-specific clusters. Most research has, for practical reasons, focused on a specific set of physiological changes that typically include cardiovascular response, respiration rate, skin conductance, and pulse (Kreibig 2010; Zhang et al. 2020); attempting to find relevant clusters among them. More recent studies in affective computing have expanded the range of signals under investigation, including pupil dilation (Kinner et al. 2017), eye movement (Lim, Mountstephen, & Teo 2020), and brain activity patterns such as EEG signals (Q. Wang et al. 2022; Zhang et al. 2020).

However, a central challenge remains: if fear is typically – but not always – associated with increased heart rate, pupil dilation, elevated respiration, and heightened skin conductance, but anger can also be associated with the same cluster of physiological changes, then physiological overlap threatens the discriminability of emotional categories. Addressing this issue, Scherer (2005) argues that the real challenge here is that a comprehensive approach to emotion measurement must assess a constellation of factors:

[...] there is no single gold-standard method for [emotion] measurement. Rather, given the component process nature of the phenomenon, only convergent measurement via assessment of all component changes involved can provide a comprehensive measure of an emotion. In other words, in an ideal world of science, we would need to measure (1) the continuous changes in appraisal processes at all levels of central nervous system processing [...], (2) the response patterns generated in the neuroendocrine, autonomic, and somatic nervous systems, (3) the motivational changes produced by the appraisal results, in particular action tendencies [...], (4) the patterns of facial and vocal expression as well as body movements, and (5) the nature of the subjectively experienced feeling state that reflects all of these component changes. Needless to say, such comprehensive measurement of emotion has never been performed and is unlikely to become standard procedure in the near future. (Scherer 2005: 709).

Without accommodating all of these points, recent advances in affective computing, particularly in multimodal emotion recognition using deep learning techniques, have significantly expanded the range of physiological data that can be gathered and analysed within a single experimental setting. Of particular promise is recent work on physiological signals, which may allow researchers to detect more nuanced physiological clusters previously inaccessible through traditional methods (see Yin et al. 2017). New ML methods, coupled with robust multimodal datasets, might open the door to a real possibility of discovering new ways to classify (and understand) emotions. For instance, it may become possible to differentiate ‘primary-level’ emotions – such as fear in response to an immediate physical threat, as emphasised by biological/evolutionary accounts⁷¹ – from more complex, ‘high-level’ emotions, such as fear of professional failure or of disappointing one’s family through personal choices (Q. Wang et al. 2022).

The capacity to distinguish between these types of emotions could enhance the predictive power of ERTs. Returning to the case of fear, primary-level fear is typically associated with strong avoidance tendencies and immediate action-readiness in the physical domain. In contrast, high-level fears often do not translate into such action tendencies and may even inhibit physical withdrawal. Recognising this distinction in between levels of a given emotion type is thus crucial for any account of emotion that aims to track the relation between felt emotion, action tendency, and social context.

⁷¹ These are often compared to almost ‘reflex-like’ types of emotions.

Moving forward, I keep open the possibility of relying on cluster-based, or other promising, approaches for ERTs. I take emotions to exhibit variability in their physiological expressions, depending on context, individual history, and social meaning. Nevertheless, I maintain that reliable patterns – especially at the level of clustered physiological changes – are likely to exist and can be meaningfully identified. These patterns, while neither universal nor exceptionless, may support robust predictions about the kinds of behaviours and concerns emotions are tied to in particular contexts.⁷² This makes evident the necessity of subjecting ERTs to critical ethical scrutiny, particularly, as I will argue in Chapters 5 and 6, regarding privacy. Tracking the emotional life does not merely amount to registering surface affective states but rather also implies the scrutiny of intimate information about a person’s evaluative stance, motivational tendencies, emotional vulnerabilities, and more, in a way and at a scale that was not even conceivable before. By tracking the affective life, ERTs risk exposing these dimensions of the self that are typically protected from unsolicited access.

3.5.4. Emotions as performative acts

Finally, a last noteworthy relation is found in emotions and performative acts. Some have argued that emotions *are* performative acts in the sense that they *do* things. Solomon argues that “[a]s a general rule, we may say that an emotional expression is action demanded by the logic and ideology of the emotion and particular circumstances” (Solomon 1976: 165). On this view, to express an emotion is to engage in a form of action shaped by the evaluative structure of the emotion itself and the specific context in which it arises. De Sousa advances a complementary account, emphasising the communicative and interpersonal functions of emotional expressions. He maintains that such emotional expressions act as communicative gestures that disclose one’s evaluative stance towards a situation or agent, thereby enabling socially coordinated action and mutual understanding (de Sousa 1987: 191).

External manifestations of emotions are thus embedded in practices of social regulation, negotiation, and recognition. As the next chapter will explore emotional expressions in greater depth, I will only note for now that emotions are not only intrapersonal phenomena but also, and maybe even mostly, interpersonal ones. In close relationships, emotions can appear strikingly transparent, so much so that they often impose themselves with such clarity that denial

⁷² I return to this point in Chapter 4.

or misinterpretation seems implausible. At times, expressions of emotion communicate more powerfully than any concurrent verbal content. When a person's emotion contradicts their spoken words, it can feel as though their verbal message is being actively undermined or invalidated. It is as if they were simultaneously affirming and denying the same claim.

In this sense, emotions, in their expressive form, function as performative acts: their occurrence enacts a specific social meaning or stance, independent of (and sometimes in conflict with) explicit linguistic content. When in contradiction, the asymmetry between emotional and verbal communication introduces a power dynamic that can significantly alter the meaning of the information conveyed and shape the trajectory of an interaction.

As shown throughout this chapter, emotions are best understood as biologically grounded, cognitively structured, and socially intelligible. Yet they do not all serve the same functions, nor can they be analysed at a single level of explanation. Some emotions, such as guilt, shame, indignation, or pride are central to sustaining social cohesion and moral accountability. Others, such as anger, fear, or compassion coordinate more immediate behavioural tendencies. Even within a single emotion-type, such as fear, fine-grained distinctions must sometimes be drawn. This variability reflects that emotions are multifaceted, context-sensitive, and plastic, meaning that they are subject to modulation, training, and strategic social management.⁷³

3.6. Conclusion

In this chapter, I have outlined a function-first and integrative perspective on emotions to better understand the relations between the highly complex and meaning-laden concepts of emotions, values, moral salience, judgements, decisions, and actions. Given the complexity of emotions – rich, multifaceted, and still only partially understood, with no consensus on their definition – I adopted such a function-first integrative view on emotions to prioritise an overview of what emotions do and what relations they maintain over what they are.⁷⁴ My focus has been on examining (i) their foundational role in morality and reasoning processes through diverse readings of classical accounts of morality and reason (Plato, Aristotle, Hume, Kant), and

⁷³ However, this also means that reliable identification will be especially difficult, even with new methods.

⁷⁴ The function-first integrative perspective is methodologically agnostic about the nature of emotions, focusing on functions instead. That said, recognising certain functions may imply that I reject some ontological views on the nature of emotions (for example, the classical rationalist account).

contemporary approaches (Greenspan, Nussbaum, de Sousa, Tappolet, Ben-Ze'ev), (ii) their role in intrapersonal life, and (iii) the functions they serve for both the self and others.

I began by analysing canonical philosophical accounts of emotion and examined their treatment of the relations between emotions and morality. This analysis underscored how the concepts of emotion, moral salience, and values are deeply intertwined and how emotions are an essential part of moral reasoning and moral actions – even within the accounts that explicitly tried to separate emotions from moral reasoning. Building on this, I expanded the discussion to include transdisciplinary insights from contemporary work in philosophy, psychology, affective neuroscience, and anthropology. To structure this broad field, I organised these contributions into three broadly conceived perspectives: biological/evolutionary, cognitivist, and social constructionist. While these perspectives are not always fully reconcilable, many recent accounts adopt hybrid or moderated positions that allow for a more compelling understanding of emotions in functional terms.

I then examined some key characteristics of emotions in relation to their expressions: action tendencies, physiological changes, and performative potential. Emotions, particularly through their anticipatory evaluations of future courses of events, are closely linked to specific action-types. However, identifying or predicting these action types is notoriously difficult, as they vary across individuals, contexts, and other situational variables. Complicating matters further, emotions rarely occur in isolation: they often appear in dynamic, overlapping forms, where the individual experiences and reflects upon multiple emotions at once, thereby blurring potential connections between a single emotion and a clear behavioural outcome. Regarding their related physiological changes, I argued that patterns of clustered physiological changes may exist, at least for certain emotion types – though these will also inevitably vary across individuals and contexts. While current methods fall short of capturing such patterns reliably, future developments may yield robust predictions by relying on them. This possibility makes it all the more urgent to subject ERTs to rigorous epistemological, ethical, and normative scrutiny. Finally, I briefly addressed the relation between emotions and performative acts. Emotions, when expressed, do not merely reveal internal states; they also often iterate something through their expressive correlate, carrying interpersonal meaning, and shaping social reality.

This chapter has focused primarily on the intrapersonal level, examining expressions of emotions only insofar as they relate to emotions themselves. I have thus not yet addressed one

of the most crucial dimensions of expression: their social functions. The relationship between expression and emotion is only the beginning of the story. In the next chapter, I will shift focus to emotional expression proper – moving from the intrapersonal to the interpersonal domain.

4. Emotional expressions in technosocial contexts: Rethinking functions and modes of recognition⁷⁵

4.1. Introduction

Chapter 3 examined the intrapersonal dimensions of emotional life, focusing on the functions of emotions and their relations to values, moral salience, judgements, decisions, and actions. This chapter shifts the focus outward, from the intrapersonal to the interpersonal dimension of emotional life. This is a necessary shift for a proper investigation of ERTs because while ERTs' ultimate aim is to access emotions, they can only attempt to do so through external means, viz., emotional expressions. Accordingly, this chapter examines how emotional expressions render affective life socially intelligible, and what epistemic and functional commitments this intelligibility presupposes. Building on these insights, it then proposes a way to understand emotional expressions that accommodates ERTs' uptake in technosocial contexts.

The chapter first advances two core claims. First, nonverbal affective communication constitutes an important channel of everyday social interaction. Second, emotional expressions play this role because they function as defeasible yet generally reliable sources of information on which emoters and perceivers rely to coordinate action, manage expectations, and regulate social relations. Emotional expressions are thus integral components of the informational infrastructure of social life.

If emotional expressions are socially central and relied upon in this way, then technologies that monitor and classify them do not merely observe affective behaviour; they intervene directly in established practices of affective coordination. Understanding how such intervention occurs, however, requires a more fine-grained account of what emotional expressions are taken to be in technosocial contexts.

This is the central task of §4.4. Building on the insights gathered in the preceding sections, §4.4 argues that emotional expressions are best understood as serving two different purposes for two different types of recipients in technosocial environments. For humans, emotional expressions function as communicative tools that enable social coordination, negotiation, and norm regulation. For ERTs, by contrast, the very same expressions are treated as diagnostic data points used to infer, classify, and predict inner affective states. This split in their functions is

⁷⁵ Parts of this Chapter are published in Prigent (Forthcoming b).

best understood as serving two distinct regimes of epistemic uptake (i.e. human and machine). §4.4 sketches the outlines how an account of emotional expressions may make sense of this functional split.

The chapter proceeds as follows. **Section 4.2** examines the nature of emotional expressions, how they relate to emotions yet also exceed them in their normative, communicative, and social functions, and maps the principal ontological debates surrounding them. **Section 4.3** situates emotional expressions within affective communication channels, focusing on probabilistic-pragmatic accounts to understand their meaningfulness and reliability in social interaction. **Section 4.4** then sketches a revised design-based account on which it is possible to ground the proposition of a bifunctionality of emotional expressions in technosocial contexts. The chapter concludes by summarising how this framework reframes the ethical stakes of ERT-based affective monitoring.

4.2. Theoretical grounding of emotional expressions

Emotional expressions (EEs) have long been regarded as direct bodily manifestations of one's inner affective states, hence their close association with emotions. Accordingly, they are often reduced to the mere “expressive component” of emotions. While this role alone is complex, EEs are much more than that, and this is rendered especially clear in technosocial contexts. In this section, I first discuss emotional expressions, identifying their main properties and causes, and situate some prototypical types of EEs within a broad spectrum. I then discuss some theories of emotional expressions and their respective view on the phenomenon.

4.2.1. Core properties of emotional expressions

This subsection identifies the main properties that accounts of emotional expressions typically acknowledge. Emotional expressions are highly heterogeneous phenomena and may take, among other things, verbal or nonverbal forms. Philosophers have dedicated considerable efforts to studying EEs through language (Peirce 1934; Wittgenstein 1953, 1980; Austin 1962; Searle 1979). On this interpretation, “emotional expressions” typically refer to emotionally charged words and sentences, where one may say to another “I’m angry”, “you are jealous”, “I hate you”, or even “what you are doing is wrong”. In this dissertation, however, I restrict my analysis to *nonverbal expressions*. Nonverbal emotional expressions can occur through different modalities, such as facial movement, bodily posture, gesture, tone of voice, but also

physiological changes such as blushing, sweating, or trembling. Emotional expressions typically vary in their dynamics, intensity, and duration. Some unfold and fade rapidly, while others develop more slowly or even tend to persist across time (Hess & Fischer 2013; van Kleef 2016).

Within the nonverbal paradigm, a core property of EEs concerns the degree of control involved in their production. This is a property that was already discussed in Chapter 3 regarding the notion of control one can reasonably hope to have over one's emotions in the context of moral life. Aristotle, for instance, believed that the fact that one gets angry about x when x happens is outside of one's control (Aristotle 1925: *NE* 1 105b28 and 1 106b9). However, one can *decide* whether one stays angry at x or not. As regards emotional expressions, the idea of control over one's emotional expressions is similar, and the answer to whether one can have reasonable control over one's EEs varies drastically depending on the definition of EEs one adopts. While emotional expressions are complex and cannot be fully grasped by simple dichotomies, it is nevertheless the case that affective researchers typically distinguish between two general types of EEs: spontaneous EEs and intentional EEs (and some will argue that the latter are not EEs at all).⁷⁶ For example, a person may smile spontaneously when feeling happy, but may also smile intentionally (e.g. when a busy mother still smiles to encourage her toddler to take a few more steps). Spontaneous EEs are generally considered involuntarily produced; therefore, the degree of control one can hope to reasonably have over them is minimal, if not null. Intentional EEs are voluntarily produced by the emoter. Consequently, the degree of control one can hope to reasonably have over them is usually considered high, with some exceptions, such as when an emotional overload defeats active emotional labour.⁷⁷

Another important (and related) concept in the literature on EEs is sincerity. In scholarship on emotional expressions, the "sincerity condition" is a property that typically tracks the relation between inner emotional states and their related emotional expressions.⁷⁸ Because spontaneous EEs are involuntarily produced, they are typically treated as sincere EEs, meaning that they are considered *genuine* outward expressions of their corresponding inner emotional state. Within the range of intentional EEs, depending on the adopted view, some may be considered sincere while others are insincere. Within the views that do consider that intentional EEs may have the property of being sincere/insincere (see Green 2007; van Kleef 2016; Scarantino 2017), the

⁷⁶ I will discuss this in more details in the remainder of the chapter.

⁷⁷ See §5.4.1 and §5.5.2 for a discussion of emotional labour. See also Hochschild (1983).

⁷⁸ This term has been borrowed from work in philosophy of language and can be notably found in the work of J.L. Austin (1962) and J.R. Searle (1979). See also Green (2007) and Scarantino (2017, 2019) for a discussion of the conceptual elements that can and cannot be reused from pragmatics in theories of emotional expressions.

threshold for sincerity is usually put where the intentional EE tracks the affective intention or inner emotion. For instance, I may be happy and *voluntarily* smile at a friend – here, the expression may be considered both intentional and sincere. In contrast, I may smile at someone despite not being happy to see her, and in fact wishing she had not come. In this case, the expression is typically considered intentional and insincere. Seen this way, intentional emotional expressions that are deemed sincere are overtly expressed in a way that enables them to successfully convey the information that the emoter wants to share.

Finally, another important aspect is that it seems that (some) EEs may be shaped by social or cultural norms or expectations (Hess & Hareli 2019a; Bar-On 2004). For instance, when Germans wave their right hand in front of their face to express their shock at how asinine a person or a situation is. Or when Italians gesture two pinecones with their hands waving up and down to express frustration at a person or a situation. However, culture is not the only relevant factor for these types of EEs, as they can also be created on what seems to be simply an ad hoc basis. For instance, one could see friends endorsing a particular behaviour for a particular emotion, such as joining and intertwining their thumbs together, pointing the rest of their fingers upwards, and the palms of their hands on their chest, to express compassion. These are examples of emotional expressions that are *learnt social behaviours*. They may be sincere or insincere, and arguably spontaneous or intentionally produced.⁷⁹

It is important to note that while these properties of EEs are typically discussed, not every theory of EEs will recognise all their variants. For instance, regarding the involuntarily/voluntarily dichotomy mentioned above (also referred to as spontaneous/intentional), certain views, such as the Basic Emotion View (see Chapter 2; Ekman 1992), or as noted by Mitchell Green (2007: 108), the Neurocultural View, hold EEs to be only *and* purely involuntary. Regarding the sincerity condition, Alan J. Fridlund and classical proponents of the Behavioural Ecology View reject the idea of an emotional expression being *sincere*. They instead argue that (facial) emotional expressions are only and always strategically produced messages in the sense that they “produce a signaling “ecology” that is analogous to the balance of resources and consumers, and predator and prey, that characterise all natural ecosystems” (Fridlund 1994: 128). Similarly to what has been discussed regarding emotions in

⁷⁹ This is so because one could argue that some socially constructed expressions, once profoundly embedded in one’s emotional expressive apparatus, might come “spontaneously” in a similar way as a smile or a frown may be spontaneously expressed (see Bar-On 2010) . For my account, I will classify these as semi-natural EEs. See Fig. 2 in §4.2.3.

Chapter 3 §4, theories of emotional expressions are also divided about whether emotional expressions may be said to be biologically caused or socially constructed. However, and as we will see in this chapter, most contemporary scholars will endorse hybrid accounts that (at least) allow for the idea that EEs may be both biologically rooted and socially mediated.

Thus, to put it in terms of opposites, EEs appear to have the potential to be: (1a) involuntarily produced or (1b) voluntarily produced; (2a) sincere or (2b) insincere; and (3a) biologically caused or (3b) socially constructed. These attributes raise challenging identification issues. For example, should frowning due to bright sunlight be considered an expression of anger if the facial configuration resembles that of genuine anger? If not, why not? Is blushing a reliable sign of embarrassment? Possibly, but how should we account for its occurrence in other contexts, such as love, shyness, or elevated body temperature? And what about someone who “fakes” an emotion by deliberately producing the prototypical expressive behaviours and physiological changes associated with typically having that emotion, despite lacking the corresponding inner state, such as actors successfully pretending to have their heart broken? These examples illustrate why formulating a general definition of EEs and their functions is so complex. While the next sections will clarify my stance on these issues, here I position myself within the broad view that embraces these dichotomies as offering ways of classifying different types of emotional expressions rather than as criteria for rejecting or excluding some expressions from counting as “emotional expression”. Now that their main properties have been discussed, the next section turns to their causes.

4.2.2. Causes of emotional expressions

Another important feature of emotional expressions concerns their causes. While the ties between emotions and emotional expressions have long been at the forefront of the research agenda, this assumption is too narrow for the present purpose which is to investigate the different functions of emotional expressions in technosocial contexts. Emotional expressions may indeed be caused by emotions, but they may also be shaped by socially situated motives (hereafter social motives), including intentions, relational goals, role obligations, and aspirational commitments.⁸⁰ Additionally, it has also been posited that physical state may sometimes be the

⁸⁰ I also include aspirational commitments in social motives because much of one’s moral and relational affective communication is aspirational. Individuals express forgiveness they are still learning to inhabit, commit to patience they know will be tested, or signal trust they hope to build. In functional terms, these expressions project a stance the individual is attempting to realise; they publicly initiate new normative expectations, both for oneself and for others, even in cases when inner emotional states that would ideally be co-present are not fully aligned yet.

cause of an emotional expression. This can be the case when physical states are tied to the dual pain-pleasure scheme; there is an agreeableness in feeling the sun on our skin or in entering a warm bath, and there is pain in putting our hands on a hot stove, lacking air, or shivering from the cold. While I do not exclude emotional expressions whose cause is a physical state from the domain of emotional expression as a whole, the remainder of this work focuses on EEs who are shaped by either emotions or social motives.

To sum up, by emotions, I mean purely affective states, namely inner affective states that are *reactional* and non-intentional states.⁸¹ They are affective reactions to internal introspections or changes in the external world. By social motives, I refer broadly to motivational/intentional states that are linked to related goals or desires. Importantly, while they are not affective reactions, they are *affectively loaded*. As explained in Chapter 3, motivation and intention are mental states intertwined with affective states, and they depend on them for their actualisation. However, in this project, I understand motivational and intentional states in functional terms, and see them as linked to emotional expressions insofar as they operate as “social motives”. The emotional expressions may thus be the channel used to obtain a certain state of affairs linked to a certain social motive, and this type of cases needs to be separated from the cases where emotional expressions are the outward expressions of emotions.⁸² Importantly, both types fall within the scope of this analysis because they reveal how EEs function in affective communication channels, linking the intricacies of affective processes and social purposes in the regulation of interpersonal life. Finally, by emotional expressions, I aim to include all physiological and non-verbal behavioural routes that may reliably lead to a person’s emotions (which means possibly also to their values, and likely future behaviours),⁸³ and/or social motives (sincere and insincere kinds), emotional expressions. Understood this way, EEs are one of the

⁸¹ Note that here, by “non-intentional”, I mean that emoters *do not have control* over the emotions that they feel. I do *not* mean that emotions do *not* have intentionality, in the sense of “aboutness”, meaning that emotions are directed at or about something in the world.

⁸² The view presented here is simplifying the relation between emotions and social motives so as to be in a position to focus on the functions of EEs. However, note that the relation between emotions and social motives is extremely complex and can be understood as complementary rather than exclusive. Emotional episodes may give rise to social motives, and conversely, social motives can elicit or modulate emotional responses. For instance, anger as an affective reaction may trigger a motive to confront, while a prior social motive to restore fairness can amplify or shape anger’s expression. But this does not mean that this co-dependency is always present. There can be EEs without a clearly articulated social motive – such as spontaneous reactions like surprise or disgust (inner emotional states) – and conversely, EEs may arise from social motives without being tied to an immediate inner emotional state – for example, a reassuring smile or a polite display of concern. However, I leave these intricacies aside since what is important for this project is that in both cases EEs function as the channel through which affective and motivational states are conveyed.

⁸³ See Chapter 3.

primary ways to interact with others, develop relationships, and shape social communication and regulation.

4.2.3. Spectrum of emotional expressions

From the properties and causes of EEs discussed above it is possible to map the different types of EEs along a spectrum (see Fig. 2). At one end of the spectrum, there are the “natural emotional expressions”, which are, from the biological/evolutionary insights, the EEs that are spontaneous and involuntary. These are usually the ones that are said to “reveal” to observers relevant information about an emoter’s inner affective state.⁸⁴

In the middle-left, we find “semi-natural EEs”. They are expressions that, as Dorit Bar-On puts it, are acquired expressions that become ‘second nature’ for an individual (Bar-On 2010: 213). I will refer to these as “semi-natural EEs” as even though they are learned expressions, once acquired and internalised, they can be expressed spontaneously and involuntarily, in a similar way to the “natural” ones. Because they are both involuntary, natural and semi-natural EEs may be said to “leak” affective information which might be picked up as meaningful and informative by perceivers.⁸⁵

“Semi-conventional” types, in the middle-right, are behaviours or gestures that are intentional and voluntary, even though they only *mean* something to a small, restricted group of people. They might appear on what looks like an ad hoc basis, *independent* of dominant norms, values, or entrenched social causes that would have supported their development.

Finally, at the opposite end of the spectrum, there would be the “conventional EEs”, which would be conventional nonverbal emotional expressions. They are also typically considered intentional and voluntary but contrary to semi-conventional EEs they are *dependent* of dominant norms, values, or entrenched social causes that would have supported their development.

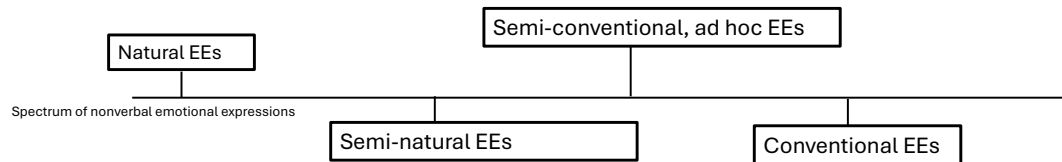
These four general types of EEs do *not* cover the full range of EEs, and the spectrum should also allow for other kinds of emotional expression, such as idiosyncratic expressions that are

⁸⁴ Natural EEs are typically said to make our mind known to others by *showing* the related affective content perceptible to observers.

⁸⁵ I come back to this point in Chapter 5.

not necessarily audience-targeted⁸⁶ but instead reflect an individual's own expressiveness. However, these four types constitute the bulk of the spectrum.

Figure 2. Spectrum of Nonverbal EEs



Despite intrinsic differences between the different types of EEs, they are all constituted by expressive behaviours and/or physiological changes. By referring to “behaviours” and “physiological changes”, I also include ones that *may not* be perceptible without access to various technologies, such as galvanic skin response, heartbeat, and the brain’s electrical signals. An important point to keep in mind, is that emotional expressions are about one’s *affective* life rather than merely about one’s *inner* life. Emotional expressions are deeply social and as such do not merely represent ‘inner’ emotional phenomena but the entire social and personal realm.

Below, I bring some nuances to this picture by examining multiple different approaches to defining what EEs are and to what extent they may qualify as sincere or insincere, spontaneous or intentional, and biologically caused or socially constructed.

4.2.4. Debates and ontological commitments

In affective research, debates around what EEs are still remain, leading to various different views. Understanding the mainstream views will be helpful to map the diverse functions of EEs within affective communication channels later. The mainstream views can be organised around two overarching families of views. The first group (Group A) treats emotional expressions as caused or part of emotions themselves. The second group (Group B) expands the reach of EEs from emotions to also include social motives. It treats EEs as elements of affective communication whose identity is set by the roles they play in coordinating interaction, managing expectations, and making affective information publicly available. In what follows I state the ontological commitments for each group and review some of their respective variants. By

⁸⁶ Bar-On (2010) gives the example of someone privately and spontaneously caressing a photo.

ontological commitments, I mean what each view claims EEs are and what makes them count as such.

Regarding Group A, as noted by Rebecca Rowson (2024), it is useful to recognise a clear split between accounts that define the nature of EEs relative to features that are *independent* of the perceiver's judgement (subgroup A1, referred to as causal-type accounts), and accounts that attempt the same aim but by holding a *strict relation* between EEs and perceivers' judgements (subgroup A2, referred to as perceiver-dependent accounts).⁸⁷ In the section below, I follow Rowson's (2024) proposed division of accounts.

Causal-type accounts (subgroup A1) typically define EEs by appeal to features that are independent of the perceiver's judgement. These accounts ground the nature of EEs in causal relations (either to underlying affective states or to historically stabilised functions) rather than in how expressions are interpreted by observers. On this view, expressions are bodily or behavioural events whose identity is fixed by their causal role, broadly construed. This general view can express the causal link in various ways, such as EEs being constitutive parts of the emotion (Part-whole accounts), by species-typical pairings with basic emotions (Natural-type accounts), or by a designed function to reveal emotion (Design-type accounts). Part-Whole accounts, for instance, consider EEs as components of emotions among other components (Parkinson & Colman 1995; Scherer 2005; Smith 2017). Hence smiling is a (physical) component of joy, frowning is a (physical) component of anger, crying is a (physical) component of sadness, etc. They thus usually consider that being happy for instance, would induce a smile, being sad would induce tears, etc (Davis 1988; Parrott 2017; Smith 2017). Natural-type accounts hold that there exist a pre-defined set of basic emotions that is universally present in humans, and this set of basic emotions has a pre-defined set of corresponding expressions, which have naturally developed to express this emotion, in a kind of one-to-one mapping (Ekman & Friesen 1971; Ekman 1992). Design-type accounts (Griffiths 1997; Green 2007)⁸⁸ usually posit that EEs are designed to communicate our emotions to others. Hence, standard design-type accounts usually propose some reasons regarding how EEs have stabilised in the way they have. The reasons often imply an account of how a particular EE transmits important information to perceivers, usually split between survival purposes (communicating fear by widening the eyes may help

⁸⁷ Note that some authors referenced here have hybrid accounts that bridge between either subtype of groups such as A1 and A2 (e.g. Griffiths 1997) or groups such as A and B (e.g. Scherer 2005; Bar-On 2004).

⁸⁸ Green's account (2007) is more complex as he distinguishes between signals and cues. On his view, only signals may fit the design-type account of emotional expressions.

others in the group react more quickly and escape the threat) and social coordination purposes (an EE of sadness may be designed to elicit care in others). Design-type accounts diverge from other causal-type accounts by grounding the identity of emotional expressions in their communicative function rather than in a fixed emotion-expression pairing.

Due to their ontological commitments, accounts from Group A1 typically struggle with intentional or motive-driven expressions that occur without or independently from the presupposed associated emotion. Design-type accounts constitute an important exception in this respect. Because they identify emotional expressions by their communicative function rather than by a fixed causal relation to a particular emotion, they can accommodate intentional and motive-driven expressions. On such accounts, a smile need not be caused by genuine joy in order to count as an emotional expression; rather, it is characterised by its role in regulating social interaction, for instance by signalling affiliative intent, appeasement, or non-hostility.

In perceiver-dependent accounts (subgroup A2), accounts typically hold that “behaviour expresses emotion only insofar as [...] it produces particular experiences in observers” (Rowson 2024: 10). Dispositional and Perceptual accounts, for instance, are examples of perceiver-dependent account-types as they both rely on some perceiver-dependent requirements for a sequence of bodily changes to be recognised as an expression of emotion. Dispositional accounts typically hold that a particular EE expresses a particular emotion if and only if that EE is *disposed to appear to express* that particular emotion to average perceivers in standard conditions (Rowson 2024).⁸⁹ This also entails that they recognise a certain sequence of bodily changes (or a certain behaviour) as an emotional expression if the receiver recognises the sequence as *representative* of a particular emotion. They also allow, similarly to the design accounts, for some contextual flexibility. For instance, Dispositional accounts may consider and take into account the receiver’s background knowledge and the context. If one sees their own brother and recognises from their experience of him being worried that he appears worried, under Dispositional accounts, their brother *is expressing* worry. An expression is thus what it is partly in virtue of how it is recognised by suitable observers in suitable contexts.

Perceptual-type accounts (Taylor 1980; Glazer 2017a) claim that if a particular behaviour can single-handedly enable the perception of a given emotion, then it is an emotional expression of

⁸⁹ “Average perceivers” and “standard conditions” are only meant to rule out situations where a particular perceiver would not be able to recognise an emotional expression for a particular reason tied to her own circumstances, and where that EE would have been commonly recognised otherwise (i.e. the sun is in their face or they are blind).

that emotion. They thus rely on a perceptibility requirement: if my cheeks blushing reveals to you my embarrassment, it is an emotional expression of my embarrassment. But on the contrary, if my heart beating faster because I am angry does not enable you to perceive my anger, because you cannot see (perceive) my heart beating faster, than it is not an expression of my emotion. The expressive status thus depends on *successful* perception by an observer. Hence, an expression count as one only if it is picked up by the right kind of uptake.

Perceiver-dependent views typically better account for intentionality and context sensitivity than causal-type accounts do but they risk falling into a relativist trap about what an expression is, and risk transferring authority from emoters to any competent perceivers, making it impossible to “objectively” determine whether something is an emotional expression or not.⁹⁰

Having outlined the strengths and limitations of Group A (A1 and A2), I now move to Group B. Accounts from Group B often rely partly on core claims from Group A, but they shift the focus away from investigating the nature of the relation between emotions and emotional expressions and instead analyse EEs as patterned behaviours that function within affective communication channels, enabling individuals to convey information about emotions, social motives, and/or physical states (Bar-On 2004; Fischer, Pauw, & Manstead 2019; van Kleef 2009, 2016; Hess & Hareli 2019a).⁹¹ EEs are thus regarded as carriers of affective meanings, which operate within affective communication systems, and whose grounding can be biological, social, or mixed. What unites these forms is their participation in a socially recognisable affective repertoire and their capacity to make affective information publicly available.

Within Group B, two main subgroups can be distinguished as well. I will refer to them as subgroup B1 and subgroup B2. Subgroup B1 considers that an EE exists by virtue of its role in coordinating individuals and making motives, intentions, and action tendencies publicly available (Frijda 1986; Fischer & Manstead 2016). Emotion-based expressions (e.g. fear or anger displays) are often biologically grounded, emerging from patterns of action readiness (Frijda 1986), whereas (social) motive-based expressions (e.g. appeasement smiles, proud postures) are socially grounded, emerging, stabilising, and sustained by learned conventions and normative

⁹⁰ By this I mean that if multiple “suitable” people arrive at different conclusions about one’s emotion, then it is impossible to discriminate or eliminate some of the conclusions until there is only one left. This is because if the expressed behaviour appears to express anger to observer A and embarrassment to observer B, then it is expressing anger and embarrassment to the respective observers. In the next section, I will expand on why this becomes highly problematic in technosocial contexts involving ERTs.

⁹¹ Physical reactions with affective valence, such as pain or startle, may enter the expressive category when they are socially interpretable as carriers of affective meaning (Hess & Hareli 2019b).

expectations (Hess & Hareli 2019a). In subgroup B1, the kindhood of an expression is therefore determined by its function within normative contexts, not by its causal link to an emotion like in most accounts of subgroup A1 for instance. Uptake by perceivers is required for coordination but does *not* constitute the existence of the expressive kind, like in accounts from subgroup A2. These approaches emphasise the mutual shaping of expression where social appraisal processes determine what expressions are appropriate or effective in a given context (Hess & Hareli 2019b; Fischer & Manstead 2016).

Finally, subgroup B2 encompasses expressive-pragmatic theories, including affective-pragmatic, neo-expressivist, and semiotic-probabilistic account-types. These approaches go a step further by sharing the view that EEs are pragmatic actions within communicative practices. On this view, to express is to *do* something with EEs. Expressions are a kind of action-type operating within an affective repertoire. A behaviour counts as an EE when it carries affective meaning and fulfils a communicative or regulatory role, regardless of whether it is caused by a felt emotion. What an EE counts as is fixed by the role it plays in the communicative practice and by the practice-related felicity conditions (Scarantino 2017). On Scarantino's Theory of Affective Pragmatics, for instance, they can be intentionally produced, modulated, or withheld to influence others (Scarantino 2017, 2019). EEs are pragmatic actions within affective communicative practices: to express is to do something with bodily behaviour. Their capacity to communicate emotion is secured by causal regularities that make them probabilistically informative about affective or motivational states (Scarantino 2019). Neo-expressivist accounts such as Bar-On (2004)⁹² presuppose a continuity with nonverbal manifestations such as crying or smiling, both of which are direct manifestations rather than descriptive reports of affective states. While some expressive forms draw on biological affordances, such as the ease with which it is possible to produce certain facial or vocal configurations, their stability across contexts depends on social learning and convention (Bar-On 2004). Semiotic-probabilistic account-types, such as the one developed by Bonard (2021), claim that expressions convey graded and defeasible affective information, with typical causal regularities providing probabilistic grounding and pragmatic use in context-fixing meaning. These expressive-pragmatic approaches typically understand EEs as communicative acts that make affective life publicly intelligible through patterns of use, meaning, and coordination.

⁹² Though her model is primarily concerned with verbal avowals, nonverbal expressions are also covered as pragmatic actions.

While this subsection has focused on the ontological commitments of the main accounts of EEs, the discussion of Group B has already led us to the epistemic and social dimensions of EEs. Theories gathered under Group B treat EEs not only as things that are, but as things that do. The next section continues in that vein and examines how EEs transmit, regulate, and structure the flow of affective information across social interactions. This is necessary to understand how affective life is shared, coordinated, and rendered intelligible within interpersonal and technological environments.

4.3. Affective communication channels

Having set out the main ontological commitments of different account-types of EEs, I now turn to their communicative dimensions. EEs circulate within what may be called affective communication channels, which are best understood as dynamic channels in which the affective life is expressed and relied upon to shape social life through expectations and coordination. Exploring how these channels operate allows a fuller grasp of how EEs acquire social intelligibility and how, in turn, social structures depend upon the regulated exchange of this expressive information.

4.3.1. Emotional expressions' functions in affective communication channels

Within affective communication channels, EEs are not only considered valuable indicators of a person's inner emotional state but also seen as signals of social motives – viz., as signals of motivations, intentions, or attitudes towards given goals or desires – and as predictors of future behaviour (van Kleef 2009, 2016; Scarantino 2017, 2019; Fischer, Pauw, & Manstead 2019). In that sense, (nonverbal) EEs are not peripheral or occasional signals, rather, they constitute a primary mode of human communication. They perform various functions that continuously shape, negotiate, and transform social life. They are richly structured, frequently deployed in everyday interactions, and crucial for social coordination. Like gestures or language, they function in patterned and intelligible ways across contexts, cultures, and relationships. In what follows, I frame EEs as core components of our communicative ecology.

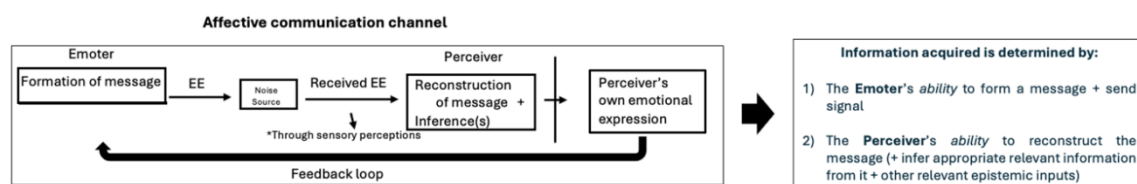
Here I move beyond the emoter and their EEs and enter the social realm of communication and interaction. To analyse EEs' communication, I adapt Shannon's (1948) model of information transmission, which originally focused on the encoding, transmission, and decoding

of messages in a technical system.⁹³ While Shannon’s model is limited in its capacity to account for meaning and intentionality, it offers a useful heuristic for mapping the underlying communicative structure of affective communication channels and the transmission of affective signals within them.

To this end, I propose a schema of affective communication (see Fig. 3) which shifts the perspective from a purely internalist or individualist conception of expression and recasts EEs as socially embedded acts that gain meaning within interactional contexts. This makes them not only an important mode of communication in daily life, but also a promising target for the monitoring of affective life by ERTs.

Figure 3 reframes Shannon’s basic communication model (1948) and applies it to the affective communication channels constituted by EEs, in order to highlight key stages in the flow of information.⁹⁴ The figure shows how, in a typical interaction, the emoter performs two tasks: producing a message (affective information) and sending a signal (emotional expression). Likewise, the perceiver performs two tasks: receiving the signal (perceiving the EE) and reconstructing the message (forming a belief about the meaning of the received affective information).

Figure 3. Standard Affective Communication Channel



In this schema, the emoter transmits information by producing an EE, which carries affective information through the affective communication channel. The perceiver – via sensory perception (e.g. visual or auditory) – decodes the signal and interprets its meaning. The perceiver

⁹³ Also heavily present in work on animal communication, which I set aside, even though I do recognise that much of the research on humans and machines has been inspired by studies on animal emotion and communication (see Panksepp 1998; Snowdon 2003).

⁹⁴ Shannon’s framework was originally meant to establish the basic elements of a communication channel, to later find the point where the communication of a signal is put in peril by the presence of noise in the channel. Hence, Shannon’s goal was to find what is now known as the signal-to-noise ratio, as well as the lowest possible signal-to-noise ratio, which is the cut-off point where the communication of the signal is successful, before being overtaken by the noise (see Shannon 1948).

thus forms a belief about the content of the message, based (in part) on the perception of the emoter's emotional expression.

I am not concerned here with whether perceptual experiences are *truly* veridical⁹⁵ but with the (practical) epistemic norm that we should – and in practice do – treat them *as* veridical unless proven otherwise. In that sense, the epistemic dimension of EEs does not require deciding the issue of (non)veridicality of EEs but rather understanding how perceivers' reliance on EEs works, practically. Hence, I am not concerned about their veridical status per se but about their *correctness* from a perceiver's point of view.

It is important to note that perceptions rarely operate in isolation in the context of affective information. They are typically accompanied by other sources of information – such as memory, contextual information, social reactions, background knowledge, or prevailing norms – which jointly contribute to the perceiver's overall assessment of a situation. Hence, EEs are both produced and perceived within social contexts and EEs' intelligibility (for perceivers) often depends on the perceivers' perceptions being supported by other epistemic inputs.

Moreover, both production and perception involve feedback loops. These mechanisms can help confirm or refute our perceptions and the inferences drawn from them, and they can in turn shape how emoter's EEs are subsequently produced thereafter (Van Kleef 2009). Accordingly, when legitimate reasons arise to doubt or revise a perceptual judgement (e.g. awareness of noise, illusions,⁹⁶ background knowledge, contextual inputs, or other reasons that contradicts the *prima facie* meaning associated with the EEs), one should correct *a posteriori* the belief that was

⁹⁵ In contrast, standard accounts of perception would question the veridicality of perceptions. According to direct realism, for instance, perceptions accurately present the world as it *truly* is. On this view, the perceiver and the perceived object stand in an unmediated relation (Snowdon 1992). In other words, when one sees a tree, one *knows* there is a tree because perception is *factive* for direct realists. Perceivers, then, have genuine epistemic access to the world as it is, which implies true knowledge of reality. At the other end of the spectrum of accounts of perception are sceptical accounts, which often invoke forms of the 'veil of perception' or brain-in-a-vat scenarios to argue that perceptions do not provide reliable access to reality. Standard accounts of perception offer little for the assessment of emotional expressions in terms of practical epistemic uptake. While direct realism may be compelling for discussions of ordinary object perception (e.g. trees, red apples, wooden tables), it is less applicable in the case of emotional expressions. This is because emotional expressions involve socially embedded meaning, whereas direct realism focuses on perceptual access to physical objects and is usually silent on the interpretive demands of affective communication. Variants of the sceptical account type are also not helpful as they treat perceptions as fundamentally unreliable or unknowable, offering no resources for explaining how emotional expressions might be reliably understood in context.

⁹⁶ If the presence of an illusion is to be found, then the perception taken as *pro tanto* veridical is rendered non-veridical. One can find the source (or cause) of an illusion that distorts human perception of reality. Categories of potential illusions are visual illusion (e.g. a *fata morgana* mirage), auditory illusion (e.g. Shepard tone), tactile illusion (e.g. phantom limb), and/or temporal illusion (e.g. telescopic effects).

constructed (at least in part) by that perception to preserve coherence with reality as it presents itself upon further scrutiny. Some of these additional elements, though, may also serve as defeaters, not of the perceptual experience itself, but of the inference drawn from it. That is, the perception remains intact, but the belief it supports may be challenged by countervailing evidence and thus be either revised or reinforced in light of these other relevant epistemic inputs. Contingent epistemic inputs thus provide a strong basis for evaluating the correctness of EEs and, when such confirmation is available, a good reason to *regard* the perceptions of emotional expressions *as pro tanto* veridical. Taking them as *pro tanto* veridical means that they are *assumed to be so unless proven otherwise*; in other words, we are justified in relying on them in the absence of defeaters.

Thus sensory perception in affective communication channels functions as *a* reason for perceivers to hold a belief x about proposition p (e.g. “If I see Anna smiling, I have a good reason to believe she is happy”). The expression is the signal, the perception is the receiver’s main tool to capture that signal, and contingent epistemic inputs often help confirm the meaning of it. Together, they constitute an affective information channel grounded in embodied, socially situated interaction.

Empirical research across social psychology and affective science substantiates this picture. For instance, we can now observe that empirical studies have, across time, consistently shown that emotional expression and emotional recognition play a crucial role in social life, helping us navigate the myriad of social relationships in which we engage daily (Van Kleef 2016; Fischer, Pauw, & Manstead 2019). EEs provide information that is reliably used by perceivers to interpret, evaluate, and respond to their social environment. Empirical studies across psychology and affective science support the view that EEs are not merely epiphenomenal signals of internal states, but are functionally informative and help perceivers make reliable inferences about others’ intentions, emotions, or likely future behaviours (Keltner & Haidt 1999; Van Kleef 2010, 2016; Scarantino 2017; Gage & Baars 2018). These inferences can regulate moment-to-moment interactions, support social cohesion or regulation, and inform appropriate responses in interpersonal contexts. For example, studies have shown that expressions of love, compassion, and care convey affective commitment and foster intimacy and cooperation in close relationships (Gonzaga et al. 2001; Ellis & Malamuth 2004). Relatedly, emotional expressions such as embarrassment and shame can signal appeasement and submission, effectively diffusing tension and reducing the risk of conflict in hierarchically structured interactions (Keltner &

Buswell 1997; Semin & Manstead 1982). Expressions of anger can rectify social misunderstandings, pinpoint norm violations, clarify expectations, or challenge status arrangements (Lerner, Goldberg, & Tetlock 1998; Averill 1982; Fischer & Roseman 2007). These findings suggest that EEs are not simply windows into inner emotional states, but rather social tools shaping interactional outcomes and allowing perceivers to act on them⁹⁷ or anticipate future outcomes (Parkinson 1996; Keltner, Haidt, & Shiota 2006; Van Kleef 2016; Fischer & Manstead 2016; Kumar & Senthil 2021). Hence, EEs serve a broad range of interpersonal and regulatory purposes.

Crucially, empirical studies also stress that the informational value of EEs is context sensitive. Numerous studies have demonstrated that the same expression can be (and should be) interpreted differently depending on relational dynamics and situational features. This means that no specific EE possesses a specific and unique meaning. For instance, the *correct interpretation* of an emoter's anger has been found to vary by context: in a parent–child setting, it may appropriately prompt a child to reflect on their behaviour; in a collaborative task, it may appropriately signal concern about failure or poor group coordination (Clark & Taraban 1991; Bourgeois & Hess 2008; Hess & Fischer 2013); in a conflict or negotiation context, an emoter's anger should be interpreted as a hardening of positions or a reduction in willingness to compromise (Van Kleef & Côté 2007; Van Doorn, Heerdink, & Van Kleef 2012).

Moreover, a growing body of research shows that individuals rely on others' emotional expressions not only to infer current affective states or social motives, but also to predict future behaviours and adjust their own responses accordingly. Emotional expressions serve a crucial regulatory function in social life by providing information that helps perceivers anticipate what others' intentions, motivations, or affective states may lead them to do in terms of future decisions or actions (Parkinson 2005; Van Kleef 2016). This predictive capacity supports social coordination by reducing uncertainty and enabling individuals to align themselves with the anticipated responses of others. For instance, in negotiation settings, whereas a correct *interpretation* of an emoter's expression of anger may be to infer that the emoter is unwilling to compromise further, a correct *anticipation* is that they will reject the offer if nothing better is done (Van Kleef, De Dreu, & Manstead 2004a, 2004b). Similarly, in interpersonal contexts, expressions of affection or compassion signal relational investment and lead the perceiver to anticipate (and expect) that the other party will show bonding behaviours and that their next

⁹⁷ For instance, if you see someone deeply sad, you may want to try to comfort them.

actions will align with this (Gonzaga et al. 2001; Van der Schalk et al. 2015). By enabling perceivers to forecast others' likely actions, emotional expressions thus serve as essential tools for social regulation and cohesion.

However, important challenges remain in the way these studies are designed (Barrett 2017), which limit the scope of their results as well as the generalisability of their findings. Nevertheless, the overarching pattern of results underscores a key point: emotional expressions are routinely used by perceivers as sources of meaningful, socially relevant information, guiding their own emotional and behavioural responses in contextually appropriate ways (Parkinson 2005; Scarantino 2017, 2019).⁹⁸ EEs thus need to be regarded as a mode of interpersonal communication that complements and often surpasses verbal language in immediacy, nuance, and embodied salience (van Kleef 2016; Fischer & Manstead 2016).

4.3.2. Meaningfulness and reliability of emotional expressions

I now turn to questions of meaningfulness and reliability of emotional expressions. I have suggested that, on the basis of the affective information carried by EEs and with support from other relevant epistemic inputs, perceivers can interpret EEs as meaningful. What kinds of meaning, then, do EEs make available, and how reliably can such meaning be extracted? These are the questions I now address.

Schlenker (2016) says that the informational content of utterances has two sources: the meaning taken from the literal words (and semantic rules) and the meaning extracted from a receiver's inferences on the speaker's motives. These two combined are often called "pragmatic meaning", and they account for what one can hope to get, in terms of information, from the content of utterances. Applied to nonverbal emotional expressions, this relates to the meaning taken from the EE itself and the meaning extracted from the inferences perceivers derive from them and other relevant epistemic inputs.

⁹⁸ Imagine a person feels anger during a work meeting after being unfairly criticised by a colleague. Their primary intention (social motive), however, is to preserve cooperative relations within the team. The spontaneous affective impulse of anger produces an action-tendency to retaliate or confront, while the social motive leads them to stay composed and to pursue affiliation behaviours. Later, the person might decide to channel that anger into a more socially acceptable form, for instance, calmly asserting their viewpoint or raising the issue later in private. In this way, the tension between inner affect and social motive performs a regulatory function and aligns individual emotional behaviour with collective expectations, enabling emotional expression to be socially shaped and to underpin social regulation and cohesion.

When one declares their love to another, something meaningful is coming from their tone of voice, the intensity of their gaze, and their rapid and nervous breath. These are often taken as cues that a recipient will typically consider when assessing whether this person is *genuinely* in love. Explicitly self-reporting one's love to another (verbally) is usually needed to confirm the receiver's inference, but rarely sufficient on its own. The utterance, by itself, is rather naked and feels empty without the *expression* to accompany it. If only from personal experience, it is easy to recall our own reliance on the emotional expressions of others. Even well into adulthood, a frown from our mother can *stop us* in our tracks, a smile from a loved one can *confirm* the robustness of our relationship, and a scream from a stranger can *petrify* us. This is because EEs constitute prevailing channels of communication, carrying reliable information on which perceivers rely to understand what is at stake, make judgements or decisions, and take future courses of actions.

To understand this phenomenon, I turn to theories of information. In philosophy, traditional theories of information – especially naturalistic ones – have typically understood information in either causal or semantic terms. For instance, Dretske (1981) defines information causally: a signal carries information about a condition only if it is reliably caused by that condition. Thus the content it carries must be true (e.g. if smoke is caused by fire, then smoke is a signal that carries natural information about fire). Floridi (2011) offers a semantic model, where information is instead seen as meaningful data that satisfies a truth-condition. Both approaches, however, share a key assumption: that for something to count as information, it must be *factive* – that is, it must be *true*. Even though this is not what Dretske or Floridi had in mind while developing their respective accounts, this maps well with subgroup A1's claims regarding the necessity of reliable causal links between emotion and EE (especially for a given EE to be considered as one). A particularly good fit between these informational approaches is to be found in (1) the theories of emotions rooted in the biological/evolutionary perspective (basic emotion theory), and (2) most of the causal-types accounts of emotional expressions (most of the subgroup A1) which treat spontaneous EEs as law-like effects of underlying inner affective states. These views would usually agree that a spontaneous EE of anger carries information about anger because it is produced by anger (or at least because the emoter *is* genuinely angry). Yet if we reserve the label “information” for strictly true signals, we risk sidelining a wide range of expressive behaviours that are socially meaningful yet defeasible. Consider a case in which someone is sincerely angry yet smiles. In such cases, these accounts struggle to explain why

genuine anger causes one to smile in this instance, whereas it causes a frown either in the same person at another time or in other individuals.

In response to the limitations of strict causal and factive accounts of information, alternative theories of information have been proposed. Millikan (2000, 2004) argues that much of the information organisms rely on is not grounded in exceptionless natural laws, but in statistically reliable correlations. Organisms “lean on” these correlations to navigate the world, even if the correlations are not universally reliable. Building on this, Scarantino and Piccinini (2010) propose a probabilistic theory of *natural information*. According to their view, a signal carries natural information about any state of affairs it reliably correlates with, relative to the perceiver’s background knowledge. This view thus rejects the factivity constraint of earlier accounts (i.e. Dretske’s and Floridi’s): what matters is not that a signal is always true, but that it raises (or not) the probability of a given state of affairs in a way that makes it plausibly (or not) reliable. This account reframes how we understand the informational role of spontaneous EEs, whose cause is an inner affective state. A spontaneous expression, like crying, may not guarantee sadness, but it increases the likelihood of sadness given alignment with appropriate justifications such as a specific context, background knowledge of that person, and/or understanding through common experience. Likewise, a grimace might suggest pain or anger depending on context. These expressions carry natural information following Scarantino and Piccinini because they enable inferences that are, more often than not, socially useful.

Still, theories of natural information cannot capture the full range of expressive meaning, since many EEs are intentional, strategic, and shaped by social norms. To account for intentionally produced types of information philosophers turn to pragmatic models of communication. Grice (1957) distinguishes between natural signs, which carry natural meaning (e.g. smoke means fire), and signals, which carry non-natural meaning through communicative intention. A signal conveys non-natural meaning when produced with the *intention* that the perceiver *recognise* that intention. On this account, a sarcastic smile, a deliberate tear, or a theatrical sigh may thus be understood as signals designed to communicate something, whether sincere or not, and do carry (non-natural) meaning, conditional on the intention being recognised by the perceiver.

Yet many EEs fall between these categories (i.e. natural and non-natural). Semi-natural and semi-conventional expressions fall in this gap. Consider expressions that may have a spontaneous, affective basis, and yet be modulated, exaggerated, or maintained for social effect. A light touch on the arm may express empathy spontaneously while drawing on normative

understanding for its meaning. An appeasement smile can combine evolved affiliative cues with learned conventions of politeness. Scholars such as Scarantino (2017, 2019), Bonard (2021), Bar-On (2004), and Green (2007), each broaden the pragmatic model of communication to accommodate this hybrid space. For them, a (semi-)intentional EE for example can convey meaningful, actionable information even without a full communicative intention and even when not recognised as such by all observers.

Scarantino's Theory of Affective Pragmatics (TAP) (2017, 2019), for instance, explicitly reconceives communication in affective terms, conceptualising EEs as pragmatic and probabilistic actions through which individuals accomplish social and motivational functions. To express is to do something with affect. A scowl, smile, or tone of voice does not merely report an inner state but performs a social act such as warning, inviting, or appeasing. Meaning arises from the role an expression plays in interaction rather than from a fixed correspondence with an inner emotion. Under TAP, EEs follow expressive regularities in a probabilistic way. What matters is not perfect accuracy but the functional reliability of the expressive signal within the practice. Hence EEs can still convey useful information even when strategically modulated, ambiguous, or partially misread. Observers can therefore infer probable affective and motivational states even without a deterministic mapping and/or explicit communicative intent. TAP is thus a great example of a model that can articulate how EEs can guide social understanding and coordination even when they are unplanned, misleading, or only partially recognised by perceivers. In this sense, Scarantino's Theory of Affective Pragmatics explains how EEs can be both performative and informational.

In what follows, I draw on the previous discussion to develop an account of EEs that is responsive to the novel challenges technosocial contexts involving ERTs pose for our understanding of affective life and its role in social interaction.

4.4. A revised design-based account of emotional expressions

Chapter 2 introduced ERTs and the prospect of fully functional systems, outlining their aims and the most promising ways in which they are expected to achieve them. Chapter 3 examined emotions as intrapersonal phenomena and their relations with motivations, moral salience, and action tendencies. The preceding sections of this chapter analysed what EEs are and how

competing accounts explain their capacity to convey meaningful information within affective communication channels.

In this section, I turn to technosocial contexts and argue that EEs perform a dual function in those contexts, respectively for humans and machines. For humans, EEs are best conceived as essential tools of social interaction, coordination, and negotiation. In technosocial contexts, however, they also serve functions for ERTs: for these technologies, emotional expressions operate as data for the diagnosis of inner affective states (see Chapter 2). This difference in purpose, communicative for humans and diagnostic for ERTs, reveals two distinct regimes of recognition. My aim is therefore to articulate a functional account of EEs that acknowledges this bifurcation of purpose. The account developed below combines a modified interpretation of the design view to explain the nature of EEs while keeping core insights from the pragmatic account of affective communication (Group B) to explain their role and reliability in social life. In doing so, it preserves a unified account of emotional expressions while recognising that their functionality differs in human and ERT-based modes of recognition.

4.4.1. Emotional expressions' epistemic functions for humans and technology

The design view is a promising starting point for explaining why EEs can serve as reliable indicators of affective life because it avoids most of the pitfalls of subgroup A1 and A2. Yet it must be revised to accommodate ERTs' diagnostic uptake. I first outline the standard view and then develop a modified version that captures both human and technological uptake.

On the design view, EEs are usually taken as behaviours or physiological changes that have stabilised over time – biologically, culturally, or conventionally – because they make emotional or motivational states perceptible to others in a reliable way (Griffiths 1997; Green 2007). A major advantage of this view is that, unlike other causal-type accounts (subgroup A1), it can bridge the biological and social perspectives by grounding the identity of EEs in their functions, accommodating both EEs that are the result of biological/evolutionary *design* and those that are the result of social or cultural *design*. For instance, the widening of the eyes in cases of genuine fear is often explained in evolutionary terms. Such a change may have stabilised over time because it reliably alerts others to potential (deathly) danger, facilitating rapid collective responses. By contrast, other EEs are best explained by appeal to social or cultural design. Consider the polite smile, which functions to signal politeness or a desire for affiliation in social interaction. Such smiles may be argued to have stabilised over time because they reliably serve

coordination and norm-regulation functions within particular social practices. Similarly, culturally specific gestures (e.g. gesturing two pinecones for Italians) can be explained as deriving their expressive status from shared conventions and learned norms.

The standard design view also provides a principled way to account for cross-context regularities while allowing expressive variation. In the case of polite smiling, for instance, the affiliative or appeasement function may recur across contexts, while the precise form, frequency, or appropriateness of the smile varies depending on local cultural pathways and norms. On this view, regularities arise from shared communicative pressures, whereas variations reflect different evolutionary or cultural solutions to fulfilling those functions.

Finally, another major advantage of the design view is that it avoids being grounded in a purely perceiver-dependent requirement, unlike accounts from subgroup A2. On pure perceiver-dependent accounts, a behaviour counts as expressing an emotion only in virtue of how it appears to a given perceiver. This suggests that an emoter may judge that their own behaviour conveys emotion E while a perceiver judges that the same behaviour conveys emotion Y instead. From this follows the possibility that the existence of emotion E is then denied, as only a perceiver's uptake truly grounds what a particular EE counts as. In technological contexts this is troubling because an emotional expression's meaning vis-à-vis a given emotion could be decided by ERTs rather than by the emoters themselves.⁹⁹

Even if it avoids some of the main challenges of subgroup A1 and A2, the design view does not escape all difficulties either. A first problem is this: if expressions are *designed* to reveal a particular emotion or social motive (i.e. they are behaviours or physiological changes that have stabilised over time, biologically, culturally, or conventionally), why are many if not all physiological changes and behavioural expressions not *exclusive* to a single emotional state or social motive? I may blush from embarrassment, heat, or anger, which seems to weaken the idea that blushing

⁹⁹ This has practical impacts. For instance, as ERTs are especially popular in the health care sector, this would also challenge the current dominant attitude in medicine and health care, which ultimately assumes that “the patient knows best”, i.e. if a patient's self-report of pain level does not match the estimate provided by a given test result, doctors should first listen to the patient's own experience of her pain and act accordingly. This could potentially fall under the type of harm that Glazer (2019) coined as “emotional misperception”, a form of epistemic violence, following Dotson's concept (2011). Epistemic violence, in the form of emotional misperception, arises, Glazer argues, when “A misreads B's nonlinguistic expression of emotion, owing to reliable ignorance, harming B.” (Glazer 2019: 60). Because only the emoter will ever have full access to the emotion, pure perceiver-dependent accounts, if endorsed, seem to be a threat to emoters' authority in asserting what they feel, and to deny the imbalance in information between the emoter (full access to information) and the perceiver (limited access to information).

is a change *designed* to signal one specific emotion or social motive (Rowson 2024).¹⁰⁰ Yet it would be counterintuitive to exclude blushing from the class of EEs simply because it is not exclusively designed to signal embarrassment, heat, or anger. A compelling account should respect this intuition and avoid ruling out such features on the grounds of ambiguity.

A second difficulty concerns perceptibility. Standard design-type accounts often imply the requirement that “to be designed” means that an expression must in principle be available for uptake by others (*viz. perceptible*). However, many bodily changes are not reliably perceptible in ordinary interaction, such as an accelerated heartbeat, pupil dilation, or sweating. It seems nonetheless also counterintuitive to exclude these changes from the class of EEs on that basis alone, especially given that affective computing and ERT research increasingly treats them as potentially informative about affective states. A robust account of EEs should therefore account for these physiological changes.

To address these difficulties, I propose two revisions to the design view. First, I suggest shifting the burden of the “design” definition from singular behaviours or physiological changes (e.g. blushing) to *patterns* of such changes. This way, it is the configuration of changes – not individual changes – that may be considered as providing the “complete” pathway between the cause of an EE and its physiological or behavioural correlates.

Second, whereas standard design-type accounts often assume that designed changes must be perceptible, I argue for a weaker requirement: recognisability rather than perceptibility. By recognisability, I do not mean that the relevant changes must be directly perceivable through ordinary visual or auditory perception. Rather, I mean that, to qualify as recognisable, a pattern of changes needs only to exist, whether or not it has (yet) been detected or interpreted by human observers. In this sense, the proposed view departs from both Group A and Group B accounts, which typically impose a condition of (average) human perceptibility. On the revised design-based account proposed here, patterns of change need not be visually or auditorily accessible in order to be considered ontologically salient features of EEs’ causes.

Shifting from singular changes to patterns of such changes implies an explanation of what would count as such patterns. I argue that to qualify as patterns, these configurations must exhibit relative stability. By relative stability, I take inspiration from Scarantino’s probabilistic

¹⁰⁰ See also Barrett (2017) on this particular point.

explanation to argue that these patterns need not be universal or strictly regular (meaning that they may vary across individuals and contexts), but they must be able, on their own, to transmit meaningful information about one's affective states or social motives. In short, a “complete” EE is a recognisable configuration that reliably indicates an affective or motivational state, even though this recognisability is not strictly dependent on ordinary human perceptibility.

With these two core revisions of the standard design view in place, I now turn to the information perceivers may expect to get from EEs. On the revised design-type account, human perceivers' uptake of an EE remains dependent on other complementary epistemic inputs such as context, background knowledge, and social norms. What we treat as “typical” EEs, such as blushing or culturally specific gestures (e.g. the pinecone gesture in Italy), may therefore have both biological and social origins, but on this view, they do not necessarily constitute expressions on their own. Rather, they function as parts of broader physiological and behavioural sets.

Importantly, when combined with other relevant epistemic inputs, a single behavioural or physiological change may nonetheless suffice for successful human uptake. For instance, if a pinecone gesture expresses angry disappointment in Italy, then it may be sufficient in that context when perceived by Italian interlocutors.¹⁰¹ Similarly, seeing my friend miss a step in a public stairway and subsequently blush may be sufficient for me to infer that she is embarrassed. A noteworthy assumption of this proposed account is thus that humans need not have complete access to the *full* pattern of physiological or behavioural changes of a given EE in order to successfully infer the cause of that EE.

For humans, a specific EE becomes more or less probable only against a background of relevant epistemic inputs. These inputs also bear on which causes (inner affective states and/or social motives) are most plausible. In most cases, humans can rely on *partial* access to the behavioural or physiological patterns of a given EE because they supplement their perceptual uptake with relevant epistemic inputs (context, social norms, and background knowledge). This explains why ordinary perception is generally a reliable mode of recognition in everyday interaction, but also why it becomes fragile when those supplementary inputs are unavailable or misleading.

¹⁰¹ This would also work, for instance, if people know that the people making the gestures are Italian and they know what this specific gesture means for Italians.

While spontaneous EEs and their associated meanings are useful for humans, intentional EEs often play a particularly important role in interactional contexts. For this reason, humans have good reasons to attend to the full spectrum of EEs. ERTs, by contrast, typically focus on inferring inner affective states, largely disregarding intentional expressions and their functions. What ERTs ‘recognise’, then, are *only* the patterns of physiological/behavioural changes that would suggest an EE whose cause is an inner affective state.

But despite this narrow take on EEs, recent advances in AI systems and multimodal emotion recognition (Pan et al. 2023) allow us to hypothesise that ERTs may incorporate *user-specific* data, mimicking a form of background knowledge that may nuance their EE identification process and model it to a single individual, perhaps grasping idiosyncrasies and other individually rooted variations as well (García-Hernández et al. 2024; Khare et al. 2024; Guo et al. 2024). This is the important distinction between *human* emotion recognition and *technology-based* emotion recognition. Accordingly, what looks to human perceivers like ‘only’ blushing as the expression of embarrassment reflects the limits of human access. ERTs need not share those limits. It is plausible that embarrassment co-occurs with a broader, probabilistically reliable, configuration of physiological and behavioural changes that humans cannot fully perceive. This working hypothesis tracks recent empirical work in both neuroscience and affective computing, which keeps improving on the discovery of these hypothesised patterns (Shu et al. 2018; Piretti et al. 2023; Yin et al. 2017).

These two revisions allow the account to accommodate the respective capacities of both human and machine “perceivers” without rendering expressive status observer-dependent, since only epistemic access varies. As a result, EEs remain empirically grounded and socially situated, while allowing for cultural variability without collapsing into relativism (as strictly perceiver-dependent accounts risk doing).

This proposed framework therefore articulates how EEs may play distinct functions in human and ERT recognition. Viewed this way, ERTs and humans may rely on the same or different physiological and behavioural patterns when interpreting a given situation. A “complete” EE is thus treated as a set of physiological and behavioural changes that has epistemic value for human

perceivers and ERTs insofar as it raises the probability of a plausible interpretation above a threshold.¹⁰²

An interesting potential advantage of this view is that it may also account for the idea that if individuals have multiple, superimposed affective and motivational states, it could in principle be the case that human perceivers and ERTs both pick up on one while both being right about its existence. As argued in Chapter 3, emotions operate as omnipresent inner states, and social motives can align with them, collide with them, and thus, often co-occur with them. Hence, let's imagine a situation where someone is angry for reason *x*, and thereby is assumed to have physiological/behavioural changes in the form of a pattern that may be correlated to her anger. Let's say she tries to suppress or at least conceal it, rendering it not visible to human perceivers for social reasons. Let's say she does so because she is seeking affiliation. To express this social motive, she produces a conventional smile. Given that, through her EEs, she is trying to conceal her inner affective state and to share her social motive, it is reasonable to assume that human perceivers are likely to notice only the smile and infer from it an affiliative motive, whereas an ERT is likely to detect the pattern correlated with the inner affective state and classify it as anger. It is thus implied that, because humans often intentionally deploy motive-grounded EEs to regulate relations, diagnostic exposure of concurrent inner affective states plausibly disrupts those social functions. The proposed design-based account helps shed light on these types of scenarios by splitting the epistemic routes through which ERTs and humans interpret EEs. In the next part of the thesis, I will explore further the consequences of such scenarios and investigate the idea that when an ERT isolates and exposes an inner affective state, it potentially disrupts the social functions of an intentional EE, impacting important foundations on which social life currently relies.

4.4.2. Potential challenges for the proposed view

The proposed account sketched above does not come without challenges of its own. First, the assumption that there *are* relevant patterns linking sets of physiological and behavioural changes to the causes of EEs is not yet supported in a compelling way by empirical work, even though some promising avenues are being explored. Second, even if such patterns can be identified in principle, EEs remain vulnerable to interference, both voluntary, such as active suppression, and involuntary, such as emotional overload, and it is not clear yet whether those interferences

¹⁰² Additionally, for humans, it is typically within a situation already shaped by context, background knowledge, and social norms.

would affect the possibility of pattern recognition, more particularly on the ERTs' side. Emotional life is complex, dynamic, and often plural. Tensions can arise between biological and social causes, or among multiple affective forces, producing atypical or idiosyncratic expressions that disrupt the stability of the hypothesised patterns.

However, I take such cases to be the exception. Their existence is troubling, but not sufficient to reject the account altogether. The literature already suggests that most EEs can be said to be humanly recognisable patterns, which is why they reliably support social interaction (Van Kleef 2009; Keltner and Lerner 2010; Hareli and Hess 2010; Scarantino 2017; Rinne 2024).

Coming back to the first concern, research is progressing, especially through multimodal ERTs. No decisive pattern has been established yet, but advances in machine learning, deep learning, and multimodal modelling make the hypothesis increasingly plausible, which is enough to treat it as a provisional assumption.¹⁰³ Moreover, the hypothesis of relevant patterns is a useful way of defining EEs given how contemporary AI works, but it is not the only one. Importantly, even though this is the way I define EEs, this definition ultimately bears low weight on the argument of the bifunctionality. For the bifunctionality to hold, it is sufficient that EEs can be, through technological means, reliably aligned with their causes, regardless of whether that alignment is humanly perceptible.

Thus, in spite of these issues, the proposed account nonetheless provides a way to understand how it may be possible to provide a functional definition of EEs that is compatible and coherent with the idea that EEs may serve a role in two different epistemic routes to recognition, as seems to be the case in technosocial contexts involving fully functional ERTs.

4.5. Conclusion

This chapter has examined EEs as the pivotal interface between the affective inner life of individuals and the social structures within which this inner life acquires social meaning. By transitioning from the nature of emotional expressions to analyses of their (social) functions, the chapter has shown that EEs are neither mere physiological symptoms nor passive reflections of emotions. They are constitutive elements of affective communication systems, embodied,

¹⁰³ See Chapter 2.

patterned, and context-sensitive modes through which emotional and motivational states become publicly intelligible.

By (1) examining the nature of emotional expressions, mapping the principal ontological debates surrounding them, and (2) situating EEs within affective communication channels, focusing on probabilistic-pragmatic accounts to understand their meaningfulness and reliability in social interaction, the chapter advanced two core claims. First, nonverbal affective communication constitutes a central channel of everyday social interaction. Second, emotional expressions play this role because they function as defeasible yet generally reliable sources of information on which emoters and perceivers rely to coordinate action, manage expectations, and regulate social relations. Emotional expressions are thus integral components of the informational infrastructure of social life.

Building on these ontological and communicative analyses, but also on the technological insights of Chapter 2, the chapter turns to the question of the functions of EEs in technosocial contexts and argues that emotional expressions are best understood as exhibiting a bifunctionality in those technosocial contexts. For humans, EEs function as communicative tools that enable social coordination, negotiation, and norm regulation. For ERTs, by contrast, the very same expressions are treated as diagnostic data points used to infer, classify, and predict inner affective states. To account for this bifunctionality, the chapter sketches the outlines of a revised design-based account of EEs. On this proposed account, EEs are relatively stable patterns of behavioural and physiological changes designed (biologically, culturally, or conventionally) to make affective and motivational states recognisable rather than perceptible. This framework offers a definition of emotional expressions that accommodates both human and machine perceivers while maintaining conceptual coherence.

The resulting picture is one where EEs serve both as instruments of social coordination and as data sources for affective inference. These two functions can operate side-by-side when emotional expressions are interpreted in line with shared social expectations. However, they may come into tension when diagnostic recognition discloses aspects of emotional life that are strategically managed or deliberately concealed. In such cases, what is communicatively appropriate for humans (e.g. an appeasing smile, a polite nod, or a controlled tone) may be placed into tension vis-à-vis the inner states identified by fully functional ERTs from the individual's physiological and behavioural patterns.

This tension between communicative and diagnostic uptake marks a critical juncture for understanding ERTs and what is at stake in the ethics of emotional AI more broadly. By showing how EEs may serve different epistemic and functional purposes, this chapter provides the conceptual bridge to the ethical inquiry that follows. The next part of the thesis takes up the question of how diagnostic detection may interfere with the interpersonal mechanisms that sustain emotional privacy (mechanisms such as affective opacity, selective disclosure, and expressive control), without which the moral and social functions of emotions and social motives cannot be properly realised.

Part Three:
The (Ethical) Analysis

5. Privacy and emotional life: Re-engineering the concept of emotional privacy

5.1. Introduction

The previous chapter argued that emotional expressions are probabilistically informative patterns of behaviour and physiological changes that provide perceivers with reliable, though defeasible, affective information. Emotional expressions thus sustain some of the core background conditions that make mutual understanding, interpretation, and warranted reliance possible in social life. They enable social orientation, decision-making, coordination, and trust, yet they remain situated, contested, and subject to deliberate modulation within context-sensitive practices of interpretation. It is this calibrated interplay between legibility and opacity that the prospect of fully functional ERTs threatens to reconfigure.

This chapter responds to that technomoral disruption by re-engineering the concept of emotional privacy. I argue that emotional privacy concerns the expressive ability of individuals to shape the legibility of affective information in interaction. This positions emotional privacy within the social perspective of (informational) privacy, according to which privacy is not primarily a retreat from others, but a constitutive condition of social life. I present emotional privacy as a subconcept of the informational privacy cluster, where the object of emotional privacy is (a type of) information. Its aim is thus not to introduce a new way of conceiving the broad concept of informational privacy, in the way that other accounts do (such as the Control-based approach, Access-based approach, and Contextual Integrity), but rather to offer a new conceptual tool within the informational privacy cluster in face of the new tech-induced conceptual gap that fully functional ERTs risk creating. After identifying the conditions under which emotional expressions can perform their social functions, I propose emotional privacy as a new conceptual tool to solve this problem.

Section 5.2 traces the conceptual evolution of privacy and shows how technological shifts repeatedly generate the need for conceptual adaptation. **Section 5.3** draws on the case of mental privacy to illustrate how subconcepts of privacy can be pro-actively added in anticipation of emerging technologies. Building on this precedent, I define two types of emotional privacy: actual emotional privacy and standing emotional privacy. While in everyday life we often lose actual emotional privacy, standing emotional privacy is typically retained. **Section 5.4** examines how emotional privacy operates within affective communication channels. It introduces

expressive conventions, affective leakages, and two types of privacy mechanisms – selective disclosure and concealment – as the dynamic tools through which individuals negotiate affective legibility. **Section 5.5** turns to the value of emotional privacy and shows that these privacy mechanisms are functionally significant at three levels of social interaction: they preserve self-presentation at the individual level; sustain role plurality and norm conformity at the interpersonal level; and contribute to social regulation and cohesion at the societal level.

5.2. Privacy

The previous chapter argued that emotional expressions (EEs) are probabilistically informative yet defeasible signals that sustain core conditions of social life. Building on that account, this section examines the relationship between privacy and emotional life, particularly in technosocial contexts. It traces the historical and conceptual evolution of privacy, showing how shifting social norms and successive technological innovations have repeatedly challenged and reshaped our understanding of both emotional life and privacy. By showing that privacy is not a static concept but one that evolves in response to technological and cultural changes, the section highlights why (inner) emotional life, once thought inaccessible, is now at risk of being exploited by ERTs.

In what follows, I first examine two relations that will help me foreground the concept of emotional privacy in light of new technological progress, namely the relation between privacy and emotional life and between privacy and technology. I then focus on the recent developments in (informational) privacy scholarship.

5.2.1. Privacy and emotional life

Although privacy was only recognised as a human right in 1948, when the United Nations enshrined it in the *Universal Declaration of Human Rights*, its concrete expression has taken various forms over the years. Particularly its related subconcepts, such as spatial privacy, proprietary privacy, bodily privacy, associational privacy, decisional privacy, and more recently, informational privacy, have been understood, recognised, and used differently over time but also across different cultures (Koops et al. 2016). For instance, whereas the bedroom is today generally regarded as a private space in French culture (spatial privacy or proprietary privacy), this was not the case for the monarchs, particularly Louis XIV, who turned waking up and going to bed into public rituals for his courtesans and subjects. Another notable example is the evolution of bodily privacy, which was shaped by the global eugenics movement and later

deconstructed and reshaped by feminist movements in the twentieth century. From the marital-rape exemption to the forced and coerced sterilisation of vulnerable groups in many Western countries, such as Denmark (Bryant 2025), the USA (Peal 2004), and Canada (Stote 2015), the concept of bodily privacy has changed profoundly in the past century, with most Western countries currently recognising bodily privacy as the capacity to control access to and interference with one's body, protecting against unwanted intrusion in medical, technological, and social contexts.¹⁰⁴

Against this backdrop of profound shifts in the conceptualisation and collective understanding of privacy, emotional life stands out as a persistent constant. Even in highly remote cultures – such as the Javanese, the Ngadju Dayaks of Borneo, the Chocó Indians of Panama, or the !Kung Bushmen of Southwest Africa – where physical, spatial, and even bodily privacy have no equivalent to the conceptions endorsed by modern Western societies, at least a part of emotional life is nonetheless voluntarily and carefully concealed under finely tuned decorum (Altman 1975: 15-16, 42).¹⁰⁵

Despite this, scholarship on the privacy of emotions remains scarce. Perhaps partly because the concept of privacy did not need to be as prominent in academic discussions as it is today, and partly because emotions were somewhat *de facto* recognised as a deeply personal and inaccessible individual phenomenon to such an extent that the violation of this sphere was deemed impossible.¹⁰⁶

The advent of new intrusive forms of technologies is what brings privacy and emotion scholarships together. In Warren and Brandeis seminal paper on privacy (1890), we can observe that their concern about public image, rising from the advent of new technologies, led them to go as far as to argue for a protection of individuals' "thoughts, feelings and sensations" (Warren & Brandeis 1890: 195). In 1988, Anita Allen expressed concerns over the individuals' ability to protect their inner life against new forms of (techno-)intrusions, specifically noting the potential access to emotions. She argued that to say that a person has privacy is to say that "in some respect and to some extent, the person (or the person's mental state, or information about the

¹⁰⁴ This excludes some countries, such as the United States, which has recently seen major setbacks on this subject, particularly with regard to reproductive rights and the recent overturning of abortion laws (see the overturning of *Roe v. Wade* by *Dobbs v. Jackson Women's Health Organization* in 2022).

¹⁰⁵ See also Westin (1967: 13-18).

¹⁰⁶ They are sometimes treated as a form of qualia (see Chalmers 1996; Nagel 1974).

person) is beyond the range of others' five senses and any devices that can enhance, reveal, trace, or record human conduct, thought, belief, or emotion." (Allen 1988: 15).¹⁰⁷

5.2.2. Privacy and technology

The rapid evolution and development of privacy scholarship in the last two centuries is mainly due to the development of new technologies that forced a reconceptualization of our conception of privacy. New technologies have consistently pushed against the boundaries of what counts as private. Each wave of innovation has unsettled prevailing social norms and revealed conceptual gaps in the understanding of privacy, forcing philosophers and jurists to re-articulate its meaning.¹⁰⁸

To reuse Warren and Brandeis's case, instant photography and the phenomenon of mass press that emerged from it were what provoked anxiety about exposing intimate life without consent, pushing for the adoption of a right 'to be let alone'.¹⁰⁹ By the mid-twentieth century, the spread of computers, databanks, and record-keeping systems shifted attention to the circulation and control of personal information. The most pressing threat was no longer photographers at the window but the invisible, large-scale capture of personal data. Concerns about government databanks and commercial profiling crystallised around the concept of informational privacy. These early concepts of informational privacy came again under strain at the turn of the twenty-first century with the rise of digital networks, online platforms, and ubiquitous surveillance. These technologies blurred the delineation of public and private spaces and cast doubt on the feasibility of both informational "control" and limited access so understood.

Every new technological wave brought with itself a need for conceptual adaptation with regard to privacy. Instant photography, x-ray devices, phone trackers, computer databanks, and today, digital networks and ubiquitous technology-based surveillance are reshaping our very understanding of what it even means to 'have privacy'. As for the mental and emotional life, the

¹⁰⁷ Interestingly, she also notes how one's emotions and other mental states may be readily available to others through external states such as emotional expressions. She observes that: "[f]acial expressions, gestures, and body movements can give us away. Happiness and anger, approval and disapproval, sincerity and deception are often discernible by external inspection" (1988: 14). Hence, "the external self can reveal the internal self, [and thus] inaccessibility to persons' bodies and products of their bodies may be needed to keep the private self private." (1988: 14).

¹⁰⁸ See Hopster & Löhr (2023) for a detailed argument that technologies create conceptual disruptions.

¹⁰⁹ In the nineteenth century, privacy was mainly understood as a protective boundary around the individual that secured autonomy, dignity, and freedom from intrusion (Mill [1859]1985; Westin 1967; Thomson 1975; Moore 2010). In this liberal tradition, the private realm functioned as the counterpart to the public, a space in which individuals could retreat, develop, and exercise self-determination (Mokrosinska 2018).

apprehensions expressed by Warren and Brandeis now find concrete echoes in contemporary forms of affective (techno-)weaponization, such as nudges (Wilkinson 2013; Thaler & Sunstein 2008, 2021), neurotechnology (Ienca 2021; Magee, Ienca, and Farahany 2024), and fully functional ERTs. Emotional life, once considered inaccessible, now risks being exploited and commodified through technologies that promise to render it transparent.

Now that I have established the links between both privacy and the emotional life, and privacy and technology, I now turn more particularly to this very concept of (informational) privacy, and establish what is the current state of affairs in terms of protection against new and invasive technologies.

5.2.3. (Informational) privacy scholarship

The deployment of technology that could process and keep detailed records of the masses in the mid-twentieth century shifted scholarly attention from the previously dominant notions of privacy that were proprietary and spatial privacy toward the emerging concept of informational privacy.¹¹⁰ To grasp what ‘informational privacy’ should entail, two general, but also competing notions were seen as plausible, namely, the control- and access-based views.

A prominent defender of the Control-based approach was Alan Westin, who defined privacy in *Privacy and Freedom* (1967)¹¹¹ as the individual’s ability to control the *disclosure* and *circulation* of personal data, in other words, the ability to determine “when, how, and to what extent information about them is communicated to others.” (1967: 7). On this view, privacy is an exercise of choice and agency, modelled on a liberal notion of individual autonomy. Informational privacy thus emerged as a response to a new socio-technical environment, where harms arose less from direct intrusion into one’s private (physical) sphere and more from the accumulation, storage, and re-use of informational traces. This now Classic Control account of privacy can be formalised as

¹¹⁰ See Roessler (2005) and Moore (2010) for modern counterarguments against this focus on informational privacy. Roessler and Moore defends the relevance of other types of privacy, particularly local and decisional privacy for Roessler and spatial and bodily privacy for Moore. Roessler, for instance, argues that “Something counts as private if one can oneself control the access to this ‘something’.” (Roessler 2005: 8). However, as I am focusing on affective communication channels and EEs, I will only gesture at these competing views and restrict my analysis to informational privacy.

¹¹¹ See also Moore (2003, 2010) and Fried (1968).

Classic Control: Privacy obtains when person A has the ability to give access to information *p*, which she possesses, to (any) person B.

The access view, by contrast, argued that (informational) privacy is not adequately protected by the notion of control and instead should be best viewed as being about the notion of ‘access’, understood in broad terms as the extent to which others can reach the individual. What I take as the now ‘Classic Access-based accounts’ thus emphasise the *passive condition of being shielded from (unwanted) access*. Ruth Gavison (1980), for instance, defined privacy as “our concern over our accessibility to others” (1980: 423). Jeffrey Reiman defined privacy as “the condition under which other people are deprived of access to either some information about you or some experience of you” (1976: 30). Consequently, the Classic Access account of privacy can be formalised as

Classic Access: Privacy obtains when information *p*, possessed by person A, is inaccessible to (any) person B.

Mainly due to their dialogue with each other where defenders from both sides restlessly pointed at the weaknesses of the other’s accounts, the Control and Access views have evolved from their respective classic forms. Hence, building on the classic form of the Control-based account, contemporary defenders, such as Mainz and Uhrenfeldt (2021) and Menges (2021, 2022) recognised that the Classic Control formulation was too imprecise on the *meaning* of control, leaving the general account weak and difficult to defend, hence pressing for a revision of the classic control formulation.¹¹² Taking up this challenge, Mainz and Uhrenfeldt (2021) propose to revise the control account along the idea of ‘Negative Control’.¹¹³ Negative Control is understood, as its name indicates, as the ability to prevent access. Contrary to the classic account of control, which present an active notion of control in which people may give or prevent access to information (presumptively at any time and under any conditions), Mainz and Uhrenfeldt

¹¹² The formulation of the Classic Control indeed leaves the door open to many cases where *a loss of privacy does not entail a loss of the ability to control the disclosure of information p*, and vice versa. For instance, as it is formulated, person A could “keep the ability to control” who will know about *p*, while person B could still correctly infer *p* from other epistemic sources. Conversely, it has also been argued that A “loses control” but not her privacy if, for instance, she attempts to give B access to her information *p*, and B *prevents* it. Following Mainz and Uhrenfeldt’s assessment for example, this does not intuitively lead to a *plausible* loss of A’s privacy.

¹¹³ They also advance the possibility of a “Republican Control” notion of privacy, in which person A’s privacy is lost simply in virtue of B’s ability to access her information. But they flag it as also problematic as this ultimately means that the mere invention of new surveillance technologies, without their use, diminishes A’s privacy, which also does not intuitively lead to a plausible violation of A’s (moral) right to privacy (Mainz & Uhrenfeldt 2021: 300). Hence, I will not consider this formulation.

notion of negative control presents control as a reactive notion which focuses on the ability to prevent access *once* someone does attempt to access. Following this notion,

Negative Control: Privacy obtains when, given an access attempt by (any) person B, person A possesses the capacity to prevent B from accessing personal information p about A.¹¹⁴

Menges (2021, 2022) recently defended another alternative on how to best conceive the notion of control, which he mainly grounds in the moral responsibility literature (2021: 395). He argues that control should be defined as

Source Control: Privacy obtains when person A is the *right kind* of source for information p's flow (if information p flows at all).

Following the Source Control account, the focus is not on the way person A would lose control over B's access to information p – what Menges relates to (effective) voluntary choice – but rather on the legitimacy of A as the *right kind* of source for information p's flow (viz., sourcehood). Hence, Menges considers other accounts to be concerned with different kinds of (effective) voluntary choice (viz., control) over the access of (any) person B to information p about person A, while the Source Control account is concerned with the legitimacy of the source (viz. the right kind) making the information flow. This view has several advantages over the Classic Control view, as it tracks back privacy loss to instances where someone else makes personal information about the agent flow without the agent's input as the source. Take the case where person B steals the diary of person A, but does not read it. In such cases, the Classic Control account struggles to explain why there seems to be no violation of privacy because even though there is clearly a loss of control on the part of person A, this loss did not result in an access to the information. In contrast, the Source Control account does not (yet) identify a privacy violation the moment person B steals the diary. This is because the focus is not on whether B has lost (effective) control over the information, but on B being the wrong kind of source for the information flow, if the information flows at all. As a result, person A,

¹¹⁴ Note that for reasons of clarity, I only refer to the (last) formulation of Mainz and Uhrenfeldt's control account, in which they say that person A's right to privacy is violated iff "Agent A has involuntarily lost Negative Control over the access to personal information P about agent A, *due to action(s) of agent B, of which B is responsible.*" (2021: 298, authors' original italics). I will return to this last formulation in the next chapter as it is intended for a normative account. Here I only want to highlight the diverse views and evolution of (informational) privacy.

presumptively the only right kind of source for this diary, has lost privacy over the diary's information *only* in the case that B does access it, since B is the wrong source of the flow.

Similarly, on the Access view, compelling alternatives to the classic formulation are to be found in Anita Allen's (1988),¹¹⁵ Tavani and Moor (2001), and Macnish's (2018) work. Allen and Tavani and Moor both propose a Restricted Access account to privacy. The restriction is mostly determined in relation to the context, the situation, and the actors present. On Allen's account, restricted access means that privacy is a condition in which access to persons and personal information is appropriately limited. On Allen's view, to say that privacy has been violated is to say that someone has been made accessible in contexts where *restricted access* is the very point of the practice.

In a similar vein, Tavani and Moor write that there are many established "zones" in which the protection of privacy is already reasonably expected and normatively protected.

The normative aspect of these private situations restricts access by individuals, groups, or governments. This restricted access expresses a right of protection. It prohibits intrusion and information processing by someone or something. [...] Although there is some conventionality in how we carve up zones of privacy in social situations including Internet use, overall the carving should produce a set of zones that offer sufficient protection of personal information. [...] In defining a private situation it is necessary to define who has access to what under which circumstances. (Tavani & Moor 2001: 7).

Tavani and Moor also raise the issue of the repurposing of the data gathered on individuals' personal information. For instance, does the fact that I have willingly granted access to my personal information *p* to person B, in circumstance *x*, also mean that now they can sell it and have it accessed by person C, in circumstance *y* (Tavani & Moor 2001: 6)? Below I have summarised the idea of Restricted Access accounts as

Restricted Access: Privacy obtains when only members of a specified set *R* have access to information *p* in the given situation, where *R* is fixed by the situation's role- and purpose-structure.

Another alternative to the Access view has recently been proposed by Kevin Macnish (2018) who insists that privacy is lost only when there is *actual* (epistemic) access. On his account, automated capture or background processing that no individual meaningfully understands does

¹¹⁵ Later, Anita Allen defines it as "limiting access to personal information and electronic communications" (2011: 196), thus slightly modifying the view she had previously defended.

not by itself reduce individuals' privacy. The loss occurs when a person who can understand, in fact accesses information *p* in a way that can be brought to bear in practice (Macnish 2018, 2020). Björn Lundgren (2020, 2021) and Peter Königs both endorse Macnish's requirement of actual (epistemic) access.¹¹⁶ Addressing more specifically the problem of government surveillance, Peter Königs (2022) builds on the Actual Access account of privacy within institutions, tying privacy loss to the point of human access, and in a second step, legitimating the loss within role-justified purposes.¹¹⁷ The reduction of privacy is bound to the actual (epistemic) access while the right to privacy is then assessed on whether the access was carried out by an authorised individual or not (Königs 2022). Hence the Actual Access account of privacy reads as

Actual Access: Privacy obtains when no person *B* with understanding has actually accessed *A*'s information *p*.

In parallel to the co-development of the Control- and Access-based approaches is notable work in what is best conceived as the social perspective of (informational) privacy. In response to new socio-technological threats, scholars such as Helen Nissenbaum (2004, 2010), Julie Cohen (2013), and Beate Roessler and Dorota Mokrosinska (2013, 2015) reconceived the value of informational privacy as a constitutive feature of social life, reshaping privacy's relevance in the face of new technologies. Leaving both the notion of control and access aside, Nissenbaum's Contextual Integrity (CI) framework, for instance, focuses instead on the appropriateness of the information flow. Among its strongest advantages is its fine-grained property, enabling it to explain, for instance, why "ten different people observing the contents of your shopping cart on ten different occasions is not the same, from a privacy perspective, as one party recording and storing it" (Nissenbaum 2019: 237, italics in the original).

Nissenbaum leaves the realm of privacy being seen as controlled or accessed by an individual, and redirects the attention towards the actual movement of information. CI starts by considering privacy as *appropriate* flow of information correlating violations of privacy as *inappropriate* flow of information (Nissenbaum 2010, 2018). This appropriateness is defined by a respect for the context's own integrity. This integrity is, in turn, itself determine through an analysis of the information flow in a given context. To be able to trace back the information

¹¹⁶ Lundgren argues for a pro tanto right (to privacy) against risks, meaning that this right can be infringed before actual access occurs in the case of serious attempts or engineered risk-impositions (Lundgren 2021).

¹¹⁷ In that sense, his view can be seen as a hybrid account between Restricted Access and Actual Access.

flow, CI focuses on the informational norms at play in a context. These informational norms are characterised by key parameters that are present in every context: the actors (senders, data subject, recipients), the transmission principles, and the attributes. These parameters respectively define the parties who are the subjects of the information as well as those who are sending and receiving it, the principles under which this information is transmitted, and the types of information at stake (Nissenbaum 2010: 142–143). Hence, flows of information are shaped by key parameters, and embedded in specific contexts. Libraries, health care system, voting stations, schools, airport security, etc., are all contexts.

Contexts are considered social domains with their own teleological values and purposes (Nissenbaum: 2010, 134). Hence, social contexts have entrenched social norms embedded in them (Nissenbaum 2010: 132-134, 146). These social norms help achieve the context's values, which take the form of goals, purposes, or ends. Prevailing values in the U.S. health care system, for instance, include alleviating physical suffering, curing illness, and promoting the health of individuals as well as collectives (Nissenbaum 2010: 134). Hence, informational norms and values are both in a co-constitutive relationship with their context, shaping it and being shaped by it (Nissenbaum 2010: 141, 180).¹¹⁸

On her account, thus, CI can be summarised as

Contextual Integrity: Privacy obtains when information p flows in a way that respects the prevailing entrenched informational norm(s) at play in a given context.¹¹⁹

If a new practice or new technology provokes a change in the flow of information, the practice is “flagged as violating entrenched informational norms, which constitute a *prima facie* violation of contextual integrity” (Nissenbaum 2010: 150). This departure from the entrenched context-relative informational norms at play in the context is assessed by drawing a “comparison [...] between [the] entrenched practice and the novel practice” (Nissenbaum 2010: 148-149).

Once the practice is flagged, it is not rejected *ipso facto* but needs to be assessed normatively. The normative assessment can be seen as the second part of the evaluation and mainly relies on

¹¹⁸ Part of this discussion is from Prégent (2025b).

¹¹⁹ I will take the moral augmentation part of the framework into account in the next chapter.

the values of the context. CI uses the values as evaluative tools to analyse the new technology. The values of a context are the result of a collective judgment that is, similar to the context-relative informational norms, also historically, culturally, and geographically shaped. Hence, a new technology that induces a departure from these entrenched norms is thus seen as an *a priori* threat, but can nonetheless be upheld against the entrenched practice if it can show that it respects and promotes the values of the context; thus maintaining the integrity of the context (Nissenbaum 2010: 186).

Let me briefly compare CI to the Source Control and Restricted Access accounts.¹²⁰ In the case of Menges' Source Control account of privacy, the account remains a control-account of privacy, only the notion of control is interpreted differently. This means that, on the Source Control account, to have privacy is to be the right kind of source of a data flow while “no receiver and no conscious agency is needed” (Menges 2022: 696).¹²¹ Hence the focus remains on the individual and her information p, contrary to CI, where the focus is on the informational flow itself and where receivers are of particular importance as they often end up being the reason of a ‘prima facie’ violation (e.g. in cases where a receiver changes, which means that a new direction is imposed on the original flow).¹²²

The Restricted Access account is also intrinsically different from CI. On the Restricted Access account, privacy is the condition of “restricted access”, which means that privacy is obtained when there is either a non-access or an authorised access. The focus is on who may look, hear, touch, smell, etc., which is dependent on role-justified permissions. On the Restricted Access account, privacy loss occurs when *unauthorised* parties gain access to information p, regardless of whether information p flows in a way that respects the context's integrity. This means that, while in cases where the flow is disrupted by a change in actor, CI and the Restricted Access account are likely to be similar in their output, in other cases, such as for instance, repurposing information p from health data to marketing data, they will likely diverge. The reason is that change of context is not in itself considered a valid ground for a loss of privacy following the

¹²⁰ For the Source Control account, see Menges (2022), footnote 2, for his full distinction between Source Control and CI.

¹²¹ Although Menges (2021, 2022) remains silent about what is entailed by the “right kind of source”, I assume that this is what he has in mind. This would mean that in the case of Source Control, a violation occurs when the authority of the source is impeded, bypassed, or defeated.

¹²² Following CI, privacy is lost (on a prima facie basis) when the information flow is altered, which may be identified through a change in actors, transmission principles, or attributes (Nissenbaum 2010: 142-143). The violation counts as a *proper* privacy violation once the moral analysis, evaluating the new information flow, fails to demonstrate that it upholds the (prevailing) values, ends, purposes, and goals of the context, in the same way – or in a better way – than the entrenched information flow did.

Restricted Access account. A loss of privacy would happen only when a person (or party), who has no authorised access, nonetheless gains access to information p . In cases where a parameter other than the ‘actors’ changes, only CI can flag the new flow.

Control-, Access-, and Contextual Integrity accounts are foundational touchstones for work on informational privacy. They shape how we think about and analyse flows of information, conditions of access, and appropriateness. Yet as mentioned by many scholars, the breadth of (technology-mediated) cases appears to resist any single, all-purpose definition (Roessler & DeCew 2023; Koops et al. 2016; Solove 2008) of informational privacy. This partly explains why, as noted in §5.2.2, conceptions of privacy are being repeatedly redefined at an ever-increasing pace. In the next section, I introduce the concept of emotional privacy as a complement to informational privacy in cases specifically tied to emotional life. I clarify the notion of tech-induced conceptual gaps and show how conceptual re-engineering can help address them. Building on DeCew’s cluster account of privacy (1997), which distinguishes bodily, spatial, and informational privacy, among others, I propose that emotional privacy is best understood as a subconcept of informational privacy. In this sense, I do not offer a competing theory of informational privacy, as Control-based, Access-based, and Contextual Integrity accounts do vis-à-vis one another. Rather, I articulate a new subconcept within informational privacy to better capture what is at stake, in privacy terms, when nonverbal emotional expressions are targeted by new and emergent technologies.

5.3. Emotional privacy

In this section, I propose to re-engineer the concept of emotional privacy by relying in particular on the case of mental privacy, which I take as an informative precedent to emotional privacy. Then, following the insights from the last chapters, I anchor emotional privacy in the social privacy perspective and define the concept.

I suggest *re-engineering* emotional privacy because in some ways “emotional privacy” is neither old nor new but at the same time it is also both an old and new concept. Philosophers have long recognised that emotional life is, at least in part, inaccessible to others, sometimes even treating emotions as qualia.¹²³ Yet recent scholars, including Menges and Weber-Guskar (2025), Akar (2024), Gremsl and Hödl (2022), and McStay (2018, 2020) argue that the exposure of the inner

¹²³ See Chalmers 1996; Nagel 1974.

affective life has become an urgent concern in the age of affective computing. The emotion recognition industry now being a fast-growing, multibillion-dollar market, with rising interest across public agencies, workplaces, education, and social media in deploying fully functional tools, only adds to this concern (Schwartz 2019; Miles 2024).¹²⁴ McStay (2020) was among the first to raise the alarm about the lack of protection surrounding (emotional) soft biometrics, identifying it as a significant regulatory gap. In response, Gremsl and Hödl (2022) and Menges and Weber-Guskar (2025) propose recognising information about emotions as ‘emotional data’ and integrating it into existing categories of personal or sensitive data, including the GDPR’s special categories. Akar (2024), while offering no explicit definition, highlights the need for a concept like “emotional privacy”. Upon closer examination, however, this emphasis on the *inner* affective life¹²⁵ overlooks a crucial dimension of emotional life and therefore fails to capture the constitutive social dimension of the concept, leaving it incomplete.

Focusing only on safeguarding inner affective experiences can allow us to focus on issues linked to the risks of certain inferences, which may be helpful in shedding light on important relations that emotions share with values, judgements, and beliefs, for instance, but it also makes us overlook the strong interplay between the inner and outer parts of emotional life (see Chapter 4). As a result, this focus blinds us to the important role emotional privacy plays in sustaining the core social functions of affective life. As much of the literature has so far focused on the former, I dedicate the remainder of this thesis to shedding light on the latter. To address this issue, I propose re-engineering the concept of emotional privacy in a way that accounts for its social dimension, as discussed in the preceding chapter.

In the remainder of this chapter, I thus go beyond current scholarship on the matter and (i) articulate what ‘emotional privacy’ is and (ii) specify precisely what it protects. This definition will help me to anchor the normative conditions of a (moral) right to emotional privacy in the next chapter. Chapter 6 will consequently formulate a proposition for a moral right to emotional privacy using the concept developed here and applying it to the case of ERTs to identify the normative conditions of its protection and distinguish wrongful violations of emotional privacy from ordinary mindreading.¹²⁶

¹²⁴ See Chapter 2.

¹²⁵ Although Akar’s analysis remains focused on the inner emotional experience, she also acknowledges the importance of social life in shaping that experience (Akar 2024).

¹²⁶ ‘Mindreading’ here is simply meant as the human ability to understand and attribute emotions and intentions to others from the interpretation of physiological and behavioural changes.

5.3.1. The case of mental privacy as a precedent to emotional privacy

We use concepts and conceptual systems to make sense of our complex reality. Concepts help us to put into neat categories different phenomena of life (Haslanger 2000; Cappelen 2018; Burgess & Plunkett 2013). Scholars working on conceptual engineering argue that, within our collective conceptual systems, *conceptual gaps* may be induced by new and emergent technologies (Hopster and Löhr 2023). “Tech-induced conceptual gaps emerge when we lack the conceptual resources to classify and conceptualise novel technologies, yielding conceptual uncertainty” (Hopster & Löhr 2023: 10). Such gaps may motivate new needs for *conceptual adaptation*, which is best understood as a response to a need for either (i) the introduction of a new concept, (ii) the elimination of an existing concept, or (iii) the preservation of a concept (Hopster & Löhr 2023: 8; see also Haslanger 2000; Cappelen 2018).

A canonical case of a tech-induced conceptual gap is the concept of “brain death”. “Brain-Death” is the result-concept of patients with severe brain injury whose brain activity is non-existent (or considered equivalent to it) while lung function is retained (indefinitely) through the invention of mechanical ventilation. The advent of mechanical ventilation created a conceptual gap in our life-or-death collective conceptual system where people were unsure about the classification of these patients: were they alive or were they dead? The ensuing uncertainty prompted the introduction and consolidation of the concept of “brain death” to respond to this new tech-induced need (Beecher 1968; Capron 1988).¹²⁷

Another illuminating example, which I take as a great precedent to emotional privacy, is the recent case of *mental privacy*. Control-based, Access-based, and, more recently, Contextual Integrity, have forced clear splits in the privacy scholars’ ideas on what it means to have informational privacy. But while the broader understanding of how we should best conceive it is still debated, the development of the subconcept of mental privacy shows that there is value in supplementing these broader conceptions with complementary subconcepts that are narrower and targeted to the exact problems that arise from particular tech-induced conceptual needs. Hence, I assume that mental privacy is, like emotional privacy, a subconcept of

¹²⁷ This situation notably provoked high moral uncertainty, and a rather obscure path towards potential moral obligations and other normative considerations that typically follow the creation of a new concept. In this case, this led to moral uncertainty regarding the moral obligations towards these patients (Bernat 2006; Truog & Miller 2014).

informational privacy. On this understanding, the object of mental privacy is a type of information pertaining to brain-related mental content.

The emergence of the concept of mental privacy has been triggered by advances in neurotechnology. The term gained traction in response to the growing ability of neurotechnologies to record, decode, and infer mental content from brain activity. These technologies include electroencephalography, functional magnetic resonance imaging, and brain-computer interfaces. Although their fully functional forms, capable of reliably “reading” thoughts, do not yet exist, their potential has already exposed a conceptual gap in privacy theory.

Traditional notions of informational and bodily privacy, which are typically designed to regulate access to external representations of a person, such as medical and financial records, location data, and to protect against physical intrusions on the body (for example, strip searches or non-consensual examinations), fail to capture the distinctive exposure created by neurotechnologies, where the information at stake is not cleanly separable from its source (as they are ongoing neural processing) and thus exemplifies what Ienca and Andorno (2017) call the *inception problem*, namely a potential direct access to mental content itself.

This perceived tech-induced need has led scholars to propose mental privacy as a new (sub)concept. Ienca and Andorno (2017) argued that the rise of neurotechnologies requires a “neurorights” framework including the right to mental privacy, defined as the right to keep one’s brain data and mental content protected from unauthorised access.¹²⁸ Similarly, Yuste et al. (2017) called for international recognition of mental privacy as a fundamental right to safeguard individuals from invasive brain-recording and decoding tools. Subsequent literature has refined the notion. For example, Lippert-Rasmussen (2017) linked mental privacy to epistemic access to neural representations, Ienca (2021) and Magee, Ienca, and Farahany (2024) expanded it to include inferences about any mental states drawn from brain-derived data, and Susser and Cabrera (2024) showed that the field remains divided over whether the target of protection is brain-related data, mental content, or the broader integrity of mental life.¹²⁹

¹²⁸ In their paper, they proposed four neurorights: cognitive liberty, mental privacy, mental integrity, and psychological continuity. Mental privacy, which is defined as the right to keep brain-related information from *unauthorised* or *actual access*, therefore directly expands on the access view.

¹²⁹ Note that it is also not fully accepted that mental privacy is best understood as a protection over *information* per se (or data), as some argue that the loss is best understood as something like a bodily violation (Barnard et al. 2024).

What unites these positions is the recognition that existing privacy frameworks were not designed for the epistemic intrusions posed by neurotechnology.¹³⁰ Control-based accounts focus on disclosure and consent, Access-based accounts attempt to pre-empt the information flow; and Contextual Integrity approaches assess the appropriateness of the information flow given the prevailing informational norms of a given context. None of these, however, can maintain a single definition of informational privacy in which neural activity or brain-signals are fully protected while other types of information are protected (conditionally) by pro tanto reason. This is so because mental privacy is typically conceived and defended as a new fundamental human right, whereas informational privacy is typically conceived as a pro tanto right.¹³¹ Mental privacy was therefore engineered as a complementary concept to respond to this new privacy need created by neurotechnology.

Conceptually, mental privacy operates as a protective condition concerning brain-derived information. This move has methodological significance. Mental privacy was not introduced because existing privacy laws had already failed, but because scholars anticipated that they would fail once fully functional neurotechnologies appeared. It represents an instance of proactive conceptual adaptation, creating conceptual resources ahead of empirical necessity to guide future normative reasoning and regulation.

This case provides a methodological precedent for the present project. It exemplifies how conceptual engineering can function as a philosophical counterpart to technological anticipation, ensuring, as Hopster and Löhr (2023) note, that our conceptual landscape evolves in step with emerging forms of human-technology interaction. Just as mental privacy was introduced to anticipate the implications of fully functional neurotechnologies, emotional privacy is introduced here to anticipate the implications of fully functional ERTs. In what follows, I thus make use of this methodological approach in the affective domain.

¹³⁰ Recent national and supranational discussions have concluded that the concept of mental privacy is needed for the proper protection of brain-related content (see for instance Council of Europe (2019)).

¹³¹ In brief, mental privacy's advocates defend the idea that there are never any "good" reasons that would outweigh a person's right to mental privacy (hence why it is often presented as a fundamental right) while there might be good reasons to outweigh someone's right to informational privacy (hence why it is usually conceptualised as a pro tanto right). Consider the workplace. As an employee, you might have a fundamental right to have your brain-related activity safeguarded from your employer or the company, but you do not hold such right regarding your salary or your name.

5.3.2. Anchoring emotional privacy in the social perspective

In the Information Age, we have been prompted to first worry about the information itself, which is in most cases the core of the problem: who loses information, who gains information, and who does what with the newly acquired information. Within this paradigm, we often forget that with the loss or illegitimate acquisition of data come other forms of harms. With emotional privacy, I attempt to show that the kind of losses that are invoked by the prospect of fully functional ERTs include unique kinds of harms, which are interconnected with the functions EEs serve in social life. These harms are to be added to the kind of harms that have already been identified, most of them relating to emotions in particular (Gremsl & Hödl 2022; Menges and Weber-Guskar 2025). To shed light on this, I turn to the social perspective of privacy.

Several scholars have argued that privacy serves interests extending beyond the individual, with works highlighting its constitutive role in specific relations to society in general (Fried 1968; Rachels 1975; Altman 1975; Cohen 2012, 2013; Roessler & Mokrosinska 2013).

Already in 1968, Charles Fried's account, though still situated within an individualistic, control-based framework, marked a clear departure from the classical liberal approach to privacy (viz. Mill [1859]/1985) by placing relationally dependent values such as love, friendship, and trust at the centre of the concept of privacy. On Fried's view, privacy cannot be reduced to a merely individual interest in freedom or autonomy conceived as a retreat from the social world. Rather, privacy constitutes the necessary foundation of social relations, for it enables the informational space on which love, friendship, and trust depend (Fried 1968: 205). The possibility of being human without these relationships, Fried insists, is hardly conceivable, since they form some of the most valuable dimensions of human life. Privacy thus safeguards what is most precious in our moral and social existence: it regulates the flow of information in social interactions, enabling individuals to determine when and to whom they reveal aspects of themselves, thereby protecting the very conditions under which intimacy, trust, and friendship can develop and endure.^{132, 133}

¹³² Note that while it is true that Fried defined privacy in terms of control-based approach (Fried 1968, 482), his insistence on the idea that individuals' having privacy is a condition for their being able to engage in interpersonal phenomena such as trust and love, opening the door to the social privacy perspective which later came to rely on his account (see Roessler & Mokrosinska 2013).

¹³³ The informational privacy perspective, especially its control-based proponents, had already incorporated some of the core elements of a conception of social privacy. Hence, similarly to Fried, Westin (1967), also described privacy as a *dynamic process* that regulates role demands and creates spaces of respite from them.

Whereas Fried establishes privacy as the condition that makes intimate relationships possible, James Rachels (1975) generalises this insight. He argues that privacy's importance cannot be limited to intimacy alone but extends to the entire web of social roles through which persons engage with others. In his essay *"Why Privacy is Important,"* Rachels argues that the significance of privacy cannot be grasped solely through the values of love, trust, and friendship as it requires an analysis of privacy's role in structuring the diversity of relationships that constitute human life. On Rachels' account, each relationship presupposes a distinct pattern of informational access. Privacy is the regulative condition that allows individuals to manage these patterns, controlling who knows what about them in ways appropriate to context: "our ability to control who has access to us and to information about us [shapes] our ability to create and maintain different sorts of social relationships with different people" (1975: 326). What one discloses to a physician about one's health differs from what one shares with an employer or confides in a close friend; without privacy, these distinctions collapse, and all relationships risk flattening into an undifferentiated field of exposure. A loss of privacy in this sense would entail the loss of the very possibility of maintaining the diversity of relationships that gives human life much of its richness. Rachels therefore argues that privacy is indispensable to the very organisation of social life as it protects the plurality of social roles.

A further development of the social perspective is found in Helen Nissenbaum's theory of Contextual Integrity (2004, 2010, 2018), which reframes privacy as the preservation of appropriate information flows within socially defined contexts. On this account, each domain of social life, such as healthcare, education or friendship, is structured by context-specific norms that govern what kinds of information may be shared, with whom, and under which transmission principles (Nissenbaum 2004, 2010). Nissenbaum's approach strengthens the insight, already present in Rachels's work, that privacy is constitutively tied to the maintenance of social roles. Yet it also shifts attention from individual preferences to the integrity of social practices themselves. What matters is whether the informational flows enable or undermine the proper functioning of the context. This structural focus is highly relevant for emotional privacy, since emotional expressions are embedded in normative expectations that regulate how affective information circulates within different social settings.

Fried's and Rachels' postulate – that the value of privacy lies primarily in the virtue of constituting social relations – has also been the stepping stones of other, more recent work, notably by Roessler and Mokrosinska (2013, 2015) and Mokrosinska (2018), who argue that

privacy should be conceived not merely as an individual or interpersonal value but also as a social and political one.¹³⁴ On their view, privacy's function is not exhausted by the management of interpersonal access to information as it also benefits the relations between individuals. Their argument (2013) pushes the social conception of privacy further by allowing privacy to protect elements *within* the relationship and *of* the relationship. In intimate relationships for example, privacy protects both the individual's ability to withdraw from forms of surveillance and intrusion that would erode their ability to form independent judgements and the capacity to "live as mature and responsible citizens in democratic societies" (2013: 775). At the same time, it also protects the social practice of intimacy itself, which "has to be protected from social and societal publicness, for the function and intention of these relationships can be realized only if their privacy is protected" (2013: 775). Privacy therefore serves both individual and collective interests; individually, it enables autonomy and the cultivation of personal relationships, while collectively it sustains social and political practices.¹³⁵ Roessler and Mokrosinska hence expand the concept of privacy from a condition of interpersonal relations to a structural requirement of social and political life (Roessler and Mokrosinska 2013; Mokrosinska 2018). While earlier relational accounts could still accommodate the early classical liberal view of the self, the social conception of privacy requires a decisive break from this atomistic view and rests instead on a social ontology of the self.

Early liberal understandings of privacy, grounded in the idea of a self that exists prior to and apart from social relations, can no longer capture the full spectrum of privacy violations that occur in contemporary, interconnected societies. As Michael Walzer noted, "[t]he individual who stands wholly outside institutions and relationships and enters into them only when he or she chooses and as he or she chooses: this individual does not exist and cannot exist in any conceivable social world" (Walzer 1984: 324). Whereas the early liberal view of privacy protected the individual's capacity to withdraw from social and political life, the social conception of privacy grows out of a non-atomistic understanding of identity (Mokrosinska 2018). On this view, identity is *constituted* through social roles, relationships, and contexts (Durkheim [1897]1952; Goffman 1959; Boyd 2008).¹³⁶ Individuals are therefore situated beings whose sense

¹³⁴ This argument does not only follow Fried and Rachels' work but also to some extent that of Gavison (1980), Reagan (1995), and Solove (2008).

¹³⁵ Mokrosinska (2014, 2018) pushes the argument in favour of privacy as a political value further, claiming that privacy preserves autonomy not as an individual good, but as a condition of democratic life. If identity and autonomy are inherently social, then privacy cannot be understood as disengagement from social life, but as its precondition.

¹³⁶ Sandel (1982), for instance, in a direct reply to Rawls' work on liberal justice, criticised the limits of liberalism by arguing for a redefinition of individual identity which should be conceived as deeply tied to social and communal

of self is relational, shaped by social norms, and influenced by the institutions and structures that organise collective life.

Social accounts of privacy take this social ontology as their foundation. The possibility and legitimacy of a social conception of privacy depend precisely on recognising that the self is socially constituted. If individuals are socially constructed, they can only exercise autonomy and develop their identity through their relationships with others rather than by withdrawing from society. Privacy's function, on this account, is to protect those relationships and the practices through which autonomy and identity are realised (Mokrosinska 2018).

Importantly, this shift does not lead social-privacy theorists to conclude that privacy does not protect individual interests. Rather, they maintain that privacy's role includes the protection of individual interests, but *extends* to the protection of social life itself. As Solove explains, privacy's value lies not only in shielding individuals but also in sustaining the activities and practices that make both individual and collective life possible. "Although privacy certainly protects individuals," he writes, "this does not mean that privacy is an individualistic right. Understanding privacy as having a social value – the benefits it confers on society by enhancing and protecting certain aspects of selfhood – does not oppose the individual to the community or value privacy solely in terms of its benefits to particular individuals. [...] The protection of privacy shields us from disruptions to activities important to both individuals and society" (Solove 2008: 92-93).

On this view, privacy is not only a personal safeguard but a constitutive condition of social life. It sustains the network of relationships upon which both individual flourishing and collective life depend. I propose to apply this rationale to the emotional life, as the emotional life is an essential part of what gives individuals a sense of self, while also being (at least partly) socially constituted. Accordingly, emotional privacy should be conceived within this broader social framework. Emotional privacy therefore is best understood as protecting individuals' expressive ability over affective information, ensuring that the *possibility* for expressive ability is preserved. Yet if we maintain this social perspective, and accept that emotional expressions and their causes are embedded in dynamic social processes, both influencing and being influenced by them, then emotional privacy cannot be regarded as something wholly possessed by individuals. Rather, it

contexts rather than independent from it. Feminists and communitarians have also criticised the classical liberal view of identity and respectively pushed for a reconceptualisation of one's identity as intrinsically social and shaped by personal and formal relationships (MacKenzie & Stoljar 2000), and have highlighted the crucial role of community, tradition, and shared values in shaping moral and political life (Etzioni 1999; Taylor 1992).

is a condition that is socially regulated, emerging from natural dynamics of affective life and from privacy mechanisms that individuals enact through their expressive capacities to modulate the flow of information about their affective states.

5.3.3. Defining emotional privacy

I have already mentioned that Control-based accounts focus on whether a person succeeds in regulating an information flow in a particular moment. Within this view, the control-based variants typically treat “control” as the central concept of informational privacy, where control is understood as the *successful regulation* of access to information p. They also usually (implicitly) assume that a person A possesses the ability required for such control. This assumption invites them to take for granted that A stands in a position, as a matter of general ability, to control access (Classic control), to block access attempts (Negative control), or to serve as the appropriate informational source (Source control). This is perfectly understandable, as outside the affective domain, it generally makes more sense to simply focus on the notion of control while implicitly assuming the ability to control. Hence, Control-based accounts operates at the control level, asking whether, in a particular moment, A succeeds or fails in regulating information p given some conditions. However, this distinction between “control” and the “ability to control” is what ultimately matters in the affective domain.¹³⁷

Building on the social functions of emotions and emotional expressions (see Chapters 3 and 4), and on recent work done in the social privacy perspective (§5.3.2), I argue that *emotional privacy* is best understood as a relational and social condition whose significance must be examined in terms of affective communication. I postulate that it is best to understand emotional privacy as operating at two interconnected levels: the actual level and the standing level. At the actual level, a person enjoys emotional privacy over a particular expressive event only insofar as she can *successfully exercise* expressive discretion in that moment. When a spontaneous EE is produced, for instance, the person loses actual emotional privacy with respect to that specific disclosure. Such losses are not only common in ordinary social life, they are also inevitable. Yet the person may retain

¹³⁷ Menges’ Source Control account is the one that comes closest to it since he distinguishes Source Control from “effective control”. However, to *be* the (right kind of) source of a given information flow is what ultimately matter for his account. Additionally, the problematic identified in his account is still found at the point of a given “actual” safeguard or lost of control over the information p’s flow (as the right kind of source of that information flow).

emotional privacy at the standing level, because she still possesses the *ability* to exercise expressive discretion. This refers to the ability of a person to use certain types of emotional expressions to disclose, conceal, reinterpret, or refute affective information. It is only when this expressive ability is weakened or defeated that emotional privacy collapses at the standing level. That normative concern will be addressed in the next chapter.

I therefore define emotional privacy as

Actual Emotional Privacy: Actual emotional privacy obtains when person A has expressive control over her affective information p.

Standing Emotional Privacy: Standing emotional privacy obtains when person A has the expressive ability to control her affective information.

This two-level structure reflects the nature of emotional life. Emotional expressions are either voluntary or involuntary. People do not enjoy full control over all their expressions, and they regularly experience affective leakages in ordinary social interactions. “Actual” losses of emotional privacy are therefore a common feature of everyday life. Yet, even when they lose (actual) control, they still retain the ability to control, as they can anticipate, manage, or refute meaning carried by EEs even when those actual losses occur.

By focusing on ‘control’ and the ‘ability to control’, emotional privacy naturally aligns with the control tradition of informational privacy in its broad sense. However, it does not align neatly with any of the Control-based variants seen above. In the affective domain, control cannot be exercised unless the person has access to a repertoire of controlled expressions and privacy mechanisms through which she can shape the affective informational environment (see following §5.4). This implies that actual control over affective information p may fail even though person A’s expressive ability is preserved, but actual control can never be retained if the expressive ability is lost. This is why emotional privacy departs from the traditional Control approaches. Emotional privacy, contrary to mainstream Control accounts, begins with the question of whether the person possesses the expressive ability that allows her to engage in expressive discretion at all.

Imagine a person sitting in a quiet corner of a café with a close friend, discussing a sensitive matter. Let's call him Daniel. Daniel chooses his words carefully, lowers his voice and keeps an eye on his surroundings. In these conditions, he retains control over who can hear the information *p* and when it becomes accessible. During the conversation he leans forward and whispers a particularly delicate remark. His friend is unsure about whether he heard him correctly and repeats Daniel's words aloud in a clearer, louder voice in order to confirm what he thinks he heard. A customer seated at a nearby table hears the repeated remark. In this scenario, Daniel loses (actual) control in the classic form, since he has not succeeded in regulating access to *p*. He loses negative control, since he was not in a position to prevent the other customer from hearing what he intended to keep private. He also loses source control in Menges' sense, since the information flow that reaches the third party does not originate from Daniel's communicative agency but from his friend's. Yet Daniel's "ability to control" remains intact. He can lower his voice again, move to another table, shift to coded phrasing or postpone the conversation. He has lost actual control over information *p*, but he retains the ability to regulate the flow of information through his speech in ordinary circumstances.

Now suppose that during another visit to the same café an employee slips a drug into Daniel's drink. As the drug takes effect, Daniel's cognitive and motor coordination begins to deteriorate. His speech becomes erratic and he begins revealing sensitive and personal information aloud. Several nearby customers hear the unintentionally loud disclosure. In this state, Daniel has lost actual control because he has lost the ability on which his actual control depends. His underlying ability to regulate his speech, to correct mistakes, or to rephrase information has been removed. Control is not simply unsuccessful in this instance. It has become unavailable to him altogether because the basic expressive resources (i.e. his expressive ability) that make control possible have been compromised by the drug's effects.

These two scenarios show the distinction between control (viz. actual emotional privacy) and the ability to control (viz. standing emotional privacy). People routinely lose actual control over their EEs in everyday social life. A blush, a wince, or a momentary tightening of the jaw can reveal more than a person intends, yet these episodes do not undermine the person's broader expressive ability. She can still reshape the interaction, reinterpret what has been seen or correct the impression through intentional expressions. By contrast, when expressive ability itself becomes unavailable, emotional privacy collapses at the standing level. This is structurally analogous to Daniel's second scenario. When a person cannot manage how affective

information is disclosed or interpreted, she lacks the ability that standing emotional privacy requires. Importantly, while expressive ability is a *necessary* (pre-)condition of actual emotional privacy, it is not sufficient. For instance, an emoter's own affective leaks can thwart their own (actual) control or perceiver's ordinary mindreading skills may also thwart that actual emotional privacy.

Chapter 6 will shed light on why this distinction becomes essential for understanding what is truly at stake on a normative level with emotional privacy and the use of fully functional ERTs, but in a nutshell, it will help us distinguish between legitimate and wrongful losses of emotional privacy.

In this section, I have briefly introduced the idea that people can manage their emotional expressions through some 'privacy mechanisms', but I have not yet defined them. I do so in the following section, where I aim at deepening this conceptualisation of emotional privacy through an analysis of its functions within affective communication channels.

5.4. Emotional privacy in affective communication channels

I have introduced emotional privacy as a subconcept of the informational privacy cluster and as aligned with the broad Control-based understanding of privacy. I have noted that Control accounts typically presuppose the ability to control (standing) while concentrating their analyses on particular instances of success or failure of control (actual). In the affective domain, by contrast, both actual and standing levels matter, even though the standing level is ultimately primary. The remainder of this chapter examines how actual emotional privacy is sustained through the functions of emotional privacy, understood as privacy mechanisms. The next chapter then explains why the standing level is normatively decisive, since the successful deployment of these mechanisms depends on it.

These functions of emotional privacy are best understood as enabling and regulating emotional expressions' functions. Whereas 'first-order' functions of emotional expressions concern what expressions do in interaction, the functions of emotional privacy concern the background conditions under which those first-order functions can succeed. In particular, emotional privacy sustains a workable space of calibrated opacity in affective interactions, maintaining neither full

secrecy nor full transparency. This space of calibrated opacity is where individuals can partially shape the legibility of affective information across settings and audiences.

Many things may cause informational opacity in affective communication channels. The obvious one is noise.¹³⁸ Others include expressive conventions, affective leakages, and privacy mechanisms. These may increase the affective *opacity* in a given communication channel, leading to a situation where the perceiver has only partial information or not enough information to successfully infer the meaning of the EEs. In turn, thus, affective *legibility* is seen as dependent on the degree of opacity present in affective communication channels.

Together with expressive conventions and affective leakages, privacy mechanisms create an interactive process that is best understood as a tense dialectic, where individuals play both the roles of emoters and perceivers in each interaction, negotiating what they reveal and reflecting on what they infer. Emoters may thus partially shape this legibility through the use of privacy mechanisms. Such mechanisms do not simply conceal or reveal; rather, they are strategic techniques regulating affective legibility; making emotional life more or less accessible. I distinguish two broad types: disclosure mechanisms, which typically decrease opacity by enabling selective access under chosen conditions, and concealment mechanisms, which typically increase opacity by limiting access to affective information.

5.4.1. Expressive conventions and affective leakages

Before turning to the privacy mechanisms that individuals may deploy to preserve emotional privacy in affective communication channels, I first overview two important background dynamics that put pressure on an emoters' own negotiation of emotional privacy within affective communication channels. First, expressive conventions need to be acknowledged as they often govern when and how emotional expressions should be produced. These conventions tell us when to smile and when to look angry, but correlatively they also guide us in anticipating when smiling and looking angry will be looked upon or disapproved. In that sense, they also participate in regulating the negotiation of emotional privacy by guiding people around what is expected but also what would be considered inappropriate.

¹³⁸ In this section, I ignore the role of noise to focus on the different types of functions. See Chapter 2, §2.2.3 for a discussion of noise in affective communication channels.

Second, while some expressions are intentionally produced to support interactional goals, others arise *spontaneously*. I have already discussed the latter in Chapter 4 (see §4.2.3), and will now refer to them as ‘affective leakages’. Affective leakages are micro, partial, or complete EEs that are not intentionally produced by the emoters, giving perceivers access to meaningful affective information that is not deliberately shared. Sudden physiological changes, such as blushing, pupil dilation, sweat, or shaking are classic examples of changes that produce affective leakages. Affective leakages may also occur through unintentional behavioural changes such as jumping or screaming out of fear, clenching one’s jaw out of anger, or even rolling one’s eyes. Some affective leakages are non-intentional in the sense that they simply cannot be controlled or suppressed by emoters, while others are simply “let out” in Green’s sense (2007), meaning that, even though they are typically classified as spontaneous, they could have been pre-empted by the emoters if they had wanted to. However, this presupposes that control over some types of affective leakages requires important “emotional labour”.¹³⁹ Following emotional labour theory,¹⁴⁰ it is possible that, given the ‘labour’ required to prevent affective leakages, in situations where they would be numerous, for instance, it is unlikely that an individual will have the emotional bandwidth to successfully pre-empt or suppress *all of them*, hence the second type may merge with the first one, transmitting meaningful information to perceivers without emoters’ deliberate intention.

This means that emoters’ management of their own affective legibility is always imperfect. Expressive conventions and affective leakages are background pressures shaping the negotiation of emotional privacy within affective communication channels. With this in mind, I now specify what privacy mechanisms are and how they function in affective communication. In a nutshell, although they rely on different strategies, they all enable and protect the negotiated and *in situ* interpretation of expressions through calibrated opacity.

5.4.2. Privacy mechanisms

The privacy mechanisms that I introduce below are inspired by Irwin Altman’s work (1975) on boundary-processes. Altman’s work is best categorised within the social perspective of privacy

¹³⁹ Emotional labour theories usually focus on the control that one can hope to have over one’s affective life through concepts such as “emotion regulation”, “internal management”, “emotional adjustment”, or “emotional contagion”, mostly studying the extent to which we possess that control. See Wilms, Lanwehr, & Kastenmüller (2020); Eisenberg (2000); Needham, Mastracci & Mangan (2017) for more on this particular point.

¹⁴⁰ An interesting insight from emotional labour theories is that the emotional labour required to control these leakages varies depending on the type of interaction, the context, and the interlocutors. These are all recognised as primary determinants of the degree of control one can hope to have over affective leakages (Xu & Fan 2023).

as he conceives privacy as a relational dynamic where privacy facilitates the management of social interactions by regulating desired and achieved levels of interaction.¹⁴¹ Altman's account, and the framework he developed from it, provides useful conceptual tools that can be reshaped into privacy mechanisms within affective communication channels. Altman argues that privacy needs to be conceived as:

[An] interpersonal boundary-control process, which paces and regulates interaction with others. Privacy regulation by persons and groups is somewhat like the shifting permeability of a cell membrane. Sometimes the person or group is receptive to outside inputs, and sometimes the person or group closes off contact with the outside environment (Altman 1975: 10).

On his account, understanding privacy as a dynamic concept is key to understand the regulation of interpersonal relationships and the self. Privacy boundaries enable us to mark a division between ourselves and the world. Altman's privacy approach does not only see privacy as a withdrawal process but as an interplay between withdrawing from social interaction and engaging in social interactions.¹⁴² His framework is conceived around an optimisation function between the desired level of contact with others and the achieved level of contact with others (Altman 1975). The desired level of contact is understood to be constantly changing over time, shaped by social contexts and types of relationships. Hence, the value of privacy is found in this role of interface between the self and the non-self. Privacy is achieved when the dialectic between the self and others – enabled by the boundary mechanisms (and their respect) – is obtained, meaning that the desired level matches the achieved level of contact. Altman's privacy-boundary account identifies specific social privacy-mechanisms that individuals, groups, and societies use as a way to express and achieve their desired level of interaction in different contexts and relationships (Altman 1975: 30). For instance, individuals can use their own body as a space-regulatory indicator during an interaction, communicating the level of desired contact they aspire to achieve. Additionally, companies, for instance, might design the office spaces in a way that shape the general work environment (i.e. flex-offices, open offices, insulated/separated offices).

I see the tense dialectic between expressive conventions, affective leakages, and emoters' privacy mechanisms in a similar way. I thus rely on Altman's core idea that individuals deploy certain

¹⁴¹ Altman's view of privacy itself builds both on Simmel and Westin, taking inspiration from Simmel's work on the "dialectic quality of social exchange" (Simmel 1950, in Altman 1975: 19) and Westin's analysis, composed of four types (solitude, intimacy, anonymity, and reserve) and functions of privacy (personal autonomy, emotional release, self-evaluation, limited and protected communication) (Westin 1967, in Altman 1975: 18). Moreover, the idea of a dynamic process regulating interaction is also found in other's work (see Smith et al. (1969), Schwartz (1968), McGinley (1959), Bates (1964), Jacobs (1961), Jourard (1966), Rapoport (1972)).

¹⁴² His framework also allows for feedback and readjustment during a given interaction (Wisniewski & Page 2022, 21).

types of privacy mechanisms to obtain the desired level of *contact* and repurpose it for the idea that individuals deploy privacy mechanisms within affective communication channels to obtain the desired level of *informational opacity*. Below I discuss the role of disclosure and concealment within the nonverbal emotional realm.

5.4.3. Disclosure privacy mechanisms

Disclosure is generally understood as sharing one's information with others. In turn, *self-disclosure* generally refers more specifically to the act of sharing one's own information, be it emotions, motivations, goals, thoughts, personal information, or past, present, and future experiences with others (Westin 1967; Derlaga & Berg 1987). Disclosure and self-disclosure are also usually seen as *intentional* revelations. But while in other contexts both disclosure and self-disclosure may be considered "full disclosure", in the realm of affects, this would be erroneous and misleading. When it concerns emotional life, disclosure is never "full", not only because the emotional experience is hardly shareable in its entirety, but mainly because individuals do not, and cannot, disclose every single emotional experience.

Consequently, from now on, I will refer to the privacy mechanism of disclosure as *selective disclosure*, to remain closer to reality. Selective disclosure thus enacts emotional privacy by making one's affective legibility conditional, contextual, and recipient-relative. It is *not* a loss of privacy but a mechanism through which privacy's broader social function may be performed, i.e. it partly governs the circulation of affective information so that interaction remains intelligible. Hence in that sense, spontaneous EEs, which are typically not controlled, are not part of selective disclosure as this mechanism is intentionally deployed.

Emotional privacy thus protects the conditions of affective legibility, and selective disclosure functions as one of its tools: it permits partial, targeted legibility for chosen perceivers, facilitating their interpretation.

5.4.4. Concealment privacy mechanisms

Concealing information is commonly regarded as *withholding* information from others. From this perspective, concealment mechanisms are often regarded with suspicion in debates about

privacy, trust, and honesty (Bok 1983).¹⁴³ Taken as such, it is easy to assume, then, that concealment simply amounts to deception. But while deception is commonly defined as inducing false beliefs, concealment in affective life typically functions differently. For one, individuals do not always aim to convince others of (affective) untruths, nor do they necessarily hide their “true” selves. Rather, they rely on this mechanism to present the self that best fits a given context or relationship. For another, despite the given fact that conveying emotions or social motives that are not currently felt or desired may sometimes mislead, it also, at least just as often, preserves social appropriateness, protects dignity, or sustains intimacy. Similarly to selective disclosure, then, the extent to which concealing one’s information is acceptable, permissible, or encouraged is thus often seen as dependent on the relationship the individual has with others (as the nature of the relationship usually models a significant part of the exchange of information (Rachels 1975; Fried 1968)) and on the context.

Within the concealment mechanisms, I distinguish two types. I refer to the first as “direct concealment”. This subcategory of concealment concerns privacy mechanisms that are used to *signal* to the other that the emoter does not wish to share inner affective states or social motives. Practical examples of these mechanisms include avoiding eye contact, neutral EEs, and increasing or maintaining physical distance. In this case, there is an overt attempt at signalling a desire to not share affective information.

Broadly conceived, this type of mechanism maintains, increases, or even decreases¹⁴⁴ the level of informational opacity presents in the affective communication channels. Moreover, it is not necessarily the case that maintaining or increasing informational opacity in the interactions indicates an unsuccessful attempt at using privacy mechanisms, as one might want to have this level of opacity.

The second subcategory of privacy concealment mechanisms I refer to as “indirect concealment”. These are cases where the emoter *does not signal* to the other that *she does not want*

¹⁴³ The “nothing-to-hide” argument against privacy, for instance, was at the centre of these debates. But as Sissela Bok rightly notes: “[...] while all deception requires secrecy, all secrecy is not meant to deceive” (Bok 1983: 7).

¹⁴⁴ Even though using concealment to decrease the level of informational opacity seem counter-intuitive, “direct concealment” may have this result, as the privacy mechanisms deployed are “out in the open”, meaning that perceivers may easily interpret that this person, for instance, “does not wish to share”. In that sense, it cannot decrease informational opacity regarding the causes of emotional expressions directly, but it may provide information on the willingness to share them (or not).

to *share* her inner affective state or social motive.¹⁴⁵ This is done by relying on the success of one's intentional emotional expressions to convey what the emoter judges as the best expression even though they do not sincerely experience them. Indirect concealment mechanisms can thus be used to follow conventional norms and express what is expected rather than what one feels, or to make clear a given stance, commitment, goal, or desire. Hence it might also be done to preserve a relationship or one's appearance. These are cases where one may *choose* to conceal their inner affective states and/or social motives and instead use an emotional expression that fits the context better. Typical cases would be to show the expression of sadness for someone's loss even though one does not feel sad, anger to support someone's side in a conflict even though one does not feel angry, or even happiness for someone's wanted pregnancy even though one does not feel happy.

These privacy mechanisms are functionally important as they regulate the inferences and predictions perceivers can make from emotional expressions, allowing individuals to manage the strategic ambiguity of their own emotional lives. In standard interactions, emotional privacy does not merely protect individuals from exposure but also secures the social and epistemic conditions under which affective legibility occurs. Emotional privacy's role is thus to safeguard emoters' ability to manage their affective legibility.

An important point is that all privacy mechanisms rely on the deployment of *intentional* EEs. I have already mentioned in Chapter 4 that *intentional* EEs serve important social functions. Empirical research has shown that EEs serve a broader range of interpersonal and regulatory purposes (Van Kleef 2016; Fischer & Manstead 2016; Kumar & Senthil 2021). Expressing love, compassion, and care, for instance, conveys affective commitment, fosters intimacy, and enables cooperation in close relationships (Gonzaga et al. 2001; Ellis & Malamuth 2004). Conversely, expressing anger can rectify social misunderstandings, pinpoint norm violations, clarify expectations, or challenge status arrangements (Lerner, Goldberg & Tetlock 1998; Averill 1982; Fischer & Roseman 2007).

Particular attention to intentional EEs is necessary to understand their role in the successful deployment of privacy mechanisms. While direct concealment may be achieved by at least allowing the emoters to remain neutral in their expressions, indirect concealment means that

¹⁴⁵ Classic examples of this type that have been largely discussed in the literature on emotional expression are the "survival" type of cases, where for instance one has *advantage* in showing anger (pretended emotion) rather than fear (true emotion) for the sake of their own survival (see Darwin [1872] 1965).

emoters are for instance able to express anger *without* experiencing it, which may, for instance, support a mother's education of her child by prompting him to reflect on his behaviour or help friends signal that they do not endorse a particular behaviour, even though in both cases the emoters are not actually angry. In selective disclosure, it allows people to emphasise and carefully craft the way they will express their emotional experience. The existence and the possibility to successfully deploy intentional EEs is what these privacy mechanisms are about.

Below I turn towards the value of emotional privacy, and overview three primary functions of the privacy mechanisms at three different levels of social interaction (individual, interpersonal, and societal).

5.5. The functional value of emotional privacy

In this section, I focus on the value of emotional privacy. I show that the privacy mechanisms are not just options at the emoters' disposition but are *integral* to the regulation of social life. They should be recognised as necessary preconditions to the enactment of expressive discretion because they are functionally essential. I argue that this is valuable on three different interaction levels, namely individual, interpersonal, and societal. At the individual level, they allow individuals to manage the self they want to present; at the interpersonal level, they enable the management of role plurality and ensure adherence to role-bound norms of expressiveness; and at the societal level, they allow *society itself* to regulate, ensuring cohesion. These different levels are not to be seen as mutually exclusive but rather as complementary, shedding light on different levels of impact.

5.5.1. Preserving self-presentation

At the individual level, philosophers and sociologists have often argued that privacy primarily helps individuals to conceal what might be called their "uncivilised inner life." By this, they do not refer to something pathological,¹⁴⁶ but to the ordinary stream of impulses, desires, and emotions that fail to conform to established social norms. A persistent assumption is thus that society depends on such concealment. For instance, Simmel said that if one's "subjective, impulsive, intimate matters [...] were exposed, they would drive everybody into the insane asylum" (Simmel [1908]1950: 311–312). Similarly, in "Concealment and Exposure", Nagel

¹⁴⁶ The concealment of pathological desires – such as those of paedophiles – cannot be assimilated into the same category. Here concealment may serve the individual, but it directly endangers others and erodes the very social goods that privacy is meant to protect.

posits that “there is much more going on inside us all the time than we are willing to express, and civilization would be impossible if we could all read each other’s minds” (Nagel 1998: 4).¹⁴⁷ It is difficult to deny that humans constantly experience emotions – envy, irritation, hostility, lust – that, if rendered fully transparent, would strip people of the possibility to decide what they choose to reveal, to whom, and when (Westin 1967). Emotional privacy ensures that the raw material of affective states and social motives (often disordered, fleeting, or socially misaligned) does not overwhelm the civility required for social life.

A second interconnected rationale is that concealment mechanisms enable moral reflection and value-commitment. As previously discussed in Chapter 3,¹⁴⁸ emotions play an essential role in individuals’ moral reasoning. Emotional privacy protects this role by giving individuals the space required to turn their attention inwards, to their inner life. At times, individuals experience emotions or have social motives that directly conflict with moral principles they hold dear. In that sense, the active suppression of such inner states can be morally admirable in itself, insofar as it demonstrates a prioritisation of enduring moral commitments. Emotional privacy thus allows people to work on themselves when they experience internal (moral) conflicts, enabling the development of important virtues. Concealing a fleeting attraction towards a friend’s partner, for instance, can develop into a morally appropriate act of self-regulation. Opacity in such a case preserves the friendship and the relationship, perhaps also forging a core value that will follow the individual for the rest of their life. By shielding “uncivilised” elements of our inner lives or interior dilemmas from immediate exposure, privacy enables individuals to reflect, recalibrate, and ultimately sustain their moral commitments within social and moral communities.

A third one is highlighted in the work of Erving Goffman (1959, 1967), who notes that individuals meticulously craft the self that they best see fit, with that ‘fittingness’ being modulated by the identity of the perceivers, the context, the time, etc. (Goffman 1959). On Goffman’s view, the purpose of self-presentation is to shape how others interpret the situation, so that they respond in ways that support one’s aims, roles, or identity in that moment.

These different types of interests are best recast as benefits that result from emotional privacy’s role in protecting expressive autonomy, which sustains the crucial ability to present the self that

¹⁴⁷ Mokrosinska (2014: 371) discusses a similar idea, citing also Murphy (1964), Merton (1964), and Altman (1975).

¹⁴⁸ See Chapter 3, §3.3.

one sees fit, regardless of the motives behind. Emotional privacy in affective communication channels is an ongoing, practice-like management of legibility and opacity. In this domain, emotional privacy safeguards those interests (hiding uncivilised inner life, enabling moral reflection and value commitments, and impression management)¹⁴⁹ not by guaranteeing secrecy, but by preserving a strategic background of partial opacity that keeps expressive meaning negotiable and answerable to context. Privacy mechanisms enact this role. Through calibrated concealment and disclosure, emoters adjust perceivers' access to affective information, exercising autonomy over their affective and social lives, and more specifically enabling individuals' self-presentation abilities.

This ultimately points to the relation between privacy and autonomy, and this relation is found at the heart of most theories of privacy. This relation, as regards to the ability to present the self that one best sees fit, is also discussed by several authors. On Fried's relational view of privacy (1968), for instance, respecting persons involves recognising their entitlement to govern access to themselves. On Reiman's account (1976), which relies on Goffman's work on self-presentation, privacy is seen as "an essential part of the complex social practice by means of which the social group recognizes and communicates to the individual that his existence is his own" (1976: 39). On Julie Cohen's view (2012), privacy "shelters" the situated boundary-work through which self-determination develops.

However, here I aim at stressing its significance within the emotional paradigm more particularly. Recontextualised to affective communication channels, self-presentation is about the maintenance of calibrated opacity that prevents others – humans or institutions – from rendering the person fully legible to others and thus affectively transparent. This preserves a person's capacity to shape how they are taken up by others across settings, retaining the ability to frame what their expressions count as *in situ*.

5.5.2. Managing social role plurality and ensuring conformity

At the interpersonal level, privacy mechanisms manage social role plurality and ensure conformity with established norms of expressiveness, often related to these roles. Most societies

¹⁴⁹ For the sake of conciseness, when referring to these different types of interests at the individual level (protecting uncivilised inner life, allowing moral reflection and value commitments, and enabling impression management), I will abridge them by saying that emotional privacy preserves "self-presentation".

prescribe specific codes of affective expressiveness that dictate how people should act and react in particular social contexts (Ekman & Friesen 1969; Matsumoto 1990; Matsumoto, Yoo, & Fontaine 2008). Such rules may call for restraint – such as suppressing irritation in professional settings – or, conversely, for expressiveness, as when service workers are expected to smile to project friendliness (Matsumoto & Kupperbusch 2001). Individuals often follow, consciously or unconsciously, the social norms of expressiveness as these rules usually sustain social practices and norms, ensuring social conformity and acceptance.

Conformity to these norms is not about hiding what one feels, but about performing a contextually appropriate self (Schwartz 1968). By maintaining partial opacity, privacy mechanisms let individuals align their emotional expressions with socially expected forms of conduct. This not only allows individuals to keep distinct role-sets (colleague, friend, client, stranger) from collapsing into one another, but it also enables the compartmentalisation and rapid role-switching without friction (Goffman 1959; Altman 1975; Stryker & Burke 2000). Relationship-specific disclosure mechanisms make this diversification actionable: who may see what, when, and in which setting (Westin 1967; Rachels 1975).

Hence it is generally accepted, and assumed, that people's behaviours, demeanour, and personality traits tend to differ from one social setting to another as people try to adjust (Goffman 1959; Matsumoto, Yoo, & Fontaine 2008). Goffman's notion of "face-work" illustrates this phenomenon by highlighting how individuals *perform* contextually appropriate selves to maintain the smooth functioning of interactions and to appear in the ways they prefer (Goffman 1967). Importantly, this entails that individuals spend lots of cognitive effort *adjusting* their EEs with the socially expected ones so as to ensure acceptance and integration (Needham, Mastracci & Mangan 2017). This emotional labour can be seen as the counterpart of affective leakages (Hochschild 1983).¹⁵⁰ Individuals present themselves differently across settings, tailoring demeanour and expression to sustain the smooth functioning of interactions (Schwartz 1968; Goffman 1959, 1967; Tcherkassof 2018). This is only possible through the strategic deployment of privacy mechanisms, which allow this role-management to exist, ensuring the preservation of social conformity within these roles.

¹⁵⁰ Hochschild's (1983) concept of emotional labour highlights how this effort often involves managing inner feelings so that outward expressions conform to institutional and cultural norms.

5.5.3. Enabling social regulation and social cohesion

Within affective communication channels, privacy mechanisms of selective disclosure and concealment serve society's interests by enabling social regulation and ensuring social cohesion (societal level). A stable and functional society requires both social regulation and social cohesion to prosper. By social regulation, I mean the ensemble of public norms, roles, and procedures that steer behaviour predictably, allocate responsibilities, and proportion sanctions so that collective goods can be produced and conflicts managed. By social cohesion I have in mind a Taylorian concept, in which there is a society-wide willingness to cooperate under disagreement, anchored in mutual trust, shared identification with common purposes, and routinised means of containing and repairing conflict (Taylor 1992). Society's interests are best served when both are jointly sustained, and I argue that at present, emotional expressions are integral to each.

As signals, they mark the salience of norms, calibrate proportional responses, and confer or withhold legitimacy on decisions and directives: indignation flags transgression and may justify censure; reassurance typically lowers social temperature and sustains compliance; remorse invites a moderated response rather than escalation (Bourgeois & Hess 2008; Hess & Fischer 2013).¹⁵¹ At the same time, shared emotional expressions underwrite social cohesion: mourning and celebration sustain common identification; apology and forgiveness rituals enable repair and re-entry after wrongdoing; (in the right cultures) everyday cues of warmth and restraint facilitate coordination and keep cooperation attractive. In short, emotional expressions do not merely reflect a society's order; they are key mechanisms through which that order is enacted and maintained.

However, the regulatory and cohesive force of EEs is conditional on how they are transmitted through affective communication channels.¹⁵² When affective life becomes fully legible, society bears a cost. Emotional mishaps are exposed, and the guiding functions of intentional expressions lose their force. As a result, regulation weakens, and social cohesion gradually erodes. Selective disclosure and concealment are the constitutive safeguards that secure the 'middle zone' society needs. Therefore, within affective communication channels, privacy mechanisms of selective disclosure and concealment serve society's interests precisely because they are constitutive of social regulation and cohesion on which healthy societies depend.

¹⁵¹ See Chapter 4, specifically §4.3.2 for more empirical work supporting these examples.

¹⁵² By this, I am referring to those socially regulated pathways that carry emotional signals across face, voice, posture, textual tone, ceremonial display, and even platform affordances.

Removal of these mechanisms would prevent EEs from performing their regulatory and cohesive work.

What these three levels of interaction highlight is how intentional emotional expressions are the important instruments of everyday coordination. Through them, people deploy privacy mechanisms to help regulate the interactive process and manage the degree of opacity in affective communication channels. This protects multiple core functions of EEs, at different interactive levels: at the individual level, it protects individuals' self-presentation, at the interpersonal level it sustains the differentiation of roles and ensures the possibility for norm conformity, and at the societal level it allows for social regulation and social cohesion. It is through this regulation, secured by emotional privacy, that the affordances needed for emotional expressions to fulfil their social functions are preserved. I will explore the potential consequences that would result from technosocial contexts in which these functions are not preserved in the next chapter.

5.6. Conclusion

This chapter has introduced and developed the concept of emotional privacy as a subconcept of informational privacy tailored to the affective domain. I argued that by protecting individuals' expressive ability, emotional privacy protects the interactive practice of emotional expression through which affective meaning is negotiated *in situ*.

I started by investigating how privacy concepts historically adapt to technological change. I showed that successive waves of innovation have repeatedly required refinement of existing privacy frameworks, prompting the introduction of new (sub)concepts where older formulations became insufficient. This established the methodological basis for treating privacy as a cluster concept open to conceptual expansion. Drawing on the case of mental privacy as an example of proactive conceptual adaptation in anticipation of emerging technologies, I defined emotional privacy as the condition under which individuals retain expressive ability over affective information, that is, when they can shape the degree of opacity in affective communication channels.

I then examined how emotional privacy functions within affective communication channels. I first identified two background pressures (i.e. expressive conventions and affective leakages) that shape how affective information becomes legible. I then introduced privacy mechanisms through which emotional privacy is enacted in practice: selective disclosure, which allows partial and context-sensitive revelation of affective states, and concealment, which regulates or redirects affective legibility. Together, these mechanisms sustain a space of calibrated opacity between full transparency and full opacity. Emotional privacy is not a barrier to social life but a condition of its intelligibility and stability. It constitutes the precondition that allows emotional expressions to perform their social functions by guiding interaction, signalling commitments, coordinating behaviour, and maintaining social order.

Finally, I showed that the value of these privacy mechanisms extends across three levels of social life. At the individual level, emotional privacy supports self-presentation by allowing individuals to shape how they appear to others. At the interpersonal level, it preserves the differentiation of social roles and expectations across contexts. At the societal level, it contributes to social regulation and cohesion, enabling affective communication to guide behaviour and maintain shared norms. The upshot is that emotional privacy is an integral part of the infrastructure of social life.

This provides the conceptual groundwork on which Chapter 6 relies on to articulate the moral conditions under which emotional privacy should be protected. It then uses these normative criteria to evaluate the prospective use of ERTs. The aim is to articulate standards by which we can judge when emotional privacy has been violated and to mark the limits that any acceptable technology must respect.

6. Emotional privacy as a moral right? Normative foundations and technological implications¹⁵³

6.1. Introduction

In the preceding chapter, I re-engineered the concept of emotional privacy and articulated its core functions. Chapter 5 showed that emotional privacy secures the conditions under which emotional expressions fulfil their communicative and regulative roles in social life. This established *why* the concept of emotional privacy matters for assessing new and emerging technologies such as ERTs, yet it made no normative proposition on how those technologies ought to be evaluated nor what type of wrongs, if any, follow from a defeat of expressive ability. This chapter takes up that normative task. I defend emotional privacy as a distinct normative concern that reveals specific structural harms overlooked by existing normative (informational) privacy frameworks.

As argued in the previous chapters, ordinary mindreading and partial emotional legibility are not only inevitable but socially necessary. The aim of this chapter is thus twofold: first, to specify what a moral right to emotional privacy should protect in terms of preserving individuals' expressive ability, so that ordinary, acceptable losses of emotional privacy (actual losses) can be distinguished from wrongful ones (standing losses); and second, to demonstrate its moral relevance by applying it to concrete technomoral scenarios and to established ethical evaluation frameworks.

The chapter proceeds in four steps. **Section 6.2** articulates a normative account of emotional privacy and specifies the conditions under which emotional privacy would be violated, namely the defeat of expressive ability. **Section 6.3** applies this account to three technomoral scenarios to illustrate the types of interests at stake and to clarify the scope of the right across three levels of interaction (individual, interpersonal, societal). **Section 6.4** shows that current dominant normative variants of informational privacy fail to fully capture these interests, thereby demonstrating that emotional privacy fills a genuine normative gap. **Section 6.5** then examines

¹⁵³ Parts of this Chapter are published in Prigent (Forthcoming a).

how the reliance on emotional privacy may recalibrate ethical evaluation frameworks, including Contextual Integrity and ALTAI-style proportionality and subsidiarity assessments.¹⁵⁴

The core claim defended in this chapter is that violations of emotional privacy give rise to a distinctive kind of harms that existing normative accounts of informational privacy do not fully grasp. Fully functional ERTs risk inducing this kind of harms by defeating people's expressive ability thereby collapsing the functions currently served by EEs (especially intentional EEs) in social life. The normative account of emotional privacy proposed in this chapter therefore provides the normative grounding needed to fully grasp this technological reconfiguration of social life.

6.2. Formulating a normative account of emotional privacy

Several privacy scholars, including Warren and Brandeis (1890), Westin (1967), Allen (1988), Taslitz (2002), and more recently, Gremsl and Hödl (2022), and Menges and Weber-Guskar (2025), have considered or defended the idea that our emotions warrant a special kind of protection, sometimes framed as a (moral) right to the privacy of emotions. My argument departs from these approaches by shifting the focus from emotions as inner or mental states to emotional life as a socially embedded phenomenon. I thus treat emotional privacy as a distinct normative dimension of social life.

Foregrounding emotional privacy as a distinct normative dimension of social life immediately raises a familiar normative challenge: what should count as an appropriate expectation of privacy for emotional expressions, given that we both involuntarily leak and deliberately share them across private and public contexts? Not every loss of privacy is normatively problematic, and not every loss should be conceptualised in terms of privacy violations. As argued in the previous chapter, ordinary mindreading and (partial) emotional legibility are not only inevitable but socially necessary because they are part of what makes (intentional) emotional expressions capable of fulfilling their communicative and regulative roles.

Emotional privacy, as I understand it, therefore cannot require individuals to retain full control over the capacity to successfully modulate when, how, and to whom one's EEs will be legible. Nor should it be understood as shielding emotional life from all forms of access. Given ordinary

¹⁵⁴ ALTAI stands for the Assessment List for Trustworthy Artificial Intelligence.

affective leaks and the inherently social nature of emotional life, such requirement would be unrealistic. Instead, the normative account of emotional privacy should only track whether individuals retain expressive discretion over the legibility of their affects within affective communication channels. Hence, in this chapter, I will only focus on the normative stakes associated with standing emotional privacy.

Expressive discretion refers to the ability one has over the legibility of their emotional expressions (notably through the use of privacy mechanisms (see Chapter 5)). Thus, expressive discretion can only exist given the presence of a structural background of opacity against which EEs can be negotiated, contested, or strategically shaped. It is this background of managed opacity that enables individuals to (i) present the self they see fit for a given audience, (ii) sustain role plurality and adhere to role-bound norms of expressiveness, and (iii) contribute to social regulation and cohesion. These interests form a downstream normative trajectory: when expressive discretion is undermined, self-presentation is first compromised, itself destabilising role plurality and adherence to role-bound norms of expressiveness, and the regulative and cohesive functions of emotional expressions begin to thin within and outside affective communication channels.

On this view, emotional privacy *obtains* when individuals retain the expressive ability to control their affective information. Emotional privacy is *lost* when this expressive ability is defeated. The key question, then, is how to distinguish ordinary, acceptable losses of emotional privacy – i.e. leaks, misfires, unusually acute human perception – from wrongful losses, especially in technologically-mediated contexts. To mark this distinction, I rely on the standing definition of emotional privacy. Emotional privacy is *defeated* when a person cannot meaningfully use their EEs and privacy mechanisms to manage the legibility of their affective life, such that emotional expressions cannot fulfil their social and pragmatic functions in that context. Defeat thus concerns the loss of expressive ability in a given configuration of the affective environment, not merely discrete instances of (unwarranted) access or momentary loss of control.

This moral right to emotional privacy can be stated as follows:

Moral right to emotional privacy: A's moral right to emotional privacy is violated
iff A's expressive ability over affective information p is defeated.

The basis of a moral right to emotional privacy thus lies in the value of protecting one's expressive ability from defeat. I will come back to what this defeat entails, but first, some clarifications on the kind of moral right in question. Emotional privacy should be understood as a distinct moral right within the normative informational privacy cluster. Consequently, I follow most of the normative accounts in informational privacy and assume here that it is widely accepted that a moral right is justified insofar as it protects certain interests of X, whose realisation contributes to X's well-being. In the case of the moral right to emotional privacy, it protects the interests of (i) self-presentation, (ii) role plurality and adherence to role-bound norms of expressiveness, and (iii) social regulation and cohesion. Within the affective communication channels, these interests cannot be secured without the individual retaining her expressive ability, hence why expressive ability is the object of protection in the case of emotional privacy. The moral right to emotional privacy is a *pro tanto*, interest-based moral right.¹⁵⁵ Moreover, since the interests follow a downstream trajectory, securing (i) is a precondition for the realisation of (ii)-(iii). The right is thus individually *held* yet it secures interests that are also societal interests as self-presentation, role plurality and norm conformity, and social regulation and cohesion are *constitutive* of both individual and societal well-being.

6.2.1. Human vs technology-based emotion recognition

While Chapter 5 demonstrates that actual emotional privacy is often lost through affective leaks and (successful) ordinary mindreading, it would be counterintuitive to argue that such losses constitute a violation of privacy as they are part of social life. Hence, not all losses of emotional privacy should be morally conceptualised in terms of privacy rights, and a robust account of privacy should meet this difficult challenge.

The distinction between actual and standing forms of emotional privacy helps to distinguish ordinary losses of emotional privacy from wrongful ones by excluding affective leakages, misfires, and (unusually) acute human perception from counting as violations of a person's

¹⁵⁵ Justifying the moral right to privacy following the argument that it protects certain interests of X, whose realisation contributes to X's well-being, subscribes to an interest-based approach to individuals' right, as opposed to a will-based (choice-based) approach to individuals' right. The latter instead defends the idea that a right is essentially an opportunity for an individual (or group) to make choices. Instead of tying rights to various aspects of the right-holder's well-being, the Will Theory of rights holds that each right gives its holder a certain measure of control over their situation. To say that someone has a right is to say that they are empowered to decide whether and how another person's duty towards them is fulfilled (see Kramer, Simmonds, & Steiner (1998) for a debate on the two approaches. See also Raz (1994, chap. 2-3) for a defense of the interest-based view and Hart (1955) for a defense of the will-based view).

moral right to emotional privacy, while including recognition by technology such as fully functional ERTs. This allows me to give a clear answer to the question of what kind of normative privacy expectations should be appropriate regarding the case of emotional privacy.

This same difficulty animates long-standing debates about surveillance. While surveillance technologies are often assessed in terms of privacy, several scholars have expressed dissatisfaction with privacy-based objections to new and allegedly intrusive technologies (Ryberg 2007, 2017; Macnish 2018). Jesper Ryberg (2007, 2017) provides one of the most influential challenges to privacy-based objections in this domain. Discussing CCTV in public spaces, he argues that privacy theorists have failed to show why CCTV is morally problematic as a privacy violation. If there is a morally relevant difference between an ordinary person observing a street from their window and a CCTV camera recording it, he claims, “it must lie inherently in the way each of the two operate” (2007: 131); yet, according to Ryberg, no such difference obtains. He later deploys a similar rationale against advocates of neurorights (2017), arguing that they have not shown why neurotechnological mindreading violates mental privacy if ordinary mindreading does not. Since much of social life depends on mutual access to affective information, and since EEs routinely reveal sensitive information (e.g. through blushing, sighing, or frowning), it might seem that technological emotion recognition is simply an extension of human recognition, and thus normatively continuous with it.

Two notable responses by Annabelle Lever (2008) and Benjamin Goold (2008) already question Ryberg’s analogy by pointing to power asymmetries and institutional context. Lever situates privacy within a broader democratic framework, arguing that state-operated CCTV differs categorically from informal observation between citizens mainly because of the coercive power of the state and the requirements of democratic justification (Lever 2008: 38). Goold similarly emphasises the institutional nature of state surveillance and its capacity for monitoring and data retention (Goold 2008: 45).^{156, 157} Yet Ryberg (2008, 2017) remains unconvinced, claiming that these responses do not demonstrate how the use of CCTV, or neurotechnological mindreading, wrongs individuals’ privacy.¹⁵⁸

¹⁵⁶ Goold’s argument also includes the possibility that the old lady is violating people’s privacy.

¹⁵⁷ Interestingly, both critics point to the risks that follow from arguments that depoliticise new technologies that are intrinsically political, such as state-operated CCTV cameras.

¹⁵⁸ Although it sometimes seems like Ryberg would want privacy scholars to provide a dichotomous privacy framework where technology-based practices are *always* wrong and human-based practices are *always* legitimate, the view I am defending is more nuanced, and, as the reader shall see, I do not exclude the possibility that technology-based recognition practices may, in some contexts, be warranted.

Below I anticipate Ryberg's challenge as a main objection for the case that ERTs may not be a threat to individuals' privacy, in the same way as CCTV cameras and neurotechnologies are described as not being a threat either. I rebut his challenge by taking up his suggestion that the relevant difference lies in the intrinsic differences between human- and technology-based emotion recognition. I show how these differences are inherent to the way the two operate and how in the case of ERTs, they become normatively salient precisely through the lens of the moral right to emotional privacy.

In Chapter 4, I argued that human and technological recognition rely on different *epistemic routes*. Human recognition typically occurs *in situ*, on a pro tanto basis, within shared contexts of meaning and expectation. It is inherently defeasible because it depends on background knowledge, social conventions, and mutual adjustment. Human perceivers can misread, but they can also revise their interpretations. Uptake is therefore both negotiated and constrained (the latter being the case notably through norms of expressiveness and affective leaks). Recognition here thus belongs to a communicative regime, in which EEs typically function as contributions to an ongoing interaction, serving the emoter, the perceiver, and society as a whole. ERT-based recognition operates differently. In Chapter 4, I also argued that, in contrast with human-based recognition, ERT-based recognition is one not of communication but of diagnostic, as it uses data points used to infer, classify, and predict inner affective states.

Innovation theory helps to further capture the fundamental differences between human-based and ERT-based recognition. In innovation theory, a common distinction is drawn between what is called incremental innovation and radical innovation. Incremental innovation involves refinements that improve performance while leaving the core concept underlying the technology intact. For instance, better lenses or higher resolution in smartphone cameras sharpen images without altering their basic purpose. Radical innovation, by contrast, overturns the core concept of an entrenched technology or practice, as with the first photographic camera replacing drawing and painting with chemically captured light, or the airplane introducing a new mode of transportation (Christensen 1997). With radical innovation, the purpose of a given technology is reshaped to satisfy new ends.

What Ryberg argues is that CCTV cameras, or even neurotechnologies (and I presume he would say the same for ERTs) should be seen as incremental improvements vis-à-vis basic human capacities and practices as they merely enhance what humans already do, but they do not

overturn the practice; surveillance is still surveillance, and mindreading is still mindreading (granted, in a much more efficient way), in a similar way as a smartphone camera is still a smartphone camera, regardless of whether it has 8 or 16 megapixels. Since it is agreed upon that human ‘surveillance’¹⁵⁹ or mindreading pose no privacy issues, to be considered a new privacy threat, privacy scholars would need to show that these technologies are something at least equivalent to radical innovations, meaning that they fundamentally depart from human practices. The problem is that, as they currently stand, they embody incremental innovations.

However, a long-standing lesson from technological history – whether in industrial automation, biotechnologies, or information systems – is that incremental innovations can also, through small, cumulative modifications, transform the very nature of a system once they pass a critical threshold (Henderson & Clark 1990; Christensen 1997; Garcia & Calantone 2002; Dosi 1982). This is what Henderson and Clark (1990) call “architectural innovation” and what more recent work describes as the emergence of new technological paradigms from the cumulation of incremental changes (Pedota et al. 2021; Berggren 2019; Hacklin, Raurich, & Marx 2004). Once such a critical threshold is passed, what initially seemed like mere optimisations reorganise the system’s functional architecture and its purpose. The shift from mobile phones to smartphones exemplifies this dynamic. When mobile phones replaced landlines, they retained the core practice of voice communication, following a traditional path of incremental changes. When smartphones replaced mobile phones, however, they redefined the device as a general-purpose, networked computing platform in which voice calls are no longer central to the core practice. While it is sometimes difficult to pinpoint the exact moment at which this conceptual discontinuity occurs in the evolution of a given technology, the idea is that the accumulative improvements ultimately produce a different technological kind, with new roles, expectations, and practices.¹⁶⁰

Atoosa Kasirzadeh (2025) identifies a parallel in AI ethics, differentiating between accumulative risks and decisive risks. Accumulative risks are the equivalent of incremental changes reaching a threshold, where the core concepts are overturned, and decisive risks are the equivalent of radical changes (i.e. general AI). Hence, she notes that alongside the risks brought about by decisive breakthroughs in AI are also accumulative changes that slowly and pervasively alter the

¹⁵⁹ As Goold points out (2008), it seems that some cases, such as a peeping Tom or a malicious stalker, should be considered as privacy-threatening. But this lies beyond the argument I am making vis-à-vis technology-based vs human-based recognition.

¹⁶⁰ This threshold is also the one on which Nissenbaum relies for determining *prima facie* privacy violations. (see Nissenbaum 2010).

world. At first, similarly to incremental changes, such accumulative changes seem continuous with what preceded them. Yet, at a certain point, the accumulation of refinements produces a qualitative shift – a reorganisation of the system’s or practice’s core logic, reaching a point where the underlying practice is ultimately transformed.

The key element in what innovation theory calls “architectural innovation” or “new technological paradigms”, and what Kasirzadeh identifies as “accumulative risks” is that if we accept the central idea, Ryberg’s methodology of presenting technologies as simple improvements of human capacities leaves his account blind to that specific kind of pervasive incremental changes.¹⁶¹ If similar changes in emotion recognition do overturn the core concepts grounding the human practice, then we have grounds for assessing them as new (technological) privacy threats in the technology-based practice.

Concerning privacy and privacy rights, I suggest that what others have attempted to point out with the use of ubiquitous CCTV cameras and anticipate with neurotechnologies is similar to what I am doing with fully functional ERTs. I am arguing that fully functional ERTs constitute a new kind of privacy threat because they have passed a decisive turning point, becoming an inherently different form of emotion recognition and posing a distinct normative kind of privacy risk from human emotion recognition. As both innovation theorists and AI researchers note, however, a central difficulty with such transformations is that, in most cases, it is almost impossible to identify the precise moment at which the technology crosses the relevant threshold and the core practice is recognised as having been overturned. The example of CCTV cameras shows clearly this challenge. It is almost impossible to pinpoint when, exactly, the practice of surveillance through CCTV cameras has stopped being simply what could be considered an extension of human surveillance and reached the point where it created a new normative paradigm of ubiquitous surveillance, but it seems fair to say that we have, somewhere down the line, crossed that threshold. This change of paradigm was pervasive and pernicious, leaving most people, including policymakers, under high uncertainty regarding the many different new ethical issues it created (including privacy issues). In the case of neurotechnologies and fully functional ERTs, what privacy scholars are attempting is to anticipate these ethical issues before they are actualised.

¹⁶¹ In innovation theory, however, architectural change is not always seen as negative, and is often even seen as a positive kind of breakthrough.

Hence, I assume that fully functional ERTs have, by definition, already reached a tipping point where the technology overturns the core concept of the human practice of emotion recognition. Under fully functional ERTs, emotional legibility is no longer situated, defeasible, and negotiated, but systematic, transparent, and unilateral. At that point, the technology ceases to be a better version of human recognition and instead becomes a different practice altogether. The underlying concept of emotional legibility changes, and with it the normative space in which emotional privacy operates.

From the perspective of emotional privacy then, the difference between human- and technology-based recognition is not merely a difference in degree but in kind. Human recognition, constrained by perceptual limits and social norms, leaves room for opacity, ambiguity, and strategic performance. It presupposes and preserves expressive discretion. Fully functional ERTs, by design, threaten to introduce a condition of systematic exposure, in which inner affective states are made legible regardless of individuals' use of expressive discretion. They enact what I refer to as the defeat of expressive discretion: the disabling of a person's ability to meaningfully manage the legibility of their EEs.¹⁶² Importantly, by 'design', I mean that this is not because an ERT bypasses expressive discretion on an x-number of occasions that at some point the ERT's legibility becomes problematic. Rather, the background conditions of affective opacity begin to erode from the first instance in which technological recognition is made reliably and continuously available. The wrong, in such cases, does not lie in a single instance of exposure, or in the accumulation of such instances of exposure, but *in the structural disruption of the social and pragmatic functions of emotional expressions*.

Once we understand the conceptual and structural differences between human and technological recognition, it becomes clear that fully functional ERTs cannot be treated as a mere extension of ordinary human emotion recognition, as Ryberg's earlier critique of CCTV cameras and neurotechnologies presumably would assume, because unlike human recognition, which is an essential part of social life, ERTs reconfigure the practice of emotional legibility itself. Ordinary human recognition and technologically mediated recognition belong to different normative kinds, and only the latter threatens the affective infrastructure on which self-presentation, role plurality, and social regulation currently rely. This is what the normative account of emotional privacy is conceptually designed to shed light on. Emotional privacy,

¹⁶² As mentioned earlier, this does not entail the successful control of EEs, but only the ability to exercise expressive discretion.

understood as a normative dimension concerned with the preservation of expressive discretion and managed affective opacity, is therefore directly threatened by fully functional ERTs because their use defeat individuals' expressive discretion.

Now that I have explained the inherent difference between human and technology-based emotion recognition, I use three technomoral scenarios to analyse the extent to which the use of fully functional ERTs may threaten the interests at stake in emotional privacy.

6.3. Illustrating emotional privacy losses through technomoral scenarios

EEs are essential elements of most of our interactive practices. Nonetheless, they are informative only conditional to the perceiver's successful interpretation of them, which is itself conditional to the perceiver's access to relevant epistemic inputs (e.g. background knowledge, context, prevailing norms).¹⁶³ Hence, ordinarily, social life proceeds under a degree of affective opacity. Emotional states and social motives are not fully legible by default and an individual's willingness to share raises or lowers the probability of their own affective legibility. The privacy mechanisms that one deploys to this end constrain the inferences and predictions others can draw from one's EEs and enable individuals to manage the strategic ambiguity necessary to navigate social life. Humans heavily rely on this expressive ability to maintain self-presentation, manage role plurality, and collectively help to regulate social life, ensuring social cohesion.

This affective infrastructure thus sustains social life across multiple levels of interaction: individual, interpersonal, and societal (see Chapter 5). Fully functional ERTs threaten to dismantle this affective infrastructure by collapsing the fragile equilibrium between social norms of expressiveness, what we choose to display, what we withhold, and how others interpret our EEs, given also possible affective leaks. By automating, accelerating, and amplifying the process of affective inference, they drastically reduce the opacity that normally obtains in affective communication channels, defeating our expressive ability. When privacy mechanisms, and, correlatively, interpretive labour, and perceptual limits no longer operate as constraints, the emotional lives of others become, in principle, continuously and effortlessly legible. What is threatened is not only "emotional data", or the sanctity of an "inner" life, but the *functional architecture of emotional life as a social practice*. Through the anticipation of fully functional ERTs, it is possible to expose the moral cost of this collapse.

¹⁶³ See Chapter 4.

In this section, I first lay out three technomoral scenarios in which a (standing) emotional privacy loss has happened due to the defeat of expressive ability by an ERT. I use these three scenarios to track the harms back, exposing the threat to our interests in self-presentation, role plurality and conformity to role-bound norms of expressiveness, and the social regulation and cohesion that stabilise collective life.

To do so and to be in a position to focus on emotional privacy violations, each scenario treats an idealised, fully functional ERT.¹⁶⁴ As a reminder, this means that the ERT is perfectly accurate, universally applicable across cultures and individuals, and capable of real-time interpretation of emotional expressions. It does not suffer from the technical limitations, bias, or contextual insensitivity that preoccupy current critics. Its inferences are flawless, context-aware, and seamlessly integrated into various social domains such as workplaces, education, healthcare, public spaces, interpersonal communication, and law enforcement.

At first glance, such a technology might appear ethically unproblematic. It produces no false positives or negatives, avoids misinterpretation, and could be deployed for seemingly beneficial purposes such as enhancing empathy, improving mental health support, or even flagging violent individuals before they commit any aggression. Yet it is precisely in this perfected form that a fully functional ERT sheds light on the moral significance of emotional privacy. The moral wrong lies not in what ERT misrepresents (or might misrepresent), but in what it renders impossible, namely, the maintenance of the functional architecture of social life sustained by emotional expressiveness.

As is often the case with social matters, the levels of interactions discussed so far, namely individual, interpersonal, and societal, are not completely impermeable from one another, and sometimes overlap. I aim thus not at attempting to show how ERTs may disrupt specific social functions of EEs while leaving the others untouched, but rather how specific functions of EEs, mainly from intentional EEs, at least at these three levels, are severely disrupted by the defeat of people's privacy mechanisms, leading to harms that affect both individuals' and society's interests.

The three scenarios follow the three levels of interaction presented in Chapter 5. Moreover, they all have a *distinct social setting* in which the ERTs are “owned” or at least in the possession

¹⁶⁴ See Chapter 2 for a detailed definition of fully functional ERTs.

of different types of individuals. Scenario I is designed to analyse issues related to ERTs that are “owned” or in the possession of the data subject. Scenario II is designed to grasp issues related to ERTs that are “owned” or in the possession of someone in a position of authority over the data subject. And scenario III is designed to grasp issues related to situations where ERTs are already a widespread and ubiquitous technology, possessed and used by everyone, similar to what phone cameras are nowadays. I first overview each technomoral scenario by explaining what would have happened in that particular situation if there had been no ERT involved. Then, I pinpoint where and in what the loss of emotional privacy consists, the harms that follow from this loss, and whether this loss appears to be wrongful or not, leaving the final verdict for the next section. While reading the scenarios, some might think that they beg a question on the role of consent. The technomoral scenarios are designed to clarify what is structurally at stake; please bear with me as I will keep the discussion around whether and to what extent consent can legitimise the use of ERTs in subsection 6.5.3.

6.3.1. Scenario I: Alfred’s self-monitoring

Alfred wears affect-sensing glasses and watch to better understand and manage his stress. At a small dinner with close friends, the system detects a sustained pattern of physiological and behavioural markers for sexual desire when his friend’s spouse speaks; on his phone, a push notification reads “arousal spike (87%)”. Alfred freezes. He also realises that two other guests have seen the notification.

Ordinarily, Alfred has the ability to, and exercises his ability to, deploy privacy mechanisms, ranging from direct concealment to selective disclosure. Even though his attempts are not always a success, he usually manages to meaningfully regulate the affective opacity in the interactions he sustains and, if at all, turns felt states or social motives into emotional expressions. Even though regular leaks are noticed, they rarely contradict his deliberate expressions in a way that would make the latter useless. The device changes this social structure. The protected space in which conflicted affects can be worked through, without becoming socially actionable, disappears.

In the technomoral scenario, Alfred loses actual emotional privacy over his sexual desire for his friend’s spouse.¹⁶⁵ By exposing his desire, the ERT dismantles the affective scaffolding that supported the dinner conversation between friends. The dinner conversation moves from negotiated meaning to device-mediated exposure.¹⁶⁶ In such a situation, the exposure of an

¹⁶⁵ As mentioned above, I will discuss the question of consent in §6.5.3.

¹⁶⁶ Note that, on the perceivers’ end, the non-acknowledgement ability is also diminished, as a formal exposure like this one makes it harder to pretend to “not notice”.

inappropriate sexual desire would likely prompt others to question Alfred, to seek explanations, or to reassess their view of him, even though he has done nothing wrongful. The privacy-preserving pathways of reflection, retrospection, and growth, that were ordinarily within Alfred's reach, are short-circuited. On the account of emotional privacy developed here, Alfred therefore loses not only actual but also standing emotional privacy. The loss is wrongful not because of the content of the desire itself, but because it results from the defeat of his expressive ability.

6.3.2. Scenario II: Marie's hospital policy

Marie is a senior surgeon. She is also in the middle of a bitter custody dispute and has sought counselling to stay on top of things. After another rough morning, she walks onto the surgery floor, happy to finally be able to clear her head at work by focusing on the procedures she has to perform. She is handed a small electronic pad that links to the hospital's affect monitors to check fitness to operate. Before she reaches the first patient file, the screen turns on and reads "Not emotionally ready to proceed." Her authority to sign pre-op orders is temporarily locked and a prompt invites her to a short well-being debrief. Nurses and doctors look up, surprised. Marie is known across the department for a cool-headed temperament and almost impeccable composure. The coordinator follows protocol and books the debrief. Marie had planned to keep the morning's dispute out of the hospital as she does not like to mix personal and professional life, but she has no other choice now, especially since the affect monitor was installed following due process.

Ordinarily, in the surgery unit, Marie sustains her professional and personal boundaries through three privacy mechanisms. Through direct concealment, she attempts to keep her inner affective life tied to personal matters away from hospital chatter. Through indirect concealment, she maintains the convention-fit calm expected of a senior surgeon, preserving both professional relationships with her colleagues and respect for her authority.

The hospital's ERT collapses this arrangement by stitching together her current two main social roles— surgeon and soon-to-be divorcee – and converting what she meant to keep opaque into a matter of professional fittingness. Direct concealment is bypassed because her effort to maintain neutral behaviours is publicly contradicted by the ERT's alert. Indirect concealment is dissolved because the system announces the very tension that her convention-fit behaviour was holding in place. Selective disclosure is pre-empted because the protocol forces her into a mandatory debrief with the coordinator. Here, too, standing emotional privacy is lost. Marie's capacity to maintain a composed professional demeanour despite experiencing a complex mix of emotions, is defeated. What is undermined is not merely the concealment of a particular feeling, but her capacity to regulate how her affective state becomes socially legible within the practice.

The moral harm, however, extends beyond the individual level. It compromises her ability to present the self she intends to be in this context and damages her ability to manage her distinct social roles and to conform, as she intended to, to the social norms of conduct of a senior surgeon. As a consequence, colleagues recalibrate their own portrait of her professional self to the device output rather than to her seniority, record, and typical expressions. Her calm is re-read as inauthentic, and her practical authority is reduced. In short, the ERT defeats the affective scaffolding that ordinarily regulates the workplace by lowering informational opacity and flagging her inner affective states as above the acceptable threshold.

6.3.3. Scenario III: The minister's public apology

A teenager is killed in a late-night knife attack. The Interior Minister schedules the typical process in such case: a closed briefing with the victim's family, a short televised address to acknowledge the harm and apologise, then a written plan for policing and prevention released the next morning. The minister begins her public apology, but this time, people use their app-based ERTs on their phone to fact-check her emotional expressions. The minister's emotional score reads "remorse (10%)". Commentators pivot to the scores rather than the reasons she is giving. The apology has not done its regulatory work as the minister's verbal apology is overridden by her emotional score, drastically altering the population's interpretation of the whole scene.

Ordinarily, the minister's address is designed to perform two intertwined social functions. It attempts to regulate society's frustration by acknowledging the breach, explaining the next steps, and anchoring a proportionate path forward. It also attempts to sustain cohesion by offering a shared moment to recognise the loss and offer reassurance. These social functions rely at least partly on selective disclosure and concealment. The televised apology is curated. Reasons are given in an order that maintains role distance and tempers any potential community outrage, before sanctions are considered.

This time, the phone-based ERT prevent the success of the typical process and collapses the minister's attempt at regulating collective outrage and ensuring the continuation of social cohesion. By rendering the minister's affects continuously legible to everyone watching in real-time, the system defeats the minister's ability to rely on privacy mechanisms to perform her emotional expressions for her public. The score appears as part of the message and the typical opacity that lets the apology play its regulative role disappears. Hence, there is loss of emotional privacy. Selective disclosure is pre-empted and the apologetic expression, that ordinarily support the verbal apology, ceases to function as a regulative practice that guides collective response. The main problem is that the new state of legibility defeats the privacy mechanisms that make

public emotional expressions reliable sources of regulation and cohesion. That shift marks a loss of regulatory traction and a thinning of social cohesion, exposing the societal-level harms caused by a defeat of emotional privacy. In this case, there seems to be a violation of the moral right to emotional privacy.

Importantly, these scenarios are not intended to suggest that emotional privacy ought to shield individuals from legibility altogether. The protection of expressive discretion does not (and is not meant to) secure an absolute state of opacity. Emotional privacy is necessarily partial and remains porous to affective leakages, misfires, and unsolicited inferences that shape how one is perceived. Within these limits, however, emotional privacy preserves a zone of strategic ambiguity in which emotional expressions can perform their social functions. It is this zone that warrants protection against technologically mediated disruption. The technomoral scenarios aim to show how fully functional ERTs can undermine this zone in distinct ways, thereby jeopardising the work of intentional EEs within social practices.

6.4. Moral right to emotional privacy protects a distinct kind of interests

The concept of emotional privacy uncovers a set of distinct interests that are not adequately captured by existing (moral) rights of informational privacy. The moral right to emotional privacy thus aims to protect these new interests whose loss risks severely damaging the current structure of social life. These interests form a downstream normative trajectory. At the upstream level is expressive discretion, the ability to manage the legibility of one's emotional expressions within affective communication channels. When expressive discretion is defeated, individuals lose the ability to present the self they judge appropriate. This upstream loss then cascades. First, it undermines self-presentation, then, it collapses role plurality and the capacity to meet the expressive expectations tied to specific social roles, and finally, it erodes the regulative and cohesive functions that emotional expressions ordinarily perform in sustaining social coordination and collective life.

The three technomoral scenarios presented above each focus on a particular stage of this progression. Alfred's scenario shows how technologically induced legibility defeated the capacity for self-presentation. Marie's scenario illustrates how this legibility collapsed role plurality and conformity with role-bound norms of expressiveness by defeating her efforts to keep personal matters away from her work. Finally, the systematic exposure of the Minister's affective states disrupts the regulative and cohesive functions of emotional expression in public

life. In each case, the defeat of expressive discretion triggers individual, interpersonal, and societal effects.

If these interests do indeed form a downstream trajectory, then a moral right designed to protect expressive discretion against defeat must also be understood as protecting the downstream interests embedded in affective communication channels. In this section, I therefore examine how the familiar variants of the moral right to informational privacy (Control-based accounts, Access-based accounts, and Contextual Integrity) fare in protecting affective information carried by emotional expressions. I argue that while affective information naturally belongs within the broader informational privacy cluster, the definition proposed by the mainstream normative accounts of informational privacy fails to secure the specific interests at stake. The conclusion is that none of them offers adequate protection to this particular set of interests, hence emotional privacy must be recognised as a moral right within the informational privacy cluster.

6.4.1. Variants of Control-based accounts

I now turn to the moral right to privacy as understood from the Control-based accounts already introduced in Chapter 5. The point I am trying to make here is *not* that these accounts *cannot* capture any privacy wrongs caused by ERTs, only that they are ill-equipped to grasp the wrongs that follow from the defeat of expressive discretion, therefore overlooking something of moral value. Hence emotional privacy is a complementary moral right.

The normative Control-based approaches (Classic Control, Negative Control, Source Control), ground the *moral right* to informational privacy in a person's authority vis-à-vis informational flows (Westin 1967; Mainz & Uhrenfeldt 2021; Menges 2021). Although the approaches disagree on what the authority is over and how it is exercised, they usually agree on the core idea that privacy wrongs stem from *failures of rightful control*.

On the Classic *normative* Control account, person A's privacy has been violated iff person A has involuntarily lost control over access to personal information p about person A.¹⁶⁷ The main problem with this formulation is that it includes ordinary affective leaks as “involuntary loss of control”, which renders any (semi-)involuntary EEs (e.g. natural EEs, semi-natural EEs, and

¹⁶⁷ I took this normative definition of the ‘classic’ control-based approach from Mainz and Uhrenfeldt (2021: 289), specifically their CA1 definition, but I have removed the “unwanted” clause to escape issues over the subjective evaluation of what is considered “unwanted access”.

affective leaks) a wrongful loss of an individual's informational privacy. This is highly problematic for obvious reasons, namely that people would have their informational privacy constantly violated, meaning that there would also be no difference between ordinary losses and wrongful losses.

On the last *normative* form of the negative control account, as proposed by Mainz and Uhrenfeldt (2021), which they characterise as the strongest one, person A's right to privacy is violated iff person A has involuntarily lost negative control over the access to (personal) information p about person A, due to action(s) of person B, of which B is responsible (2021: 298). Recall from Chapter 5 that by "negative control" the authors mean "the ability to prevent".

This formulation captures clear cases in which a person is subjected to access they cannot block, for example coercive interrogation or unauthorised surveillance (Mainz & Uhrenfeldt 2021). However, it does not translate well to cases involving emotional expressions. The central difficulty is that the account assumes a discrete and attributable act of access by a specific B whose responsible action causes A's involuntary loss of control. Emotional expressions do not function in this way in ordinary social life. People routinely perceive and interpret the emotional expressions of others, and while this does result in a loss of negative control, B's behaviour consists merely of ordinary looking, listening, or attending. It is a normal feature of interpersonal engagement rather than an attempt to invade someone's privacy. Without any technological intervention, Alfred's friends may notice his attraction at dinner, Marie's colleagues may see that she is tense, and attendees present at the Minister's apology may form judgements about her sincerity. In each of these situations, someone accesses information about another person and the loss of control on the part of the person observed is involuntary, yet none of these instances are plausibly classified as wrongful privacy violations. The negative-control account therefore risks treating routine and socially appropriate forms of affective uptake as violations simply because they involve an involuntary loss of control caused by another's ordinary actions.

The account also encounters difficulties when applied to technologically mediated emotional legibility. In environments where fully functional ERTs are continuously operating, as in Marie's hospital scenario, emotional legibility results from the design of the environment rather than from any single agent's deliberate action. If we insist on identifying responsible agents in such cases, responsibility becomes so widely dispersed across engineers, managers, system operators, and others that the notion loses any meaningful grip. The alternative is to conclude that no

violation occurs because no single act of access can be attributed to a specific B. Yet the harm in Marie's case clearly does not stem from any specific observer B actively acquiring more information than she intends to reveal. It stems from the technological configuration of the environment itself. Once the system is installed, her expressive discretion is undermined regardless of what any individual colleague or supervisor does. A framework that recognises only discrete and agent-driven acts as privacy-relevant cannot account for this structural defeat of expressive discretion.

The result is that emotional expressions generate two types of losses of negative control. The first is benign and unavoidable in ordinary interaction. The second is wrongful and arises when emotional legibility is engineered into the environment through technological means. On Uhrenfeldt and Mainz's account, a person's privacy is violated when they involuntarily lose negative control over access to their personal information due to actions attributable to another agent (Uhrenfeldt & Mainz 2021). Framed in this way, the account appears both overinclusive and insufficiently precise regarding the scope of wrongful privacy violations. It risks labelling ordinary interpersonal perception as a privacy violation, while also failing to register violations in technologically-mediated contexts where emotional legibility is structurally imposed. For these reasons, the negative-control account cannot capture the practice-disrupting defeat of expressive discretion brought about by fully functional ERTs and cannot distinguish such cases from the ordinary losses that structure everyday life. It therefore cannot absorb the normative work that the moral right to emotional privacy must perform.

Finally, Menges' source-control account is, as I understand it, developed into a normative account called the Deep Self View (DSV). On the descriptive account of source control, Menges holds that privacy obtains when person A is the *right kind* of source for information p's flow (if information P flows at all). Building on this descriptive foundation, Menges formulates DSV as a normative refinement of the control tradition by grounding the moral right to privacy in the protection of a person's deep self as the legitimate moral source of information about them. On the normative account, he thus holds that:

A has a right to privacy regarding personal information p just in case others have the pro tanto obligation directed at A to not make p flow unless (i) this flow is an expression of and (ii) does not conflict with A's deep self. As long as A has neither waived nor forfeited the right to privacy, B infringes A's right to privacy regarding p just in case B makes p flow even though it is not an expression of or conflicts with A's deep self (Menges 2024: 7).

The “deep self” can be summarised as a person’s “set of [relatively] stable intrinsic desires” (Menges 2024: 9).¹⁶⁸ Examples of such intrinsic desires include wanting one’s children to flourish or wanting to leave future generations a world worth living. According to the DSV, information p should fall under A’s moral right to informational privacy when, in a given context, it is instrumentally, representatively, or symbolically valuable that A’s deep self be the source of any associated informational flow. The moral right therefore imposes on others a *pro tanto* obligation not to initiate informational flows about A unless doing so expresses, and does not conflict with, A’s deep self. The decisive normative question is not whether A voluntarily chooses the flow, but whether the flow properly originates from A’s genuine and relatively stable intrinsic commitments.

Although the view is internally coherent, certain tensions arise. First, Menges holds that “one does not need to be aware of one’s intrinsic desires” (Menges 2024: 9) but also requires that others have a *pro tanto obligation* not to make information about A flow unless it expresses (and does not conflict) with A’s deep self. While Menges’ account does not demand omniscience but only a reasonable-belief standard on person B, this does not entirely resolve the tension. Given this epistemic opacity, it is unclear how others can reliably fulfil the *pro tanto* obligation not to make information flow unless it expresses or aligns with A’s deep self. Even with a reasonable-belief standard and a default presumption against flow, if we accept that seemingly banal facts, such as whether one has two hands,¹⁶⁹ can possess instrumental, representative, or symbolic significance in particular contexts, the DSV places a considerable practical burden on others. They must determine, often immediately and without proper contextual insights, whether a given informational flow would appropriately express A’s deep self or instead conflict with it. This problem becomes even more acute in large-scale or multi-agent technological environments, for example those involving ERTs in classrooms, hospitals, or airports. In such settings, it is effectively impossible to assess, person by person, whether a specific informational flow would or would not originate in the individual’s deep self.

But the central difficulty for DSV in relation to emotional privacy is that it evaluates the legitimacy of the flow of information through the lens of authenticity and alignment with stable

¹⁶⁸ I will not go into the details of the notion of deep self here, but note that these stable desires constitute, in a sense, their sense of agency; “the agent’s ability to write their own life stories” (Menges 2024: 16). See also Menges (2024: 9-11) for a detailed explanation of the deep self.

¹⁶⁹ Let’s say that I know person A has two hands, and a couple days later our country declares mandatory military service for everyone who is apt to fight, which implies that the people need to have two hands, this suddenly becomes highly relevant for A’s deep self.

intrinsic desires, whereas emotional privacy evaluates the preservation of expressive discretion within affective communication channels. As a result, the two frameworks do not track the same kinds of harm.

The DSV view thus allows for a flow to align with one's deep self and yet undermine one's ability for expressive discretion. The DSV can judge an informational flow to be legitimate even when that flow undermines the individual's ability to manage the opacity and selectivity of their emotional expressions. Marie's case illustrates this problem. Suppose that part of Marie's deep self, as a committed surgeon, includes the intrinsic desire to safeguard patient well-being and to maintain high professional standards. If so, the informational flow produced by the hospital's ERT system can plausibly be interpreted as expressing or at least not conflicting with Marie's deep self. Under the DSV, this flow would not constitute a privacy infringement. Yet emotional privacy is clearly violated in Marie's case. Her expressive discretion is defeated by the system's continuous extraction of affective information, which prevents her from managing her professional self-presentation in context. The DSV therefore fails to protect her ability for expressive discretion because in this case it does not directly correlate with her deep self but instead concerns the functional and relational significance of emotional expressions in everyday interactions.

In other words, the DSV operates at the level of the individual's deep, relatively stable commitments, while emotional privacy operates at the level of social practices in which expressions serve regulative and coordinative functions. The DSV therefore overlooks something of independent moral importance, namely the preservation of expressive discretion in affective communication channels. Consequently, the harms that emotional privacy seeks to protect against are not adequately captured by the DSV.

6.4.2. Variants of Access-based accounts

The normative Access-based approaches (Classic Access, Restricted Access, Actual Access), ground the *moral right* to informational privacy in a person's information accessibility given some conditions (Gavison 1980; Allen 1988; Macnish 2018). Despite their internal differences, all normative access-based accounts of privacy share a structural commitment. What is protected, in each case, is some informational *p* about person *A*, and correlatively, what is condemned is *an act of access to that information under the wrong conditions*. Whether they focus on unjustified

intrusion (Gavison 1980), contextually inappropriate accessibility (Allen 1988), or unauthorised epistemic access (Macnish 2018), these accounts locate privacy's moral content in the act or condition of access across an informational boundary.

On the normative Classic Access-type accounts, following Gavison (1980) and Reiman (1976), person A's informational privacy is violated iff A is subjected to unwanted access to information p by some person B. The central idea is that privacy protects individuals against being subjected to access they do not want. Applied to paradigmatic cases, this account-type can explain the wrong of intrusive surveillance when A never had any meaningful control over the situation in the first place, for instance when B observes A through a hidden camera or wiretaps a telephone line. In these cases, it seems clear that A's privacy is violated because B exposes A to a sustained unwanted access.

However, once we turn to emotional expressions, the strengths of the Classic Access account become a source of difficulty. Applied to the technomoral scenarios, this account faces an immediate problem. Unwanted access to affective information is part of ordinary social life (and even important to it). If the classic access account treats all such unwanted access as a violation, it collapses benign and even valuable forms of emotional perception into the category of privacy harms.

In technology-free versions of the scenarios for instance, Alfred may not want his friends to notice his attraction, yet their uptake of his demeanour is clearly part of the normal dynamics of friendship and social attunement. Likewise, Marie's colleagues may notice that she is unusually tense, and in the Minister's case, members of the public will inevitably form views about her sincerity. In each situation, someone accesses affective information that the person would rather not share, and yet these instances of unwanted access do not seem to be wrongful privacy intrusions. They are ordinary losses of emotional privacy that are woven into the fabric of interpersonal and public life. Hence, the Classic Access account cannot respond to Ryberg's challenge because it does not have the tools to differentiate ordinary mindreading from ERT-based recognition. What emotional privacy highlights is that the difference between ordinary mindreading and ERT-mediated access does not lie only in the presence or absence of an unwanted act of access. It lies in the way ERTs structurally defeat expressive discretion by making affective states continuously and determinately legible, and by transforming the practices

within which emotional expressions operate. For this reason, the Classic Access account cannot absorb the moral role of emotional privacy.

On Restricted Access-type accounts, person A's informational privacy is violated iff A is subjected to unauthorised access to information p by some person B. In this tradition, inspired by Allen's (1988) and by Tavani and Moor's Restricted Access account (2001), the key question is not simply whether access is wanted, but whether it is normatively permitted in light of social and institutional rules (Allen 1988; Tavani and Moor 2001). What matters is not whether A wants access to occur, but whether B is authorised (institutionally, socially, or even morally) to acquire the relevant information. Privacy is therefore understood as a system of restricted access zones: certain agents may access certain information only within the limits defined by social roles, institutional norms, or moral duties. In this respect, the Restricted Access account has an important advantage over the classic access view. Whereas the classic view relies on unwantedness and thereby risks treating any unpreferred perceptual uptake as a privacy violation, the Restricted access view replaces subjective attitudes with objective permissions. This allows it to distinguish between unwanted but permissible access, such as a colleague noticing fatigue, and unwanted but impermissible access, such as an employer secretly reading private emails. It also explains why some forms of surveillance constitute privacy violations even when individuals have little to no control (against weaknesses of the Control-based accounts) and even when they might not object in a subjective sense. By grounding privacy in socially recognised boundaries rather than personal preference, the Restricted Access framework is better equipped to analyse privacy in institutional contexts where entitlement, not desire, determines who may know what about whom.

However, once applied to emotional expressions and fully functional ERTs, this advantage becomes insufficient. Consider Marie's case. In the technology-free version, some colleagues might notice that she seems unusually tense before surgery. Their uptake of this information is not only socially authorised but also expected, since mutual attunement is part of professional collaboration. On the Restricted Access account, no privacy violation occurs, and this seems correct. In the ERT-mediated version, the hospital installs a system that continuously analyses Marie's EEs and outputs a classification that she is not fit to work. However, since the system operates within proper institutional authorisation, the Restricted Access accounts register no wrongful loss of privacy, even though Marie's expressive discretion is defeated in a way that deeply affects her professional and personal roles. Emotional privacy, by contrast, identifies this

as a central harm. The mere fact that the hospital is authorised to monitor her EEs does not mitigate the architectural change in the conditions under which her emotional expressions function nor the defeat of her expressive ability.¹⁷⁰

A third variant within the access-based tradition strengthens the definition of a privacy violation by requiring not merely unwanted or restricted possibility of access but Actual (epistemic) Access (see Macnish 2018). On this view, person A's informational privacy is violated iff A is subjected to actual access to information p by some person B. This account narrows the scope of privacy violations, since mere opportunity for access or the mere increase in accessibility does not suffice. There must be an identifiable moment in which B comes to possess, apprehend, or register the information in question (Macnish 2018). The strength of this account lies in its ability to avoid treating only hypothetical or merely potential access as wrongful. If surveillance cameras are installed but no one watches the footage, or if a database is created but never queried, the mere possibility of knowing is not automatically treated as a privacy harm (Macnish 2018). This makes the Actual Access account stricter in its violation-threshold than both the classic and restricted variants, especially in situations where informational risk is high due to the automatic collection of information, but no knowledge has yet been formed because no agent has accessed it yet.¹⁷¹

Applied to emotional expressions, however, if actual access means *genuinely* acquiring information p about A, it seems that the Actual Access account gives us the opposite result than the one we are looking for intuitively. Because emotional expressions are inherently public-facing and partially involuntary forms of behaviour, and ordinary uptake of them is not only unproblematic but essential to social life, B acquires information p about A in ordinary settings, all the time, subjecting A to actual access over their information p. Let's call this one interpretation 1. If we do treat human-based inference as actual access, then ordinary perceivers in the technomoral scenarios often count as having actual access, and the view would then condemn human-based uptake over the practice-disrupting effects of ERTs. Yet it would be implausible to treat these episodes as privacy violations.

¹⁷⁰ But as I will discuss in the next section, this authorisation may be used to warrant the use of ERTs against the pro tanto reasons in favour of emotional privacy.

¹⁷¹ It has also been used in neurorights debates to distinguish between technologies that “actually access” inner mental content and ordinary interpersonal inference, suggesting that some forms of technologically mediated knowledge acquisition are categorically different from ordinary observation (Yuste et al. 2017; Lavazza 2018; Ienca 2021).

One might reply that Actual Access is best understood as a strong kind of epistemic penetration that ordinary perceivers lack. Let's call this one interpretation 2. On this interpretation, though, then both ERTs and humans would never "actually" access emotional states, because the Access view already discard technological uptake from "actual", human agent, uptake. This results in no privacy violations ever, which also seems implausible. On another interpretation, we might say that Actual Access is only happening when human agents actually access ERTs' output (interpretation 3). Let's consider this interpretation in the light of the technomoral scenarios. In Marie's hospital, the ERT system continuously analyses behavioural and physiological changes, storing affective classifications even if no supervisor personally looks at the data. Marie's expressive discretion is already defeated at this stage, because her emotional life is rendered structurally legible in a way she cannot meaningfully regulate. Yet no discrete moment of actual access by a human observer need occur for the harms to happen. When the "Not emotionally ready to proceed" notification popped up on the screen, people around Marie did not have access to the 'actual' state of her inner emotions¹⁷² and yet this was sufficient to defeat her expressive ability. This means that if we adopt the last interpretation of the Actual Access view, many privacy violations are overlooked and the harms resulting from those violations are left unflagged.

The Actual Access account therefore ties privacy violations to the instant at which information is obtained (no matter how we conceive the conditions that needs to be met for this obtention to realise), whereas emotional privacy concerns the degradation of the background conditions that make emotional expressions meaningful, interpretable, and socially functional. Ordinary mindreading in the technomoral scenarios involves actual access to some EEs (If we accept the first interpretation of the Actual Access view) but does not undermine these conditions. ERT-mediated environments defeat expressive discretion even if no individual ever accessed the information produced by the system. Hence, regardless of the interpretation we use for understanding the Actual Access view, it cannot systematically identify practice-level transformations in affective communication channels where the wrong (*viz.* the defeat of expressive ability) arises prior to, and independently of, any discrete moment of actual epistemic access. It either takes every, ordinary, human uptake as a privacy violation (interpretation 1), no human or technology-based uptake as privacy violation (interpretation 2), or only human uptake of a technology-based output (interpretation 3), where the latter results in overseeing privacy

¹⁷² For the sake of this argument, let's hypothesise that such thing could have been possible if they had taken a look at the fine-grained data provided by the system rather than just the final decision based on those results.

violations that arise despite this condition being met. What emotional privacy highlights as harmful is not the moment at which a person actually acquires information about another person. It is the architectural transformation of the communicative environment that makes expressive discretion impossible, independent of any actual (epistemic) uptake of a human on a technology-based output.

6.4.3. Contextual Integrity

Helen Nissenbaum's framework of Contextual Integrity (CI) (2010, 2019) offers one of the most influential normative approaches to privacy. Unlike control- or access-based theories, CI grounds privacy in the *appropriateness* of a given information flow relative to a given context. Privacy is preserved when information flows conform to the prevailing entrenched informational norms of a context, specified by roles (sender/data subject/recipient), attributes (information types), and transmission principles (e.g. consent, confidentiality, reciprocity).

A deviation from those contextual norms is first flagged as a *prima facie* violation of privacy and subsequently confirmed (or refuted) as *genuine* once a moral assessment has deemed that the new flow does *not* uphold the context values, ends, and purposes in an equivalent or better way than the entrenched flow. CI's strength lies in its ability to connect privacy with social life, recognising that informational norms are grounded in the values and functions of distinct social contexts such as medicine, education, or politics.

This makes CI exceptionally well suited for evaluating technological practices that alter social life. The framework is therefore capable, at least in principle, of recognising new informational norms that arise in response to changing conditions in the flow of information.

Yet when the time comes to use the CI heuristic on new and emergent technologies, many challenges arise. First, scholars have pointed out its inherent conservatism (Rule 2019; van de Poel 2022). The inherent conservatism of CI makes it possible to justify problematic technologies when they follow established norms (Austin 2003; van de Poel 2022). The case of Marie is a paradigmatic example of this type of situations, where the ERT has *already* been incorporated into hospital's protocols, ensuring patient-safety and professional excellence. Marie's case assumes that the ERT has already been incorporated into hospital's protocols. This matters for CI because institutional contexts are precisely those where entrenchment is strong.

This means that the norms about roles, attributes, and transmission principles are formally codified and reinforced by hierarchy, compliance systems, and professional standards. Once an informational practice becomes part of the institutional fabric, CI risks classifying it as already appropriate, on the grounds that the relevant flows now match the prevailing norms.¹⁷³ By grounding its analysis in entrenched norms, CI struggles to ratify harmful transformations once they have become “normal.” In Marie’s case, CI could conclude – if we follow the official norms strictly – that ERT-mediated affect assessments are contextually appropriate because they are authorised by the institution, justified by a mission of patient safety, and embedded in professional role expectations. But this verdict would miss the core harm. Marie’s loss is not primarily tied to the context’s values, goals, or ends, but about the structural defeat of *her* expressive discretion as a human, which undermines her ability to sustain the self that she sees best fit, but also that her role and authority expect.¹⁷⁴ The hospital’s safety aim must be weighed against the fact that emotional discretion affords individuals like Marie the capacity to sustain a plurality of roles (mother, divorcee, surgeon) and to follow role-bound norms of expressiveness, such as those tied to her role as a surgeon (see §5.5.2).

Emotional privacy, by contrast, does not treat existing norms as the ultimate criterion. Instead, it evaluates whether the introduction of ERTs in a specific context undermines individuals’ ability to control their affective information and whether the defeat of intentional EEs’ functions disrupt the social practices and their own societal benefits.

In the case of Marie, by defeating the functions of intentional emotional expressions, the technology risks weakening social practices that rely on them. Emotional privacy is sensitive to disruptions caused within affective communication channels. Hence, it can explain why some use of ERTs are problematic even in contexts that appear formally acceptable. By contrast, CI’s reliance on entrenched norms can normalise harmful practices once they become institutionalised.

Even when access, purpose, and authorisation are satisfied, the defeat of expressive discretion damages the practice itself through the undermining of individuals’ expressive discretion. Focusing only on the context’s internal structure, without an interest to the individuals’ well-being, result in missing harms that affect the very practice under evaluation.

¹⁷³ See Austin (2003), on the argument of 24/7 CCTV-camera surveillance.

¹⁷⁴ See Gstrein & Beaulieu (2022) for a critique of CI as unable to protect individuals’ rights.

Moreover, as pointed out by Rule (2019), CI also makes it hard to reach a normative diagnostic when, contrary to the hospital's case, the context is ill-defined, and/or partly merging with other context(s). This is the case of Alfred's situation, where the ERT seems to be both in the context of friendship, which would include everyone physically present at the table (or at least the ones that saw the push notification), and the context of self-care, which focuses only on Alfred's personal assessment via the use of the ERT. This merging of contexts, which could in some other cases be neatly divided, is in this case produced by the social disruption of the technology that is under investigation. This disruption causes, as I have argued elsewhere (Prégent 2025b), the context to undergo significant reshaping, which may also involve, alongside a severe collapse with other contexts, a re-hierarchisation of the prevailing norms of the new (post-disruption) context. But focusing only on the fact that oftentimes, contexts such as friendship or self-care are often ill-defined, we can see that in Alfred's case, there are no codified informational norms governing *affective legibility* among friends, and the values at stake are at best implicit rather than formally articulated such as in Marie's scenario. CI can flag that a self-care emotional inference ("arousal 87%") becoming visible in a friendship setting is a *prima facie* violation, since the flow deviates from the informal expectation of defeasible emotional interpretation. But when CI reaches its normative stage, it must determine whether the practice should evolve to accommodate this new flow. Here the framework lacks the resources to explain why this new degree of legibility is objectionable because CI ties its assessment to entrenched norms. In informal settings, such as friendship, where norms are weak, shifting, or tacit, CI has difficulty identifying the decisive contextual value that must be upheld and hence it is difficult to determine whether this new degree of legibility is upholding the context's integrity. As Rule (2019) and van de Poel (2022) warn, CI risks either normative drift, which would allow the practice to evolve with a harmful technology, or normative hesitancy, which would result in a final decision that lacks moral support.

Finally, the last challenge pertains to the fact that some contexts, even though in principle formally regulated, struggle with never-ending disagreements on the norms that should regulate the flow of information. This is what is happening in the case of the Minister, and ultimately, the question of whether and to what degree public authorities should have privacy when in office. Some hold that transparency maximises accountability; others hold that maintaining privacy or even secrecy is essential to public trust, political unity, and institutional dignity (Hood 2006; Rosen 2001). CI performs poorly in such cases because its normative tool depends on the

very contextual values that are under dispute. When the context's purpose is contested, CI can flag that a new flow deviates from entrenched norms, but it cannot decisively say whether those norms ought to change.

In the next section, I will propose a way to supplement the moral part of the CI framework with emotional privacy, when relevant, to help remedy some of its internal tensions regarding the assessment of ERTs.

In sum, while emotional privacy is largely compatible with the different variants of informational privacy, none of these variants can on their own cover what emotional privacy protects. Hence, in the context of ethical analysis concerning ERTs and technologies of this kind, the moral value of emotional privacy can be added to the already existing weight of privacy concerns. This is what I turn to now.

6.5. Assessing emotional privacy's moral weight in future societal trade-offs

There are two reasons to define the moral right to emotional privacy as an interest-based, *pro tanto* moral right. First, establishing it as a moral right secures its status. It ensures that its ethical and normative weight must be acknowledged in relevant analyses. Second, identifying it as a *pro tanto* right has practical implications. As with other moral rights in (liberal-democratic) societies, emotional privacy may be subject to limitations when it conflicts with other important rights or interests. This approach is consistent with the standard treatment of rights conflicts in modern liberal-democratic societies and with how privacy is usually conceived in normative analyses,¹⁷⁵ where limitations may be placed upon it, but any compromise must be reasoned, proportional, and accountable.

One of the aims of this chapter is to reconsider the weight of privacy in the inevitable trade-offs that liberal-democratic societies will have to make vis-à-vis the use of new technologies that profoundly reshape our social life. I thus here pre-empt techno-optimists' doubts on the actual "gravity" of privacy violations compared to the potential uses and opportunities that these new technologies will bring to the table. In the particular case of ERTs, at first sight, the prospect of exposing people's emotional lives may appear beneficial. One might think, at least on a *prima facie* basis, that such transparency would foster (or force) honesty, eliminate the space for

¹⁷⁵ See Moore (2010: 99) and Allen (2015: 174).

manipulation, and effectively counteract in real-time the everyday strategies of deceit or concealment that often complicate human interactions. ERTs, in this techno-optimist view, would cut through the noise of social performance and expose the ugly but nonetheless authentic picture of how people truly feel. Such exposure could seem especially welcome in contexts where sincerity is deemed valuable – in politics, in the workplace, or even in intimate relationships. By piercing through emotional expressions' shield and exposing concealed emotions and social motives, ERTs might appear as a new way to eliminate ambiguity and ensure social order grounded in (forced) honesty and trust.

Yet this line of thought now not only has to dialogue with the “mainstream” privacy losses that are the (il)legitimate collection, storage, and commodification of individuals' (emotional) data, it also needs to demonstrate that ERTs' use is not going to disrupt society's current (affective) social equilibrium. Ethical assessments typically weigh competing values involved in using technological tools. Within the cluster of privacy, emotional privacy brings new reasons to favour privacy in those settings. Below I aim to show practically what the proposed concept and the following moral right to emotional privacy may bring to the overall value of privacy in relevant trade-offs. Importantly, in these trade-offs, ERTs' advocates will have to justify the use of the technology by weighing societal interests against another. This is because the interests served by emotional privacy are held by individuals but are also constitutive of society's well-being. Trade-offs involving ERTs therefore no longer oppose individuals' (privacy) interests to society's interests (e.g. security, efficiency, or community), instead, they require an explicit balancing between competing societal interests.

By extending privacy as not only an individual but a societal interest, and by carefully unfolding the various functions of privacy in sustaining essential practices of social life (and the consequences of a disruption of such practices), emotional privacy adds a significant weight to the more general value of privacy in our techno-social world. What is protected by emotional privacy and threatened by ERTs is nothing less than the affective infrastructure that supports the self and social life.

If the core interests at risk also include the ones inherent to social practices constitutive of a liberal-democratic society, then the evaluation of ERTs must ask: Does the deployment preserve enough selective opacity for expressions to fulfil their communicative and regulative roles? If not, there is, at minimum, a *prima facie* infringement of a moral right to emotional privacy.

Consequently, claimed benefits of such exposure must be weighed against the constitutive social goods at risk. Nevertheless, whether to uphold other values against (emotional) privacy or not is a decision that needs to be justified by taking into account the societal benefits currently obtained under emotional privacy.¹⁷⁶

With the concept of emotional privacy and the proposed moral right to emotional privacy in view, I now test whether the weight of privacy changes (when relevant) vis-à-vis other values in anticipation of future societal trade-offs. To do so, I propose to see how emotional privacy may substantiate two different types of privacy assessments. First, I slightly modify CI to enable the CI heuristic to take into account new moral rights to privacy such as mental privacy and emotional privacy, and then investigate whether these rights may recalibrate privacy's overall weight against other important values such as security, safety, and efficiency. Second, I analyse how emotional privacy may supplement informational privacy's reasons in a classic ethical assessment of new and emergent technologies, by relying on the HLEG ALTAI ethical framework, which recommends relying on criteria such as prevention of harm, proportionality, subsidiarity, and respect for privacy and data governance.

6.5.1. Substantiating the CI moral part of the heuristic

Several scholars have already attempted to modify the CI heuristic to achieve broader or other aims than the ones deriving from the “original” CI. Natalia Criado and Jose M. Such (2015) propose a computational model of Implicit CI, in which the idea is to use CI's rationale on *implicit* informational flows (rather than entrenched and collectively recognised ones) in ill-defined contexts such as the ones present in Online Social Networks. Elizabeth O'Neill (2022, 2023) suggests a General Contextual Integrity account in which the norms followed are expanded from privacy norms to societal norms, in an attempt to evaluate technological change. Sebastian Benthall and Ido Sivan-Sevilla (2024) propose a formulation of CI, called Regulatory CI, in which informational flows are modelled and audited using Bayesian networks and causal game theory.

¹⁷⁶ Having said that, note that the loss of expressive discretion does not always lead to an (emotional) privacy violation. For example, consider a theme-park mascot performer in a bulky costume who must constantly display “cheerfulness” by clapping the two paws of the costume together. Here expressive *discretion* is entirely eliminated, yet no (emotional) privacy interest is violated, because the constraint on emotional expressions prevents the mascot's emotional expressions from participating in the ordinary affective communication channels through which expressions typically carry information about inner emotional states and/or social motives. In this case, there is no risk of either actual or standing emotional privacy loss.

My proposal follows this trajectory but with a narrower purpose. I aim to supplement CI's moral stage with emotional privacy when the technology under evaluation directly targets emotional legibility. As noted in the previous section, CI usually adequately flags departures from entrenched flows, but at the normative stage it struggles whenever entrenched norms are absent, already reshaped by institutional power, or deeply contested. By introducing the concept of emotional privacy as a background condition without which core practices and interests cannot be realised through affective communication channels, CI gains a stable normative anchor from which to reach firm, morally supported decisions with technologies such as ERTs.

In Alfred's informal friendship setting, CI can identify a *prima facie* violation (an *a priori* self-care directed inference becomes visible in a friendship context), but the norms that should govern affective legibility among friends are such that a final diagnosis is hard to achieve. Once emotional privacy is recognised as a constitutive value of the practice, protecting expressive discretion as the condition for self-presentation, CI's moral stage has another moral tool to identify and reject the practice of forced legibility. If Alfred or his friends use a CI-supplemented heuristic, he (they) may better see why this ERT, and in this case, the push notification more specifically, might be harmful and detrimental to Alfred's interests, but also for the interests of the group itself. Hence, he may re-evaluate the "cost" of being self-aware of his affective states via the use of ERTs. On a design-level, perhaps it is better to avoid externalising push-notifications in favour of private, self-directed feedback and overview charts through, for instance, voluntary and deliberate access to the app or platform.

In Marie's institutional setting, hospital protocols have already adopted ERT-mediated affect assessments under the principle of safety and professional excellence. Here CI may be tempted to treat the flows as presumptively appropriate. Yet Marie's core loss is not misuse of data but the structural defeat of expressive discretion, the very capacity that sustains her authority as a surgeon, trust, and role-appropriate composure. By treating emotional privacy as a background condition of professional practice (i.e. because it protects expressive discretion which itself enables the maintenance of role plurality and allows individuals to adhere to role-bound norms of expressiveness), CI's moral stage can perhaps withstand institutional entrenchment by delivering the following supplemented verdict: where ERTs disable the expressive preconditions of role-appropriate composure, the practice is harmed and the flow must be prohibited or radically redesigned (for example, strictly local self-feedback to the surgeon or no mandatory debrief).

In the Minister's public context, the problem for CI is likely to be the persistent contestation over the purpose of the practice itself. Some argue that maximal transparency best serves accountability while others hold that privacy or secrecy is necessary for the regulative and cohesive functions of public address (Hood 2006; Rosen 2001; Mokrosinska 2015). CI's moral proscription weakens when the very values that should ground its appraisal are under dispute. The framework can register deviation from existing norms but struggles to convincingly adjudicate between rival value-interpretations of the practice (Rule 2019). Adding emotional privacy to the evaluation strengthens the value of privacy against competing values. In the case of ERTs, CI gains a moral ground for a default presumption in favour of managed affective opacity where the practice's success depends on it (e.g. apologies that mediate collective anger and symbolically reaffirm civic bonds).¹⁷⁷ The burden of justification then shifts. Those who would entrench real-time (audience) recognition must show *why* the loss of social regulation and cohesion is *justified* by countervailing reasons.¹⁷⁸

Emotional privacy supplies CI with a background condition that must be preserved if the practice's core functions are to be realised. With this supplement in place, evaluators can use CI to reach morally supported decisions on ERTs.

6.5.2. Evaluating the moral weight of emotional privacy on the HLEG ALTAI framework

The second test for assessing emotional privacy's moral significance in relevant trade-offs targets its capacity to recalibrate ethical assessments of new technologies beyond CI, particularly in

¹⁷⁷ It could be argued that the practice of apology itself depends on the apology being sincere. Hence, ERTs that reveal a lack of genuine sincerity would only help the practice. Although on a *prima facie* basis this seems intuitively right, closer examination of the practice of apology reveals that it is not always the case that its success depends on the apology being genuine. There are many occasions where a non-genuine apology may serve the broader interests of the practice. Even though an apology is typically recognised as an acknowledgement of a wrong, it is not always the case that the person has fully accepted this wrong nor that they regret their actions. Additionally, people who apologise on behalf of others, like in the Minister's case, do not always feel genuine regret or remorse vis-à-vis the given apology (perhaps something closer to a sense of duty instead). This does not mean that the practice of public apology is then unsuccessful, on the contrary. Delivering a convincing public apology when needed, despite not experiencing the genuine emotional correlate, may also be a way to protect one's well-being. Think of funeral workers, soldiers, police officers, first responders, doctors, nurses, and others who have to deliver (too) many 'formal' apologies to families of the deceased. It seems only natural that they sometimes detach themselves from experiencing genuine remorse, regret, or emotions of powerlessness, although this does not seem to take anything away from the recipient's side when it is performed well, and this symbolic and performed apology retains its value.

¹⁷⁸ Here there remains a practical problem of feasibility. Even though such justification could win a CI moral evaluation, the reality is that the Minister's public discourse might always be assessed through ERTs if fully functional ERTs become available via a phone-based app, for instance.

widely used ethical evaluation frameworks. The High-Level Expert Group on Artificial Intelligence (HLEG) recommends in its Ethics Guidelines for Trustworthy AI that emerging technologies be assessed through criteria such as prevention of harm, proportionality, subsidiarity, and respect for privacy and data governance (HLEG 2019). These criteria articulate the need to minimise intrusions, avoid unnecessary inferences on individuals, and ensure that any encroachment on fundamental interests is strictly justified. Yet, as with CI, these assessments depend for their force on the underlying reasons we have for valuing the interests protected (viz. self-presentation, role plurality and role-bound norms of expressiveness, and social regulation and cohesion). Where those reasons are thin or underdeveloped, privacy tends to be outweighed by competing goals – efficiency, optimisation, safety, transparency, or innovation – especially in the context of predictive or diagnostic technologies, as the benefits of these kinds of technology are often sold under the branding of societal interests. This dynamic mirrors well-known critiques in AI ethics where frameworks often identify what must be weighed, but lack the resources to explain *why* certain values should carry more normative force in hard cases.

If we take ERTs and run them through an ALTAI-style evaluation without emotional privacy in hand, it might not do justice to the interests protected by privacy simply because the evaluators do not possess the conceptual tools to fully grasp the whole range of interests at stake. Under the criterion of prevention of harm, for instance, ERTs may be judged favourably if they are thought to pre-empt risk, facilitate early detection of emotional crises linked to violent acts, or enhance interpersonal (or self-) understanding. Under proportionality, the argument might lean towards the view that modest privacy losses are acceptable to win important gains in safety, productivity, or transparency. And under subsidiarity, it may seem that no less intrusive alternative exists for accessing the relevant emotional insight that put us in a situation where we can obtain these gains. In short, without a robust account of why emotional privacy matters for social life, an ALTAI-like assessment risks reproducing the same results already seen in public discourse, where the main question concerns how emotional data are handled, and privacy becomes a narrow data-protection constraint rather than a substantive social value.

Emotional privacy might help recalibrate these trade-offs, since it enriches the meaning of the privacy-related criteria¹⁷⁹ and connects them more tightly to harm prevention, proportionality, and subsidiarity. In the case of ERTs, this leads to different practical outcomes. Under

¹⁷⁹ See HLEG (2019), requirement #3: Privacy and data governance.

prevention of harm, emotional privacy expands the privacy-harm landscape beyond data misuse, discrimination, or chilling effects to include the structural harm of undermining expressive discretion, from which both individual and social harms stem. This reframing shifts the burden of justification. ERTs' advocates must now show not only that their systems produce benefits that outweigh individual and data-focused privacy interests, but that those benefits justify displacing the social interests protected by managed opacity.

Under proportionality, emotional privacy raises the justificatory threshold, since losses to expressive discretion cannot be dismissed as minor or “merely informational,” because they undermine the basic architecture of interpersonal and collective life. It becomes harder to claim that so-called ‘small’ intrusions are a price worth paying if those intrusions undermine the conditions of self-presentation, role plurality and role-bound norms of expressiveness, and/or social regulation and cohesion. It also recasts proportionality as a comparison of societal interests. Real-time affect recognition may fail the proportionality test unless the intrusion is strictly necessary and the benefits cannot be achieved by less intrusive means.¹⁸⁰

And under subsidiarity, emotional privacy forces the analysis to move to the design choice itself, rather than only to the mitigation of its consequences. Instead of asking whether there is a less invasive way to access the same emotional information, we may ask whether emotional legibility should exist at all in the practice. And in cases where monitoring is genuinely desirable,¹⁸¹ we must ask how to architect it so that opacity is maximised within affective communication channels. This places the design choice itself under ethical scrutiny. Subsidiarity now better supports design choices that preserve managed affective opacity by default, rather than merely limiting data sharing or retention. The ethical assessment thus shifts from a logic of (emotional) privacy trade-off in service of greater interests to a logic of (emotional) privacy as a constitutive condition of social life that can be overridden only under compelling justification.

Seen through this lens, some changes are in order within the three scenarios. In Alfred's dinner, the acceptable design becomes one that avoids externalising push-notifications in favour of private, self-directed feedback and overview charts through, for instance, voluntary and

¹⁸⁰ An example of such technology would be real-time affect recognition in classrooms to monitor student “engagement”. Upon ethical assessment, it might be decided that the technology does not pass the proportionality test and that there are less intrusive ways to achieve the same end.

¹⁸¹ An example of such technology would be the use of fully functional ERTs in residential dementia care, where they could be used to detect escalating distress in residents who can no longer reliably verbalise pain or fear, prompting faster comfort or medical intervention (see Gerbaudo-González et al. 2025).

deliberate access to the app or platform. In Marie's hospital, unless the institution can demonstrate that only legible affect achieves the safety aim, the design must shift to private or confidential checks that keep expressive discretion intact. Hence a general alert, visible to everyone, needs to be replaced by a less exposing system. The Minister's case enables us to collectively reflect on the impact of a widespread endorsement of ERTs and what this normalisation of the use of ERTs would potentially have on some EEs' social functions on which democratic societies currently rely. It brings the individual level and the societal level of interaction developed in this work closer together, showing how cases such as Alfred's self-monitoring and the Minister's apology will likely be co-existing. The Minister's apology is, ultimately, rendered legible by people who, like Alfred, will (want to) possess such technology, perhaps for their own self-assessment at first but also probably for the assessment of others as well soon after.

The takeaway is thus one that situates itself mainly in the practical domain. Pre-emptive ethical evaluations of ERTs should ask whether the use of the technology impinges on expressive discretion and where expressive discretion is recognised as a constitutive good of social life, justify why the loss of emotional expressions' social functions would be warranted. Only where justification succeeds should deployment proceed, and then only with safeguards that preserve as much selective opacity as the practice can bear. Design-level limits on affective legibility should be implemented to prevent the use of ERTs from impinging on expressive discretion. This preserves the individual and societal interests secured by emotional privacy while leaving room for warranted infringements in clearly defined contexts.

Taken together, the supplemented CI analysis and the recalibrated ALTAI assessment show that emotional privacy is strong enough to potentially tip future ethical assessments of ERTs in favour of privacy or at least increase the moral significance of (emotional) privacy against other competing values. It does so by introducing emotional privacy as a practice-sustaining value, capable of grounding firm normative conclusions within both contextual and general ethical frameworks. In this way, emotional privacy provides precisely what anticipatory analyses of socially disruptive technologies require, namely, moral justification that pre-emptively resists the pervasive normalisation of affective surveillance.

6.5.3. Emotional privacy and the question of consent

Before closing, and as promised above, I will say a few words on the question of consent and its potential power in legitimising what would otherwise be considered as wrongful losses of emotional privacy. A natural objection to the claim that fully functional ERTs constitute a distinct privacy threat often appeals to the notion of informed consent. If individuals knowingly and voluntarily authorise affective monitoring, it may seem that no wrongful loss of emotional privacy occurs. On this view, consenting to the deployment of an ERT amounts to a justified waiver of one's (moral) right: the individual has decided to lower their affective opacity, and therefore cannot subsequently complain that it has been reduced.

At first glance, this objection appears quite powerful against emotional privacy since I have situated it within the broader Control-based view of privacy. If privacy consists in controlling access to information about oneself, then authorising affective monitoring appears to be a paradigmatic exercise of that control (and is indeed, how many consent-based approaches are justifying themselves). Consent is typically theorised within an individualistic paradigm and tied to the notion of control over discrete informational transactions (Faden & Beauchamp 1986; O'Neill 2003). It enables a person to approve, permit, or waive a right in well-defined and temporally bounded circumstances. The notion of consent, when grounded in Control-based accounts of privacy, is a way to provide individuals with the tool to protect their autonomy and life choices. Consent would then be enough to justify the use of ERTs because it would enable individuals to voluntarily allow (or refuse) the disclosure of their inner emotional life under chosen conditions.¹⁸²

However, to be considered “valid consent”, this line of reasoning must presuppose that what is at stake is merely *actual* emotional privacy, which is concerned with the success or failure of control of discrete episodes of disclosure. Since informed consent is valid when given by individuals regarding particular transfer of information, it assumes that individuals give their consent regarding the decision of whether to reveal particular affective states on particular occasions. Yet the argument developed in Chapters 5 and 6 shows that the core normative concern raised by fully functional ERTs lies at the standing level. What these systems threaten, by design, is not only the success of specific privacy mechanisms but the background expressive ability that makes the deployment of such mechanisms possible (and meaningful).

¹⁸² See Chapter 5 for the discussion on disclosure mechanisms.

The relevant question, therefore, is not simply whether individuals can consent to a particular act of monitoring. It is whether they can waive the structural conditions that sustain expressive discretion itself. To consent to the exposition of (inner) affective information p at a given time is one thing. To consent to the systematic defeat of one's ability to regulate her affective legibility altogether is another. The latter implicates the whole communicative infrastructure within which emotional expressions perform their social functions.

One way to answer this question would then be to say that whether consent may be used as a form of waiver depends on whether intentional emotional expressions are peripheral or constitutive of the practice itself. Where emotional expressions play only an instrumental or marginal role, a tightly bounded waiver (through consent) of emotional privacy may be defensible. But where calibrated opacity is constitutive of the practice, consent is not fitting since the consequences extend past individual's autonomy and concern societal well-being. This is why consent cannot automatically legitimise the defeat of emotional privacy.

Hence, in contexts where emotional expressions are *not* constitutive elements of the social practice, consent may indeed operate as a valid waiver. Consider settings in which intentional EEs are largely instrumental. In such cases, authorising ERT monitoring may resemble other forms of informational trade-off. Provided that consent is informed, specific, and revocable, individuals may lower affective opacity through consent.

Airport security screening illustrates this possibility. Suppose travellers are informed that affective analysis will be used during checkpoint control and are asked to consent as a condition of expedited screening. The monitoring is temporally limited, confined to a narrowly defined interaction, and justified by security risk assessment. Crucially, airport screening is not structured around reciprocal emotional exchange in the way that friendship, teaching, or professional collaboration might be. Although consent in such contexts still remains imperfect given potential background pressure associated with travel needs, it may be possible to treat (standing) emotional privacy as waivable through consent within this narrowly delimited setting.

Yet the airport case is likely the exception rather than the rule. I mentioned that consent, when valid, enables a person to approve, permit, or waive a right in well-defined and temporally bounded circumstances, meaning that individuals may consent (or not) to the disclosure of particular affective states on particular occasions. Emotional privacy, by contrast, concerns not

merely specific disclosure but the standing ability to regulate how one's affective life becomes socially legible across contexts. The overlap between these two frameworks is therefore limited. Consent may legitimise specific informational transfers whereas emotional privacy protects the background conditions that make such transfers meaningful in the first place. When we return to the three technomoral scenarios, the limits of consent as a way of waiving emotional privacy right become more apparent.¹⁸³

In Alfred's case for instance, the ERT is embedded in ordinary social interaction through his personal smart glasses and smartwatch. By consenting to the ERT's monitoring, Alfred does not consent to disclosing particular affective information on particular occasions to particular recipients but in fact "consents" to the total defeat of his own expressive ability. Social life relies on calibrated opacity which allows Alfred to conceal irritation, mask anxiety, or reveal his vulnerability selectively. In this situation Alfred's "consent" is not limited to a single disclosure but amounts to authorising a standing condition in which his expressive ability is systematically bypassed. In such circumstances, the structural harm extends beyond consenting to the disclosure of particular affective information on particular occasions to particular recipients. Hence, the notion of consent is ill-fitted to legitimise on its own a potential trade-off of emotional privacy.

Marie's case presents a related but distinct problem. She consents (at least formally) to institutional monitoring in her professional role as a surgeon. Setting aside concerns about coercion within hierarchical employment relations, a deeper issue arises regarding specifically emotional privacy. Marie's affective life is not neatly divisible into role-specific compartments. Algorithmic inferences conducted within the hospital inevitably also covers her broader personal circumstances. The ERT collapses distinctions between Marie-the-surgeon and Marie-the-divorcee, disclosing affective information about Marie-the-divorcee within the context of Marie-the-surgeon when she sought to keep it concealed. Although she has consented to monitoring within a defined role, the technology defeats her ability to regulate which aspects of herself become socially legible in that professional context. In practice, consent to institutional monitoring of her professional role as a surgeon will always *ipso facto* also be consent to institutional monitoring of her entire self. There can be no possibility to maintain role-bound norms of expressiveness once the conditions for role-bound expressiveness have been structurally undermined. The loss of standing emotional privacy occurs because ERTs'

¹⁸³ On the limits of informed consent, see Onora O'Neill (2003) "Some limits of informed consent".

inferences bypass the privacy mechanisms that ordinarily sustain professional boundaries. This is, once again, an issue that is much broader than what individual consent can hope to legitimately cover.

A similar dynamic is visible in the Minister's case. Even as a public official speaking in a public setting, she relies on selective disclosure and calibrated emotional presentation to perform the regulative and cohesive functions of her role. When members of the public use phone-based ERT applications to render her affective states continuously legible, the system does not merely reveal emotions tied to her official capacity. It exposes her as a whole person, collapsing the distance that makes public emotional performance an act that is socially functional. Here too, consent would not be sufficient by itself to legitimise the defeat of expressive discretion.

Across these scenarios, the core insight is that waiving standing emotional privacy is not equivalent to authorising a discrete informational flow. It is to permit a transformation of the communicative environment in which emotional expressions operate altogether, which is a much broader issue. Hence whether consent may be used as a valid waiver depends on how ERTs' use impacts the social functions of EEs in particular situations and whether these functions are constitutive to the practice itself. Where expressive discretion is constitutive of that practice, its systematic defeat cannot be legitimised solely by individual consent.

Consent, then, remains relevant in some contexts but in a limited manner. It may justify specific, bounded uses of ERTs vis-à-vis specific individuals but only in contexts where justifications have already been given regarding the role of intentional EEs' functions. It cannot single-handedly justify the restructuring of affective infrastructure on which social life depends. Recognising a moral right to emotional privacy therefore places principled constraints on the extent to which inner affective life may be rendered fully and continuously legible, and the decision about whether we have good reasons to waive this moral right extend way beyond consent, to the assessment of existing social goods sustained by managed opacity and proposed counter benefits of affective transparency.

6.6. Conclusion

This chapter asked whether emotional privacy could ground a distinct moral right and whether recognising that right changes how we ought to evaluate technologies like ERTs. To answer these questions, I formulated a moral right to emotional privacy, grounded in the protection of

expressive discretion as a constitutive condition for self-presentation, role plurality and role-bound norms of expressiveness, and the regulative and cohesive functions of emotional expression. I distinguished acceptable losses of emotional privacy from wrongful ones through the notion of defeat, which marks the point at which expressive ability is defeated.

I then demonstrated, via three technomoral scenarios, how fully functional ERTs can trigger that violation at three levels of interaction. Alfred's case showed the loss of self-presentation in ordinary interaction (individual level). Marie's case displayed the erosion of role plurality and adherence to role-bound norms of expressiveness (interpersonal level). And the Minister's case illustrated the collapse of public emotional expressions' regulative and cohesive functions (societal level). In each, the primary harm arose from an imposition of legibility that disables expressive discretion and managed affective opacity.

I then turned to neighbouring moral rights and leading privacy theories (access-based, control-based, Contextual Integrity) and argued that they fail to capture the harms caused by the collapse of managed affective opacity. Control-based and access-based variants target boundary crossings and authorisations rather than the conditions that make expressions socially functional. And Contextual Integrity tends to detect departures in informational flows rather than architectural defeats of expressive discretion. Emotional privacy therefore fills a genuine normative gap.

Finally, I examined the practical implications for evaluation and governance. Supplementing Contextual Integrity with emotional privacy furnished CI's moral stage with a background condition (expressive discretion) that may help to stabilise the evaluations in informal, institutionally entrenched, and contested contexts. Bringing emotional privacy into ALTAI-style assessment reframed prevention of harm, proportionality, and subsidiarity, shifting the focus from "data handling" to design choices about legibility, and from "individual vs. society" to "society vs. society" trade-offs, recalibrating the assessment between existing social goods sustained by managed opacity and proposed counter benefits of affective transparency. Lastly, I address the question of informed consent as a way to waive the moral right to emotional privacy and argue that informed consent can only play a limited and secondary role, which must come after the ethical assessment of the existing social goods sustained by managed opacity and proposed counter benefits of affective transparency. The result is a higher justificatory threshold for ERT deployment and a principled preference for opacity-preserving technological design.

Thus, the chapter established that (i) fully functional ERTs threaten a distinct kind of moral interests; (ii) emotional privacy protects this distinct kind of moral interests in a way that broader accounts of informational privacy cannot; and (iii) recognising a moral right to emotional privacy strengthens leading evaluative frameworks and affects governance recommendations. Protecting emotional privacy is thus not only conceptually coherent but normatively and practically necessary for the governance of socially disruptive affect-surveillance technologies.

7. General conclusion

This dissertation has examined the ethical, social, and normative implications of fully functional ERTs. While current systems remain flawed, contested, and in many cases demonstrably unreliable, the ongoing developments in affective computing, together with sustained commercial, political, and institutional interest in ERTs, suggest that fully functional ERTs may be successfully developed in the near future. In light of this trajectory, the guiding approach of the thesis has been anticipatory. Rather than examining the ethical, social, and normative consequences of existing systems, the dissertation has asked what would follow if ERTs were to work as intended and which social and normative structures would be challenged by technologies capable of reliably detecting affective states.

By assuming the possibility of fully functional ERTs, the thesis has shown that their integration into everyday social life would risk unsettling many of the core expressive, epistemic, and normative conditions on which interpersonal and societal functioning depend. A close examination of existing privacy frameworks reveals that mainstream accounts of informational privacy currently lack the conceptual resources to capture these harms. This exposes a conceptual and normative gap in contemporary privacy theory. The dissertation addresses this gap by introducing and defending emotional privacy as a subconcept within the informational privacy cluster. Emotional privacy protects individuals' expressive ability, and thereby the conditions under which individuals can modulate and negotiate the legibility of their affective life *in situ*.

This protection secures at least three core socio-normative interests that are dependent on individuals' expressive ability. These are (i) self-presentation (individual level), (ii) role plurality and adherence to role-bound norms of expressiveness (interpersonal level), and (iii) social regulation and cohesion (societal level). By articulating these functions, the thesis demonstrates that emotional privacy identifies a class of harms that is distinct from, yet compatible with, standard informational privacy concerns.

On this conceptual foundation, the dissertation develops emotional privacy into a candidate moral right. Through the use of technomoral scenarios and the evaluation of ERTs via Contextual Integrity and ALTAI-style assessments, the thesis shows that recognising a moral right to emotional privacy alters the normative weight of privacy in ways that can reshape or even shift the ethical verdicts reached about affective technologies.

The central claim defended throughout the dissertation is thus the following: when technologies can reliably infer emotional states, the resulting harms are not confined to data privacy, manipulation, or the exploitation of sensitive information. A distinct, but interconnected, set of harms emerges, centred on the expressive, relational, and normative structures on which social life depends. Ethical and regulatory assessments of ERTs must therefore consider not only familiar informational privacy risks but also the risks of undermining the social infrastructure that makes self-presentation, role negotiation, and social cohesion possible.

Main contributions of the dissertation

One central contribution of the dissertation is its conceptual and ethical focus on fully functional ERTs. By shifting attention from current systems to fully functional ERTs, the dissertation moves beyond current discussions, which largely criticise existing ERTs for their unreliability, bias, and lack of scientific grounding. To date, little scholarship has examined what would follow, ethically, if these technical limitations were overcome. Hence, this dissertation is one of the first to question the ethical concerns that would follow from these systems working as intended, providing ethical guidance ahead of technological development.

A second contribution is the development of an integrative and function-first account of emotional life, articulated across Chapters 3 and 4. This account provides the theoretical foundation necessary for understanding what is at stake when technologies target affective states. At the core of this contribution is the proposal of EEs as serving two different kinds of functions for humans and machines. While humans rely on embodied, contextual, and norm-grounded informational signals, ERTs are assumed to rely on probabilistic regularities in sets of physiological and behavioural changes, thereby creating a second, distinct mode of recognition. This establishes that ERTs operate through a fundamentally different epistemic route into affective life. This bi-functionality thus creates a conceptual bridge between work in philosophy of emotion and the ethics of emerging technologies.

A third contribution lies in the introduction and defence of emotional privacy as a subconcept within the informational privacy cluster. Using conceptual engineering, the dissertation argues that emotional privacy protects not only actual control over particular disclosures but also the background expressive ability that enables individuals to modulate what their expressions count as across contexts. This dual-level structure clarifies why the harms posed by fully functional

ERTs arise not only from the loss of actual control over one's emotional expressions but also from the defeat of the expressive ability that underlies the production of these emotional expressions.

A final major contribution of this thesis is the normative development of emotional privacy into a candidate moral right within the informational privacy cluster. By comparing emotional privacy to existing theories of informational privacy, contextual integrity, and mental privacy, the dissertation shows that emotional privacy captures a distinct normative domain: the maintenance of expressive discretion against technologies that make affective life continuously legible. Fully functional ERTs, on the analysis offered here, restructure the affective and expressive fabric of social interaction, rather than merely extending ordinary human mindreading. The recognition of a moral right to emotional privacy thus adds a new normative dimension to the ethical and regulatory evaluation of affective technologies.

Some limitations and conceptual challenges

The dissertation faced some limitations and conceptual challenges. First, the potential impact of this thesis may seem, and is to a certain extent, restricted by its core assumption concerning the prospect of fully functional ERTs. It remains possible that technological development will never reach the level of reliability assumed here, or that progress unfolds along trajectories that diverge from those reconstructed in this work. Nonetheless, these limitations are not fatal to the central argument, since the conceptual and normative findings apply to any technology that produces a comparable increase in affective legibility, irrespective of whether it ultimately matches the functional assumptions described in the preceding chapters. This further implies that while removing the issues related to the technical limitations of current ERTs was necessary to isolate the specific kind of harms at stake, the concept of emotional privacy may also be relevant for the analysis of partially functional ERTs, provided that they increase affective legibility in ways that pressure existing expressive practices.

A second limitation concerns the adoption of an integrative perspective on emotions and emotional expressions. In unifying insights from divergent disciplinary approaches, the dissertation necessarily forgoes deeper engagement with foundational debates about the nature of emotions and emotional expressions, debates for which the “right” answers remain contested across philosophy, psychology, and affective science. The integrative and function-first account

developed here recognises those quandaries but deliberately brackets most of them in order to concentrate on their functional roles. While this choice was necessary to pursue the aims of the dissertation, it leaves unresolved ongoing questions about the nature of emotions and emotional expressions.

A further conceptual challenge arises from the tension between efforts to empirically ground the assumptions about fully functional ERTs and the anticipatory orientation of the project. Since it is still unclear how fully functional ERTs will be developed, I have been careful to stay at a high level of technical explanation regarding the recognition process of fully functional ERTs. Nonetheless, the anticipatory orientation forces the thesis to commit to some views concerning the technical development of technology-based emotion recognition, which might turn out to be erroneous or irrelevant in the near future. I have taken this level of uncertainty into account, and this is why Chapters 4 and 5 do not tie their conclusions to any particular technique regarding the recognition process, only to the idea that the desired outcome is reached, namely that the technology works as intended.

Potential directions for future research

Future research is needed in affective computing and regarding the use of fully functional ERTs. In direct continuation to this thesis, one interesting and needed line of research concerns the legal and political implications of a moral right to emotional privacy. While the dissertation establishes its normative foundations in privacy theory, further work should assess how such a right can be operationalised within existing legal frameworks, particularly in relation to data protection law, human rights instruments, and national and supranational regulations. Questions concerning enforcement mechanisms, emotional privacy as a legal subcategory of the human right to privacy, and the balance between emotional privacy and other collective interests (for example, in public health, education, or security contexts) appear to be primary candidates for further investigation. Addressing these issues would provide more concrete guidance for policymakers and regulators navigating the rapidly evolving landscape of affective technologies.

A second avenue for future research concerns the empirical and socio-technical dimensions of emotional privacy. The thesis has offered a largely Westernised philosophical and conceptual analysis; it would be valuable to investigate how individuals in different social and cultural contexts understand, value, and attempt to secure emotional privacy in practice. Empirical

studies on users', developers', and regulators' perceptions of affective legibility could help refine the concept of emotional privacy, test its intuitiveness, and reveal further mechanisms through which expressive discretion is preserved or undermined in everyday technosocial environments.

A third direction lies in extending the analysis to other affect-sensitive technologies beyond ERTs, such as social robots, conversational agents, and immersive environments that adapt to users' moods. These systems may not always be marketed as "emotion recognition" tools, yet they increasingly rely on affective inferences to guide interaction. Given the rise of affective computing in HCI, exploring how emotional privacy applies to these neighbouring technologies, and whether additional conceptual resources are needed, would deepen our understanding of the broader affective ecosystem in which future interactive technologies are likely to operate.

In closing, this dissertation has argued that as technologies grow ever more capable of rendering affective life transparent, the task of safeguarding the conditions under which we appear to one another as social and moral agents becomes increasingly urgent. Emotional privacy, as developed here, is not merely a barrier against intrusion but a tool to protect the opacity, ambiguity, and selective self-disclosure that are integral to humans and society's well-being. Whether fully functional ERTs ultimately emerge or not, the questions raised in this thesis speak to a broader challenge of our time: how to preserve the spaces in which persons can shape who they are, how they are seen, and how they relate to others within increasingly mediated worlds. Safeguarding those spaces is not only a matter of protecting information from malicious use but of protecting the social fabric in which meaning, agency, and shared life are formed.

Bibliography

- Affectiva. (n.d.). *Affectiva: Humanizing technology with emotion AI*. <https://www.affectiva.com> (Accessed on 12th of October 2024).
- Ahmad, Z. & Khan, N. (2022). A survey on physiological signal-based emotion recognition. *Bioengineering*, 9(11), 688.
- Akar, E. (2024). Emotional artificial intelligence: Introducing the concept of ‘emotional privacy’. In: Prifti, K., Demir, E., Krämer, J., Heine, K., Stamhuis, E. (eds) *Digital Governance, Information Technology and Law Series*, 39 (pp. 65-87). Asser Press.
- ALIZ-E Project. (n.d.). ALIZ-E: Adaptative strategies for sustainable long-term interaction. <http://aliz-e.org/> (Accessed on 4th of March 2022).
- Allen, A. (1988). *Uneasy access: Privacy for women in a free society*. Rowman & Littlefield Publishers.
- Allen, A. (1999). Coercing privacy. *William & Mary Law Review*, 40(3): 727-757.
- Allen, A. (2011). *Unpopular privacy: What Must We Hide?*. New York: OUP.
- Allen, A. (2015). Privacy, democracy, and freedom of expression. In Rössler, B. & Mokrosinska, D. (eds), *Social Dimensions of Privacy* (pp. 162-180). Cambridge University Press.
- Altman, I. (1975). *The environment and social behavior : privacy, personal space, territory, crowding*. Brooks/Cole.
- Amazon Web Services. (n.d.). *Amazon Rekognition*. <https://aws.amazon.com/rekognition/> (Accessed on 1st of December 2025).
- American Polygraph Association. (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, 40(4): 194-305.
- Amnesty International. (2023). The digital border: Migration, technology, inequality. *Amnesty International*.
- Aristotle. (1925). *The Nicomachean Ethics of Aristotle* (Ross, D. (ed. & trans.)). Oxford University Press.
- Armon-Jones, C. (1985). Prescription, explication and the social construction of emotion. *Journal for the Theory of Social Behaviour*, 15(1): 1–22.
- Austin, J. L. (1962). *How to do things with words*. Harvard University Press.
- Austin, L. (2003). Privacy and the Question of Technology. *Law and Philosophy*, 22(2): 119-166.
- AVATAR. (n.d.). AVATAR: Automated Virtual Agent for Truth Assessments in Real-Time. <https://discernscience.com/avatar.html> (Accessed on 1st of December 2025).
- Averill, J. R. (1980). A constructivist view of emotion. In Plutchik, R. & Kellerman, H. (eds.), *Emotion: Theory, research, and experience* (pp. 305–340). Academic Press.
- Averill, J. R. (1982). *Anger and aggression: An essay on emotion*. Springer-Verlag.
- Baker, R. (2013). *Before bioethics*. Oxford University Press.
- Bandes, S. A. (2014). Remorse, demeanor, and the consequences of misinterpretation, *Journal of Law, Religion and State*, 3: 170-199.
- Barnard, G., Berger, L., Dolezal, E., Gremsl, T., Jarke, J., et al. (2024). *The protection of mental privacy in the area of neuroscience: Societal, legal and ethical challenges* (Study No. EPRS_STU(2024)757807). European Parliament, Panel for the Future of Science and Technology (STOA).
- Bar-On, D. (2004). *Speaking my mind: Expression and self-knowledge*. OUP.
- Bar-On, D. (2010). Expressing as “showing what’s within”: on Mitchell Green’s, self-expression OUP 2007. *Philosophical Books*, 51: 212-27.

- Bar-On, D. (2013). Origins of meaning: Must we 'go Gricean'?. *Mind & Language*, 28(3): 342-375.
- Barrett, L. F. (2006). Are emotions natural kinds?. *Perspectives on psychological science*, 1(1): 28-58.
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin-Harcourt.
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest: a journal of the American Psychological Society*, 20: 1-68.
- Barrett, L. F. & Satpute, A. B. (2019) Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience letters*, 693: 9-18.
- Bedford, E. (1957). XII.—Emotions. *Proceedings of the Aristotelian Society*, 57(1): 281-304.
- Beecher, H. (1968). Ad Hoc Committee of the Harvard Medical School to Examine the Definition of Brain-death: A definition of irreversible coma. Special communication: report of the Ad Hoc Committee of the Harvard Medical School to Examine the Definition of Brain Death. *JAMA*, 205: 337-340.
- Benthall, S. & Sivan-Sevilla, I. (2024). Regulatory CI: Adaptively regulating privacy as contextual integrity. *ACM Symposium on Computer Science & Law*.
- Ben-Ze'ev, A. (2000). *The subtlety of emotions*. The MIT Press.
- Berggren, C. (2019). The cumulative power of incremental innovation and the role of project sequence management. *International journal of project management*, 37(3): 461-472.
- Bernat, J. L. (2006). The whole-brain concept of death remains optimum public policy. *Journal of Law, Medicine & Ethics*, 34(1): 35-43.
- Bittle, J. (2020). Lie detectors have always been suspect. AI has made the problem worse. *MIT Technology Review*. <https://www.technologyreview.com/2020/03/13/905323/ai-lie-detectors-polygraph-silent-talker-iborderctrl-converus-neuroid/> (Accessed on 16th of April 2025).
- Blum, L. (1982). *Friendship, Altruism, and Morality*. Routledge and Keagan Paul.
- Boden, M. (2018). *Artificial Intelligence: A very short introduction*. Oxford University Press.
- Bonard, C. (2021). *Meaning and emotion*. Doctoral dissertation, University of Antwerp.
- Boffey, D. (2018). EU border 'lie detector' system criticised as pseudoscience. *The Guardian*. <https://www.theguardian.com/world/2018/nov/02/eu-border-lie-detection-system-criticised-as-pseudoscience> (Accessed on 14th of August 2024).
- Bok, S. (1983). *Secrets: On the ethics of concealment and revelation*. Pantheon Books.
- Borak, M. (2024). Is the EU AI Act leaving a backdoor for emotion recognition?. *Biometric Update*. <https://www.biometricupdate.com/202402/is-the-eu-ai-act-leaving-a-backdoor-for-emotion-recognition> (Accessed on 10th of October 2025).
- Bota, P. J., Wang, C., Fred, A. L., & Da Silva, H. P. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE access*, 7: 140990-141020.
- Bourgeois, P., & Hess, U. (2008). The impact of social context on mimicry. *Biological psychology*, 77(3): 343-352.
- Boyd, D. (2008). Why youth (heart) social network sites: The role of networked publics in teenage social life. In Buckingham, D. (ed.), *Youth, Identity, and Digital Media*. The MIT Press.
- Brey, P. A. (2012). Anticipatory ethics for emerging technologies. *NanoEthics*, 6(1): 1-13.
- Briggs, J. L. (1971). *Never in anger: Portrait of an Eskimo family*. Harvard University Press.

- Bryant, M. (2025). Denmark issues first apology over forced contraception of Greenlandic women. *The Guardian*. <https://www.theguardian.com/world/2025/aug/27/denmark-pm-apologises-over-physical-and-psychological-harm-caused-by-iud-scandal> (Accessed on 26th of September 2025).
- Budner, P., Eirich, J., & Gloor, P. A. (2017). Making you happy makes me happy: Measuring individual mood with smartwatches. *arXiv preprint*. <https://arxiv.org/abs/1711.06134> (Accessed on 14th of August 2024).
- Burgess, A., Cappelen, H., & Plunkett, D. (2020). *Conceptual engineering and conceptual ethics*. Oxford University Press.
- Burgess, A. & Plunkett, D. (2013). Conceptual ethics I. *Philosophy Compass*, 8(12): 1091-1101.
- Burgess, M. (2024). Amazon-powered AI cameras used to detect emotions of unwitting UK train passengers. *Wired*. <https://www.wired.com/story/amazon-ai-cameras-emotions-uk-train-passengers/> (Accessed on 14th of August 2024).
- Cabitza, F., Campagner, A., & Basile, V. (2023). Towards a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 6860-6868
- Cabitza, F., Campagner, A., & Mattioli, M. (2022). The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data & Society*, 9(2).
- Cambridge Analytica (n.d.) Cambridge Analytica: Understanding the power behind modern data. <https://cambridgeanalytica.org/> (Accessed on 1st of December 2025).
- Candiotto, L. & Renaut, O. (eds.). (2020). *Emotions in Plato* (vol. 4). Brill.
- Cannon, W. B. (1914). The interrelations of emotions as suggested by recent physiological researches. *Journal of American Psychology*, 25: 256-282.
- Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.
- Capron, A. M. (1988). The report of the President's Commission on the Uniform Determination of Death ACT. In Zaner, R. M. (ed.), *Death: Beyond Whole-Brain Criteria* (pp. 147-170). Springer.
- Carlsen, H., Dreborg, K. H., Godman, M., Hansson, S. O., Johansson, L., et al. (2010). Assessing socially disruptive technological change. *Technology in Society*, 32(3): 209-218.
- Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Chen, J. X., Zhang, P. W., Mao, Z. J., Huang, Y. F., Jiang, D. M., et al. (2019). Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. *IEEE Access*, 7: 44317-44328.
- Chen, S., Jiang, K., Hu, H., Kuang, H., Yang, J., et al. (2021). Emotion recognition based on skin potential signals with a portable wireless device". *Sensors*, 21(3):1018.
- Chiffi, D., Moroni, S. & Zanetti, L. (2022) Types of technological innovation in the face of uncertainty. *Philosophy & Technology*, 35, 94.
- Clark, A. (1997). *Being there: Putting mind, body, and world together again*. MIT Press.
- Clark, A. (1999). An embodied cognitive science?. *Trends in cognitive sciences*, 3(9): 345-351.
- Clark, M. S., & Taraban, C. (1991). Reactions to and willingness to express emotion in communal and exchange relationships. *Journal of experimental social psychology*, 27(4): 324-336.
- Cognovi Labs. (n.d.) AI-powered emotional intelligence. <https://www.cognovilabs.com/> (Accessed on 27th of September 2025).
- Cohen, J. E. (2012). *Configuring the networked Self: Law, code, and the play of everyday practice*. Yale University Press.
- Cohen, J. E. (2013). What privacy is for. *Harvard Law Review*, 126(7): 1904-1933.
- Collier, M. (2011). Hume's science of emotions: Feeling theory without tears. *Hume Studies*, 37(1): 3-18.

- Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.
- Converus. (n.d.). CASE STUDY: University of Florida Police Dept. <https://converus.com/wp-content/uploads/2024/07/Case-Study-University-of-Florida-Police-Dept.pdf> (Accessed on 1st of December 2025).
- Cooper, J. M. (1999). *Reason and emotions*. Princeton University Press.
- Council of Europe (2019). Strategic Action Plan on Human Rights and Technologies in Biomedicine (2020-2025), adopted by the Committee on Bioethics (DH-BIO) in November 2019. <https://rm.coe.int/strategic-action-plan-final-e/1680a2c5d2>.
- Cornelius, R. R. (1996). *The science of emotion: Research and tradition in the psychology of emotions*. Prentice Hall.
- Crawford, K. (2021) *The atlas of AI: Power, politics and the planetary cost of artificial intelligence*. Yale University Press.
- Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., et al. (2019). AI Now 2019 Report. *AI Now Institute*. https://ainowinstitute.org/AI_Now_2019_Report.html
- Criado, N. & Such, J. M. (2015). Implicit contextual integrity in online social networks. *Information Sciences*, 325: 48-69.
- Christensen, C. M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail*. Harvard Business School Press.
- Cross, M. P., Acevedo, A. M., Leger, K. A., & Pressman, S. D. (2023). How and why could smiling influence physical health? A conceptual review. *Health psychology review*, 17(2): 321–343. <https://doi.org/10.1080/17437199.2022.2052740>
- Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., et al. (2024). Use of AI in mental health care: Community and mental health professionals survey. *JMIR mental health*, 11.
- Damasio, A. (1994). *Descartes' error: emotion, reason, and the human brain*. Avon Books.
- D'Arms, J., & Jacobson, D. (2006). Anthropocentric constraints on human value. *Oxford studies in metaethics*, 1: 99-126.
- Darwin, C. ([1872] 1965). *The expression of the emotions in man and animals*. University of Chicago Press.
- Davis, W. A. (1988). Expression of emotion. *American Philosophical Quarterly*, 25: 279-291.
- Day, M. (2019). Amazon is working on a device that can read human emotions. *Bloomberg*. <https://www.bloomberg.com/news/articles/2019-05-23/amazon-is-working-on-a-wearable-device-that-reads-human-emotions> (Accessed on 17th August 2024)
- DeCew, J. W. (1997). *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*. Cornell University Press.
- Devlin, H. (2025). Brain implant that could boost mood by using ultrasound to go under NHS trial. *The Guardian*. <https://www.theguardian.com/science/2025/jan/20/brain-implant-boost-mood-ultrasound-nhs-trial> (Accessed on 5th of December 2025).
- Discern Science Inc. (n.d.). Discern Science: Uncovering hidden deception. <https://discernscience.com/> (Accessed on 1st of December 2025).
- Dobbs v. Jackson Women's Health Org., 597 U.S. 215 (2022). <https://supreme.justia.com/cases/federal/us/597/19-1392/>
- Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research policy*, 11(3): 147-162.
- Dotson, K. (2011). Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*, 26(2): 236-257.
- Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.
- Dreyfus, H. (1992). *What computers still can't do: a critique of artificial reason*. MIT Press.

- Dudzik, B., Hrkalic, T. M., Hao, C., Raman, C., & Tsfasman, M. (2024). Indeterminacy in affective computing: Considering meaning and context in data collection practices. In *12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 181-185). IEEE.
- Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., et al. (2021). The rise of affectivism. *Nature Human Behaviour*, 5: 816-820.
- Durkheim, E. ([1897]1952). *Suicide. A study in sociology* (Spaulding, J. A. & Simpson, G. trans.). Routledge.
- Eibl-Eibesfeldt, I. (1989). *Human ethology*. Aldine de Gruyter.
- Eisenberg, N. (2020). Findings, issues, and new directions for research on emotion socialization. *Developmental psychology*, 56(3), 664.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6: 169-200.
- Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48: 384-392.
- Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115: 268-287.
- Ekman, P. & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1): 49-98.
- Ekman, P. & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2): 124.
- Ekman, P., Friesen, W.V. (1976) Measuring facial movement. *Journal of Nonverbal Behaviour*, 1: 56-75.
- Ekman, P., & Friesen, W. V. (1978). The Facial Action Coding System. In Ermiane, R. & Gergerian, E. (eds.), *Atlas of facial expressions/Album des expressions du visage*. La Pensée Universelle.
- Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. Pergamon.
- Ekman, P., Friesen, W. V. & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, 54: 414-420.
- Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, 77(5): 757-767.
- Elgin, C. Z. (2008). Emotion and understanding. In Brun, G., Doğuoğlu, U. & Kuenzle, D. (eds), *Epistemology and Emotions*. Ashgate Publishing Company.
- Elkins, A., Golob, E., Nunamaker, J., Burgoon, J., & Derrick, D. (2014). Appraising the AVATAR for automated border control: Results of a European Union field test of the AVATAR system for interviewing and passport control. *National Center for Border Security and Immigration*. European Parliament. [https://www.europarl.europa.eu/RegData/questions/reponses_qe/2019/002653/P9_RE\(2019\)002653\(ANN3\)_XL.pdf](https://www.europarl.europa.eu/RegData/questions/reponses_qe/2019/002653/P9_RE(2019)002653(ANN3)_XL.pdf)
- Ellis, B. J., & Malamuth, N. M. (2000). Love and anger in romantic relationships: A discrete systems model. *Journal of Personality*, 68(3): 525-556.
- Ellsworth, P. C. & Scherer, K. R. (2003). Appraisal processes in emotion. In Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (eds.), *Handbook of affective sciences* (pp. 572-595). Oxford University Press.
- Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in psychology*, 14, 1199058.
- Empatica. (n.d.). Empatica: Advancing medical research. <https://www.empatica.com> (Accessed on 25th of September 2025).
- Emteq Labs. (n.d.). Emteq labs: Measure what matters. <https://emteqlabs.com> (Accessed on 25th of September 2025).

- Engstrom, S. (2010). Reason, desire, and the will. In Denis, L. (ed.), *Kant's metaphysics of morals: A critical guide* (pp. 28-50). Cambridge University Press.
- Erickson, K. & Schulkin, J. (2003). Facial expressions of emotion: A cognitive neuroscience perspective. *Brain and Cognition*, 52(1): 52-60.
- Escobar, M. C. (2019). Kairos launches camera featuring facial recognition and emotion detection. *Hospitality Technology*. <https://hospitalitytech.com/kairos-launches-camera-featuring-facial-recognition-and-emotion-detection> (Accessed on 15th of August 2024).
- Etzioni, A. (1999). *The limits of privacy*. Basic Books.
- Evans, D. (2001). *Emotion: The science of sentiment*. Oxford University Press.
- Faden R, Beauchamp T. (1986). *A history and theory of informed consent*. Oxford University Press.
- Fehr, B. & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113(3): 464-486.
- Find Solution AI. (n.d.) *Find Solution AI*. <https://www.findsolutionai.com/> (Accessed on 27th of May 2025).
- Fischer, A. H. (ed.). (2000). *Gender and emotion: Social psychological perspectives*. Cambridge University Press.
- Fischer, A. H., & Manstead, A. S. (2016). Social functions of emotion and emotion regulation. In Lewis, M., Haviland-Jones, J. & Barrett, L. F. (eds.), *Handbook of emotions* (4th ed.) (pp. 424-439). Guilford.
- Fischer, A., Pauw L.S., & Manstead A. S. R. (2019). Emotion recognition as a social act: The role of the expresser-observer relationship in recognizing emotions. In Hess, U. & Hareli, S. (eds.), *The Social Nature of Emotion Expression*. Springer Nature.
- Fischer, A. H., & Roseman, I. J. (2007). Beat them or ban them: the characteristics and social functions of anger and contempt. *Journal of personality and social psychology*, 93(1), 103.
- Fisher Phillips (2024). AI programs in Japan are forcing workers to smile more – would that work in the U.S.?. *Fisher Phillips Insights*. <https://www.fisherphillips.com/en/news-insights/ai-programs-japan-forcing-workers-smile-would-work-in-us.html> (Accessed on 13th of July 2025).
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2).
- Floridi, L. (2011). *The philosophy of information*. Oxford University Press.
- Frank, R. H. (1993). *Passions within reason: The strategic role of the emotions*. W. W. Norton & Company.
- Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, 60(2), 229–240.
- Fridlund, A. J. (1994). *Human Facial Expressions: An evolutionary view*. Academic Press.
- Fried, C. (1968). Privacy. *The Yale Law Journal*, 77(3): 475-493.
- Frijda, N. H. (1986). *The emotions*. Cambridge University Press.
- Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43(5): 349-358.
- Gage, N. M. & Baars, B. (2018). *Fundamentals of cognitive neuroscience: A beginner's guide*. Academic Press.
- Garcia, R., & Calantone, R. (2002). A critical look at technological innovation typology and innovativeness terminology: a literature review. *Journal of Product Innovation Management: An international publication of the product development & management association*, 19(2): 110-132.

- García-García, J. M., Penichet, V. M., Lozano, M. D., & Fernando, A. (2022). Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions. *Universal Access in the Information Society*, 21(4): 809-825.
- García-Hernández, R. A., Luna-García, H., Celaya-Padilla, J. M., García-Hernández, A., Reveles-Gómez, L. C., et al. (2024). A systematic literature review of modalities, trends, and limitations in emotion recognition, affective computing, and sentiment analysis. *Applied Sciences*, 14(16), 7165.
- Gavison, R. E. (1980). Privacy and the limits of Law. *The Yale Law Journal*, 89(3): 421-471.
- Gerbaudo-González, N., Condorí-Fernández, N., Catalá, A., & Gandoy-Crego, M. (2025) AI-based emotion recognition in dementia through facial expression: A scoping review. *Ageing Research Reviews*, 114.
- Gilbert, D. T. (2006). *Stumbling on happiness*. Knopf.
- Gillum, J. & Kao, J. (2019). The aggression detector: The unproven, invasive surveillance technology schools are using to monitor students. *ProPublica*. <https://features.propublica.org/aggression-detector/the-unproven-invasive-surveillance-technology-schools-are-using-to-monitor-students/> (Accessed on 14th of August 2024).
- Glazer, T. (2017a). Looking angry and sounding sad: The perceptual analysis of emotional expression. *Synthese*, 194: 3619-3643.
- Glazer, T. (2017b). The semiotics of emotional expression. *Transactions of the Charles S. Peirce Society*, 53(2): 189-215.
- Glazer, T. (2019). Epistemic Violence and Emotional Misperception. *Hypatia*, 34(1): 59-75.
- Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.
- Goffman, E. (1967). *Interaction ritual: essays on face-to-face interaction*. Aldine.
- Goldie, P. (2000). *The emotions: A philosophical exploration*. Oxford University Press.
- Gonzaga, G. C., Keltner, D., Londahl, E. A., & Smith, M. D. (2001). Love and the commitment problem in romantic relations and friendship. *Journal of personality and social psychology*, 81(2), 247.
- Goold, B. (2008). The difference between lonely old ladies and CCTV cameras: A response to Ryberg. *Res Publica*, 14(1): 43-47.
- Grand View Research. (2023). Emotion detection & recognition market size (2024 – 2030). *Grand View Research*. <https://www.grandviewresearch.com/industry-analysis/emotion-detection-recognition-market-report> (Accessed on 16th of April 2025).
- Green, M. S. (2007). *Self-expression*. OUP.
- Green, O. H. (1992). *The emotions: A philosophical theory*. Kluwer Academic/Plenum Publishers.
- Greenemeier, L. (2012). Avatar officer installed at Arizona-Mexico border station. *Scientific American*. <https://www.scientificamerican.com/article/avatar-officer-installed-mexico-border/> (Accessed on 16th of April 2025).
- Greenspan, P. (1980). A case of mixed feelings: ambivalence and the logic of emotion. In Rorty A. O. (ed.) *Explaining emotions* (pp. 223-250). University of California Press.
- Greenspan, P. (1988). *Emotions and reasons: An enquiry into emotional justification*. Routledge.
- Gremsl, T., & Hödl, E. (2022). Emotional AI: Legal and ethical challenges. *Information Polity*, 27(2): 163-174.
- Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66(3): 377–388.
- Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories*. University of Chicago Press.
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3): 271-299.

- Gstrein, O. & Beaulieu, A. (2022). How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches. *Philosophy & Technology*, 35(3).
- Guo, R., Guo, H., Wang, L., Chen, M., Yang, D., et al. (2024). Development and application of emotion recognition technology – a systematic literature review. *BMC psychology*, 12(1), 95.
- Guyer, P. (2010). Moral Feelings in the Metaphysics of Morals. In Denis, L. (ed.), *Kant's Metaphysics of Morals: A Critical Guide* (130-151). Cambridge University Press.
- Hacklin, F., Raurich, V., & Marxt, C. (2004). How incremental innovation becomes disruptive: the case of technology convergence. In *2004 IEEE International Engineering Management* (pp. 32-36). IEEE.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4): 814-834.
- Hall, S. (2024). AI and the election. *The Link*. <https://magazine.cs.cmu.edu/ai-and-the-election> (Accessed on 16th of April 2025).
- Halsey, A. (2013). House member questions \$900 million TSA 'SPOT' screening program. *Washington Post*. https://www.washingtonpost.com/local/trafficandcommuting/house-member-questions-900-million-tsa-spot-screening-program/2013/11/14/ad194cfe-4d5c-11e3-be6b-d3d28122e6d4_story.html (Accessed on 14th of August 2024).
- Hamzah, A. (2024). Japan's SoftBank introduces AI that makes calls from angry customers sound 'calm'. *The Straits Times*. <https://www.straitstimes.com/asia/east-asia/japan-s-softbank-introduces-ai-that-makes-calls-from-angry-customers-sound-calm?ref=inline-article> (Accessed on 10th of August 2024).
- Hareli, S., & Hess, U. (2010). What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. *Cognition and emotion*, 24(1): 128-140.
- Harré, R. (ed.). (1986). *The social construction of emotions*. Blackwell.
- Harris, M. (2018). An eye-scanning lie detector is forging a dystopian future. *Wired*. <https://www.wired.com/story/eye-scanning-lie-detector-polygraph-forging-a-dystopian-future/> (Accessed on 1st of December 2025).
- Hart, H. L. A. (1955). Are there any natural rights?. *Philosophical Review*, 64(2): 175-191.
- Harwell, D. (2019). A face-scanning algorithm increasingly decides whether you deserve the job. *Washington Post*. <https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/> (Accessed on 14th of August 2024).
- Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be?. *Noûs*, 34(1): 31–55.
- Hassija, V., Chamola, V., Mahapatra, A. et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1): 45–74.
- Hassounieh, A., Mutawa, A. M., & Murugappan, M. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked*, 20, 100372.
- Häuselmann, A., Sears, A. M., Zard, L., & Fosch-Villaronga, E. (2023). EU law and emotion data. In *2023 11th international conference on affective computing and intelligent interaction (ACII)* (pp. 1-8). IEEE.
- Heaven, D. (2020). Why faces don't always tell the truth about feelings. *Nature*, 578(7796): 502-505.
- Helm, B. W. (2009). Emotions as evaluative feelings. *Emotion Review*, 1(3): 248-255.
- Henderson, R. M. (2021). Innovation in the 21st century: Architectural change, purpose, and the challenges of our time. *Management Science*, 67(9): 5479-5488.
- Henderson, R. M. & Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35(1): 9-30.

- Herman, B. (1993). *The practice of moral judgement*. Harvard University Press.
- Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and social psychology review*, 17(2): 142-157.
- Hess, U. & Hareli, S. (2015). The influence of context on emotion recognition in humans. In *11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, 3(1): 1-6.
- Hess, U. & Hareli, S. (2019a). *The social nature of emotional expression: What emotions can tell us about the world*. Springer Nature Switzerland.
- Hess, U. & Hareli, S. (2019b). The emotion-based inferences in context (EBIC) model. In Hess, U. & Hareli, S. (eds.) *The social nature of emotion expression: What emotions can tell us about the world* (pp. 1-5). Springer International Publishing.
- Hess, U., Kafetsios, K., Mauersberger, H., Blaison, C., & Kessler, C. L. (2016). Signal and noise in the perception of facial emotion expressions: From labs to life. *Personality and Social Psychology Bulletin*.
- High-Level Expert Group on Artificial Intelligence (HLEG). (2019). *Ethics guidelines for trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (Accessed on 3rd of December 2025).
- Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences*, 16(2): 107-117.
- Hjortsjö, C. H. (1969). *Man's face and mimic language*. Lund University, Studentlitteratur.
- Hochschild, A.R. (1983). *The Managed Heart: commercialization of human feeling*. University of California.
- Hollanek, T. & Nowaczyk-Basińska, K. (2024). Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry. *Philosophy & Technology*, 37(2), 63.
- Honts, C. R. (2004). The psychophysiological detection of deception. In Granhag, P. & Strömwall, L. (eds.), *Detection of deception in forensic contexts* (pp. 103-123). Cambridge University Press.
- Honts, C. R., Thurber, S., & Handler, M. (2021). A comprehensive meta-analysis of the comparison question polygraph test, *Applied Cognitive Psychology*, 35: 411-427.
- Hopster, J. (2021). What are socially disruptive technologies? *Technology in Society*, 67, 101750.
- Hopster, J., & Löhr, G. (2023). Conceptual engineering and philosophy of technology: Amelioration or adaptation?. *Philosophy & Technology*, 36(4), 70.
- Hood, C. (2006). Transparency in historical perspective. In Heald, D. & Hood, C. (eds.), *Transparency, the Key to Better Governance?* (pp. 3-23). Oxford University Press.
- Hornyak, T. (2010). Robot companions to befriend sick kids at european hospital. *IEEE Spectrum*. <https://spectrum.ieee.org/robot-companions-to-befriend-sick-kids-at-european-hospital> (Accessed on 14th of August 2024).
- Houssein, E. H., Hammad, A., & Ali, A. A. (2022). Human emotion recognition from EEG-based brain-computer interface using machine learning: a comprehensive review. *Neural Computing and Applications*, 34(15): 12527-12557.
- Hume, D. ([1739] 2003). *A Treatise of Human Nature*. Project Gutenberg.
- Hursthouse, R. (1999). *On Virtue Ethics*. Oxford University Press.
- Ienca, M. (2021). On neurorights. *Frontiers in Human Neuroscience*, 15.
- Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life sciences, society and policy*, 13(1), 5.
- iBorderCtrl Project. (n.d.). iBorderCtrl Project. <https://www.iborderctrl.eu/> (Accessed on 27th of September 2025).
- iMotions. (n.d.). *iMotions: Powering Human Insights*. <https://imotions.com> (Accessed on 10th of March 2025).

- Information Commissioner's Office. (2022). *Biometrics: Insight*. Innovation & Technology.
- Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11).
- Ishiguro, K. (2021). *Klara and the Sun*. Knopf.
- Izard, C. E. (1969). The emotions and emotion constructs in personality and culture research. In Cattell, R. B. (ed.), *Handbook of modern personality theory*. Aldine Press.
- Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts.
- James, W. ([1884]/1948). What is emotion?. In Dennis, W. (ed.), *Readings in the history of psychology* (pp. 290–303). Appleton-Century-Crofts.
- Javelosa, J. (2017). New lie-detecting robot security agent could help secure borders. *Futurism*. <https://futurism.com/new-lie-detecting-robot-security-agent-could-help-secure-borders> (Accessed on 14th of August 2024).
- Jiang, Y., W. Li, M.S. Hossain (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*. 53: 209-221.
- Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., et al. (2020). Dfaw: A large-scale database for recognizing dynamic facial expressions in the wild. *Proceedings of the 28th ACM international conference on multimedia* (pp. 2881-2889).
- Johnson-Glenberg, M. C. (2018). Immersive VR and education: Embodied design principles that include gesture and hand controls. *Frontiers in Robotics and AI*, 5.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2): 263-291.
- Kairos. (n.d.). *Kairos: Human analytics platform*. <https://www.kairos.com> (Accessed on 17th of June 2024).
- Kant, I. ([1785] 1998). Groundwork of the metaphysic of morals (Gregor, M. (ed. & trans.)). Cambridge University Press.
- Kasirzadeh, A. (2025). Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, 182: 1975–2003.
- Keltner, D., & Buswell, B. N. (1997). Embarrassment: its distinct form and appeasement functions. *Psychological bulletin*, 122(3), 250.
- Keltner, D. & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition & emotion*, 13(5): 505-521.
- Keltner, D., Haidt, J., & Shiota, M. N. (2013). Social functionalism and the evolution of emotions. In Schaller, M., Simpson, J. A., & Kenrick, D. T. (eds.), *Evolution and social psychology* (pp. 115-142). Psychology Press.
- Keltner, D., & Lerner, J. S. (2010). Emotion. In Fiske, S. T., Gilbert, D. T., & Lindzey, G. (eds.), *Handbook of social psychology* (5th ed.) (pp. 317–352). John Wiley & Sons, Inc..
- Keltner, D., Oveis, C., Van Kleef, G. A., & Luo, Y. (2014). The science of emotion: Exploring the basics of emotional intelligence. *Current Directions in Psychological Science*, 23(1): 10-15.
- Kendrick, M. (2019). The border guards you can't win over with a smile. *BBC*. <https://www.bbc.com/future/article/20190416-the-ai-border-guards-you-cant-reason-with> (Accessed on 15th of August 2024).
- Kennedy, P. (2014). Artificial intelligence lie detector developed by imperial alumnus. *Imperial*. <https://www.imperial.ac.uk/news/144486/artificial-intelligence-detector-developed-imperial-alumnus/> (Accessed on 15th of August 2024).
- Kenny, A. (1963). *Action, emotion and will*. Routledge & Kegan Paul.
- Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information fusion*, 102.

- Kim, D. & Lee, B. (2023). Voice analysis of people with dementia: Towards to AI-based emotion detection for people with dementia". *Alzheimer's Dement*, 19.
- Kinner, V. L., Kuchinke, L., Dierolf, A. M., Merz, C. J., Otto, T., et al. (2017). What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes. *Psychophysiology*, 54(4): 508-518.
- Königs, P. (2022). Government surveillance, privacy, and legitimacy. *Philosophy & Technology*, 35(1), 8.
- Kosgaard, C. (1998). Introduction. In Gregor, M. (ed. & trans.) *Groundwork of the metaphysics of morals* (pp. vii-xxx). Cambridge University Press.
- Koops, B. J., Newell, B. C., Timan, T., Skorvanek, I., Chokrevski, T., et al. (2016). A typology of privacy. *U. Pa. J. Int'l L.*, 38(2): 483-575.
- Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in facial expression recognition: A Survey of methods, benchmarks, models, and datasets. *Information*, 15(3): 135.
- Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24): 8788-8790.
- Kramer, M., Simmonds, N., & Steiner H. (1998). *A debate over rights: philosophical enquiries*. Oxford University Press.
- Kreibig S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biological psychology*, 84(3), 394-421.
- Kriegel, U. (2014). Towards a new feeling theory of emotion. *European Journal of Philosophy*, 22(3): 420-442.
- Kudina, O., & Verbeek, P. P. (2019). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, 44(2): 291-314.
- Kumar, T. S., & Senthil, T. (2021). Construction of hybrid deep learning model for predicting children behavior based on their emotional reaction. *Journal of Information Technology*, 3(1): 29-43.
- Kundera, M. ([1978]/1987). *The book of laughter and forgetting* (trans. Heim, M. H.) Faber & Faber Limited.
- Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, 17: 22–26.
- Lange, C. G. (1885). *The emotions*. William & Wilkins.
- Larsen, J.T., Berntson, G.G., Poehlmann, K.M., Ito, T.A., Cacioppo, J.T. (2008). The psychophysiology of emotion. In Lewis, R., Haviland-Jones, J.M., & Feldman-Barrett, L. (eds.), *The Handbook of Emotions* (3rd ed.). Guilford.
- Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1): 3-32.
- Lavazza, A. (2018). Freedom of thought and mental integrity: The moral requirements for any future neurotechnology. *Frontiers in Neuroscience*, 12, 82.
- Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. *American Psychologist*, 37(9): 1019-1024.
- Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.
- Lazarus, R. S. & Lazarus, B. N. (1996). *Passion and reason: Making sense of our emotions*. Oxford University Press.
- Le Breton, D. (1998). Sociologie des émotions: Critique de la raison darwinienne: Le corps. Entre émotion et rationalité. *Recherches sociologiques*, 29(1): 37-54.
- Ledoux, J. (1996). *The emotional brain: The mysterious underpinnings of the emotional life*. Simon & Schuster Paperbacks.
- Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober second thoughts: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin*, (24): 563-574.

- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, 66: 799-823.
- Lever, A. (2008). Mrs. Aremac and the camera: A response to Ryberg. *Res Publica*, 14(1): 35-42.
- Lewis, E. (2022). Microsoft is removing emotion recognition features from its facial recognition tech. *NBC News*. <https://www.nbcnews.com/tech/tech-news/microsoft-removing-emotion-recognition-features-facial-recognition-tec-rcna35087> (Accessed on 15th of August 2024).
- Li, C., Xu, C., & Feng, Z. (2016). Analysis of physiological for emotion recognition with the IRS model. *Neurocomputing*, 178: 103-111.
- Li, J., & Washington, P. (2024). A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: Machine learning study. *JMIR AI*, 3(1).
- Liddell, H. G. & Scott, R. (1940). *Greek-English lexicon, with a revised supplement*. Clarendon press.
- Lighthart, S., Ienca, M., Meynen, G., Molnar-Gabor, F., Andorno, R., et al. (2023). Minding rights: Mapping ethical and legal foundations of 'neurorights.' *Cambridge Quarterly of Healthcare Ethics*, 32(4): 461-481.
- Lim, J. Z., Mountstephens, J., & Teo, J. (2020). Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors*, 20(8), 2384.
- Lippert-Rasmussen, K. (2017). Brain privacy, intimacy, and authenticity: Why a complete lack of the former might undermine neither of the latter!. *Res Publica*, 23(2): 227-244.
- Lively, K. & Weed, E. A. (2016). The sociology of emotions. In Barrett, L., Lewis, M., & Haviland-Jones, J. M. (eds.), *The Handbook of Emotions* (4th ed.). Guilford.
- Lorenz, H. (2006). *The brute within: appetitive desire in Plato and Aristotle*. Oxford University Press.
- Lundgren, B. (2020). A Dilemma for Privacy as Control. *The Journal of Ethics*, 24: 165–175.
- Lundgren, B. (2021). How we can make sense of control-based intuitions for limited access-conceptions of the right to privacy. *Journal of Ethics and Social Philosophy*, 20(3): 382-391.
- Lutz, C. A. (1988). *Unnatural emotions: Everyday sentiments on a Micronesian atoll & their challenge to Western theory*. University of Chicago Press.
- LuxAI. (n.d.). QTRobot: Expressive humanoid robots. <https://luxai.com/> (Accessed on 23rd of November 2025).
- Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist Journal*, 29(10): 725-739.
- Macarol, J. (2025). Honda 0 series – a new era of emotional cars. *City Magazine*. <https://citymagazine.si/en/honda-0-series-new-era-of-emotional-cars/> (Accessed on 14th of April 2025).
- MacAskill, E. (2015). British and Dutch researchers develop new form of lie-detector test. *The Guardian*. <https://www.theguardian.com/science/2015/jan/04/british-dutch-researchers-new-form-lie-detector-test-polygraph> (Accessed on 14th of August 2024).
- Mackenzie, C. & Stoljar, N. (eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.
- Mackie, J. L. (1980). *Hume's moral theory*. Routledge.
- Macnish, K. (2018). Government surveillance and why defining privacy matters in a post-Snowden world. *Journal of Applied Philosophy*, 35(2): 417-432.
- Magee, P., Ienca, M., & Farahany, N. (2024). Beyond neural data: Cognitive biometrics and mental privacy. *Neuron*, 112(18): 3017-3028.
- Mainz, J. T., & Uhrenfeldt, R. (2021). Too much info: Data surveillance and reasons to favor the control account of the right to privacy. *Res Publica*, 27(2): 287-302.

- Mantello, P. et al. (2023). Bosses without a heart: socio-demographic and cross-cultural determinants of attitude towards Emotional AI in the workplace. *AI & Society*, 38: 97-119.
- Marda, V. & Jakubowska, E. (2023). Emotion (mis)recognition: is the EU missing the point?. *Article 19*, <https://www.article19.org/resources/eu-emotion-misrecognition/> (Accessed on 23rd of November 2024).
- Martinez, H. P., Yannakakis, G. N., & Hallam, J. (2014). Don't classify ratings of affect; rank them!. *IEEE transactions on affective computing*, 5(3): 314-326.
- Marston, W. M. (1917). Systolic blood pressure symptoms of deception. *Journal of Experimental Psychology*, 2(2), 117.
- Mascellino, A. (2022). China develops face biometrics, AI tool to test Party member loyalty. *Biometric update*. <https://www.biometricupdate.com/202207/china-develops-face-biometrics-ai-tool-to-test-party-member-loyalty> (Accessed on 22nd of November 2025).
- Maschke, G. W., & Scalabrini, G. J. (2000). *The lie behind the lie detector*. AntiPolygraph.org.
- Mattioli, M. & Cabitza, F. (2024). Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. *Machine Learning & Knowledge Extraction*, 6(4): 2201-2231.
- Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and emotion*, 14(3): 195-214.
- Matsumoto, D. (1999). American-Japanese cultural differences in judgements of expression intensity and subjective experience. *Cognition & Emotion*, 13(2): 201-218.
- Matsumoto, D., Yoo, S. H., & Fontaine, J. (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of cross-cultural psychology*, 39(1): 55-74.
- Matsumoto, D., & Kupperbusch, C. (2001). Idiocentric and allocentric differences in emotional expression, experience, and the coherence between expression and experience. *Asian Journal of Social Psychology*, 4(2): 113-131.
- Mauss, I.B. & Robinson, M.D. (2009). Measures of emotion: a review. *Cognition & Emotion*, 23(2): 209-237.
- McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016). AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems* (pp. 3723-3726).
- Mcintosh, J., M. Rodgers, B. Marques, and A. Cadle (2019). The use of VR for creating therapeutic environments for the health and wellbeing of military personnel, their families and their communities. *Journal of Digital Landscape Architecture*, 4: 185-194.
- McMullan, T. (2016). How a robot could be grandma's new carer. *The Guardian*. <https://www.theguardian.com/technology/2016/nov/06/robot-could-be-grandmas-new-care-assistant> (Accessed on 14th of August 2024).
- McStay, A. (2016). Empathic media and advertising: Industry, policy, legal and citizen perspectives (the case for intimacy). *Big Data & Society*, 3(2).
- McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage Publications.
- McStay, A. (2020). Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society*, 7(1).
- McStay, A., & Urquhart, L. (2019). "This time with feeling?": Assessing EU data governance implications of out of home appraisal based emotional AI. *First Monday*, 24(10).
- Meijer, E. H., & von Koppen, P. J. (2008). Lie detectors and the law: The use of polygraph in Europe. In Canter, D. & Zukauskienė, R. (eds.), *Psychology and the law: Bridging the gap* (pp. 31-50). Taylor & Francis.
- Menges, L. (2021). A defense of privacy as control. *The Journal of Ethics*, 25(3): 385-402.
- Menges, L. (2022). Three control views on privacy. *Social Theory and Practice*, 48(4), 691-711.

- Menges, L., & Weber-Guskar, E. (2025). Digital emotion detection, privacy, and the law. *Philosophy & technology*, 38(2), 77.
- Merton, R. (1964). *Social Theory and Social Structure*. Free Press of Glencoe.
- Metz, C. (2016). Apple buys AI startup that reads emotions in faces”, *Wired*. <https://www.wired.com/2016/01/apple-buys-ai-startup-that-reads-emotions-in-faces/> (Accessed on 18th of July 2023).
- Metzger J., and Kirchengast, T. (2025). Why a US court allowed a dead man to deliver his own victim impact statement – via an AI avatar. *The Conversation*. <https://theconversation.com/why-a-us-court-allowed-a-dead-man-to-deliver-his-own-victim-impact-statement-via-an-ai-avatar-259045> (Accessed on 13th of July 2025).
- Miles, N. C. (2024). Are you 80% angry and 2% sad? Why ‘emotional AI’ is fraught with problems. *The Guardian*. <https://www.theguardian.com/technology/article/2024/jun/23/emotional-artificial-intelligence-chatgpt-4o-hume-algorithmic-bias> (Accessed on 18th of August 2024).
- Mill, J. S. ([1859] 1985). *On liberty*. Ticknor and Fields.
- Millikan, R. G. (2000). *On clear and confused ideas: An essay about substance concepts*. Cambridge University Press.
- Millikan, R. G. (2004). *Varieties of Meaning: The 2002 Jean Nicod Lectures*. Bradford Books.
- MindShift CBT. (n.d.). Take control of your anxiety with MindShift CBT. <https://mindshiftcbt.com/> (Accessed on 13th of April 2025).
- Minsky, M. (1986). *Society of minds*. Simon & Schuster Paperbacks.
- Montag, C., & Panksepp, J. (2016). Primal emotional-affective expressive foundations of human facial expression. *Motivation & Emotion*, 40(5): 760-766.
- Mokrosinska, D. (2015). How much privacy for public officials? In Roessler, B. & Mokrosinska, D. (eds.) *Social dimensions of privacy: interdisciplinary perspectives* (pp. 181-201). Cambridge University Press.
- Mokrosinska, Dorota (2018). Privacy and autonomy: On some misconceptions concerning the political dimensions of privacy. *Law and Philosophy*, 37(2):117-143.
- Moore, A. D. (2003). Privacy: Its Meaning and Value. *American Philosophical Quarterly*, 40(3): 215-227.
- Moore, A. D. (2010). *Privacy rights: Moral and legal foundations*. Penn State Press.
- Merton, R. (1964). *Social Theory and Social Structure*. CFree Press of Glencoe.
- Murphy, R. F. (1964). Social Distance and the Veil. *American Anthropologist* 66: 1257–74.
- Naar, H. (2021) The fittingness of emotions. *Synthese*, 199: 13601–13619.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4): 435-450.
- Nagel, T. (1979). *Mortal questions*. Cambridge University Press.
- Nagel, T. (1998). Concealment and exposure. *Philosophy & public affairs*, 27(1): 3-30.
- Needham, C., Mastracci, S., & Mangan, C. (2017). The emotional labour of boundary spanning. *Journal of Integrated Care*, 25(4): 288-300.
- Neurologica. (n.d.). *Empowering AI with emotional insights*. <https://www.kopernica.ai/>
- Nickel, P. J. (2020). Disruptive Innovation and Moral Uncertainty. *NanoEthics*, 14(3): 259-269.
- Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review* 79, 119.
- Nissenbaum, H. (2010) *Privacy in context: Technology, policy and the integrity of social life*. Stanford Law Books.

- Nissenbaum, H. (2018). Respecting context to protect privacy: Why meaning matters. *Science and engineering ethics*, 24(3), 831-852.
- Noldus Hub. (n.d.). *Noldus Hub: Simplify your multimodal research*. <https://noldus.com/noldushub> (Accessed on 1st of December 2024).
- Nordmann, A. (2007). If and then: A critique of speculative nanoethics. *Nanoethics*, 1(1): 31-46.
- Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70: 1373-1411.
- Nussbaum, M. C. (2001). *Upheavals of thought: The intelligence of emotions*. Cambridge University Press.
- O'Hear, M. M. (1996). Remorse, cooperation, and acceptance of responsibility: the structure, implementation, and reform of Section 3E1. 1 of the Federal Sentencing Guidelines. *Nw. UL Rev.*, 91, 1507.
- O'Neill, E. (2022). Contextual integrity as a general conceptual tool for evaluating technological change. *Philosophy & Technology*, 35(3), 79.
- O'Neill, E. (2023). Balancing caution and the need for change: The general contextual integrity approach. *Philosophy & Technology*, 36(4), 68.
- O'Neill, O. Some limits of informed consent. *Journal of Medical Ethics*, 29(1): 4-7.
- Oatley, K. (1992). Integrative action of narrative. In Stein, D. J. & Young, J.E. (eds.), *Cognitive science and clinical disorders* (pp.151-170). Academic Press.
- Oatley, K. (2004). From the emotions of conversation to the passions of fiction. In Manstead, A. S. R., Frijda, N., & Fischer, A. (eds.), *Feelings and emotions: The Amsterdam symposium* (pp. 98–115). Cambridge University Press.
- Owen, R. J., Bessant, J. R., & Heintz, M. (eds.). (2013). *Responsible innovation: Managing the Responsible Emergence of Science and Innovation in Society* (Vol. 104). Wiley.
- Paiva-Silva, A. I., Pontes, M.K., Aguiar, J. S. R., de Souza, W. C. (2016). How do we evaluate facial emotion recognition?. *Psychology & Neuroscience*, 9(2): 153-175.
- Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561.
- Panksepp, J. (1998). *Affective neuroscience: The foundation of human and animal emotions*. Oxford University Press.
- Pantic, M. & Rothkrantz, L. J. (2000). Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11): 881-905.
- Parkinson, B. (1996). Emotions are social. *British Journal of Psychology*, 87(4): 663-683.
- Parkinson, B. & Colman, A. M. (1995). *Emotion and Motivation*. Longman.
- Parkinson, B., Fischer, A. H., & Manstead, A. S. R. (2005). *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology Press.
- Parrott, M. (2017). The Look of Another Mind. *Mind*, 126: 1023–61.
- Paul Ekman Group. (n.d.). *Facial Action Coding System*. <https://www.paulekman.com/facial-action-coding-system/> (Accessed on 15th of August 2024).
- Peal, T. R. (2004). The continuing sterilization of undesirables in America. *Rutgers Race & L. Rev.*, 6, 225.
- Pedota, M., Grilli, L., & Piscitello, L. (2021). Technological paradigms and the power of convergence. *Industrial and Corporate Change*, 30(6): 1633-1654.
- Picard, R. W. (1997). *Affective computing*. MIT Press.

- Piretti, L., Pappaiani, E., Garbin, C., Rumiati, R. I., Job, R., et al. (2023). The neural signatures of shame, embarrassment, and guilt: A voxel-based meta-analysis on functional neuroimaging studies. *Brain sciences*, 13(4): 559.
- Parliamentary Office of Science and Technology (2011). Detecting deception (POSTNOTE-375). https://www.parliament.uk/globalassets/documents/post/postpn375detecting_deception.pdf (Accessed on 18th of August 2024).
- Peirce, C. S. (1934). *Collected papers of Charles Sanders Peirce* (Vol. 5). Harvard University Press.
- Powell, M. (2025). Privacy: Concept, right, and underlying values. *Philosophy & Technology*, 38(3), 104.
- Price, A. W. (2009). Emotions in Plato and Aristotle. In Goldie, P. (ed.), *The Oxford Handbook of Philosophy of Emotion* (pp. 121–142). Oxford University Press.
- Price, T.F. and Harmon-Jones, E. (2015). Embodied emotion: the influence of manipulated facial and bodily states on emotive responses. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(6), 461-473.
- Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford University Press.
- Prégent, A. (2021). *Informatique affective: L'utilisation des systèmes de reconnaissance des émotions est-elle en cohérence avec la justice sociale?*. Mémoire de maîtrise. Presse de l'Université Laval.
- Prégent, A. (2025a). Is there not an obvious loophole in the AI act's ban on emotion recognition technologies?. *AI & Society*, 40, 5581–5582.
- Prégent, A. (2025b). Why you should not use CI to evaluate socially disruptive technology. *Philosophy & Technology*, 38(1), 6.
- Prégent, A. (Forthcoming a). Re-engineering Emotional Privacy in the face of emergent technologies: Forecasting the need for a normative proposal. In Isaac, M. G., Frank, L. E., & Hopster, J. (eds.), *Conceptual Engineering for Emerging Technologies*, Routledge.
- Prégent, A. (Forthcoming b). Technosocial risks of ideal emotion recognition technologies: A defense of the (social) value of emotional expressions. In Bojic, L. & Erić, Z. (eds.), *Ethics of AI Alignment: Rethinking Society Beyond the Human*. The International Library of Ethics, Law and Technology, Springer.
- Rachels, J. (1975). Why privacy is important. *Philosophy & public affairs*, 4: 323-333.
- Ratcliffe, M. (2005). *The feeling of being: Phenomenology, psychiatry and the sense of reality*. Oxford University Press.
- Raz, J. (1994). *Ethics in the public domain: essays in the morality of law and politics*. Oxford University Press.
- Regan, P. M. (1995). *Legislating privacy: Technology, social Values, and public policy*. University of North Carolina Press.
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).
- Reiman, J. H. (1976) Privacy, intimacy, and personhood. *Philosophy & Public Affairs*, 6(1): 26-44.
- Realbotix. (n.d.). Realbotix: Robots for human interaction. <https://www.realbotix.com/> (Accessed on 4th of April 2025).
- Rhue, L. A. (2018). Racial influence on automated perceptions of emotions. *Social Science Research Network*, 3281765.
- Ricœur, P. (1990). *Sois-même comme un autre*. Seuil.

- Riggins v. Nevada, 504 U.S. 127, 142 (1992). (Kennedy, J., concurring). Presumably, defendants are deprived of a fair trial by those things that interfere with a jury's perceiving remorse. In Barrett, L. F. (2017) *How emotions are made: The secret life of the brain*. Houghton Mifflin-Harcourt (p. 398).
- Rinne, I. (2024). Interpersonal scaffoldings for shared emotions: how social interaction supports emotional sharing. *Phenomenology & the Cognitive Sciences*, 1: 1-25.
- Robins-Early, N. (2024). New GPT-4o AI model is faster and free for all users, OpenAI announces. *The Guardian*. <https://www.theguardian.com/technology/article/2024/may/13/openai-new-chatgpt-free> (Accessed on 15th of August 2024).
- Rocksheng, Z., Baranoski, M., Feigenson, N., Davidson, L., Buchanan, A. et al. (2014). So you're sorry? The role of remorse in criminal law. *Journal of the American Academy of Psychiatry and the Law*, 42(1): 39-48.
- Roe v. Wade, 410 U.S. 113 (1973). <https://supreme.justia.com/cases/federal/us/410/113/>
- Roessler, B. (2005). *The value of privacy*. John Wiley & Sons.
- Roessler, B. & DeCew, J. (2023). Privacy. In Zalta E. N. & Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*.
- Roessler, B. & Mokrosinska, D. (2013). Privacy and social interaction. *Philosophy & Social Criticism*, 39(8): 771-791.
- Roessler, B. & Mokrosinska, D. (eds.). (2015). *Social dimensions of privacy: Interdisciplinary perspectives*. Cambridge University Press.
- Rosen, J. (2001). *The unwanted gaze: The destruction of privacy in America*. Vintage.
- Rowson, R. (2024). The dispositional account of emotional expression. *The Philosophical Quarterly*, 0(0): 1-25.
- Rubio-Licht N. (2024). Snap patent brings emotion detection to workplace surveillance. *The Daily Upside*. <https://www.thedailyupside.com/technology/social-media/snap-patent-brings-emotion-detection-to-workplace-surveillance/> (Accessed on 14th of July 2025).
- Rule, J. B. (2019). Contextual integrity and its discontents: A critique of Helen Nissenbaum's normative arguments. *Policy & Internet*, 11(3): 260-279.
- Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, 110(3): 426-450.
- Ryberg, J. (2007). Privacy rights, crime prevention, CCTV, and the life of Mrs Aremac. *Res Publica*, 13(2): 127-143.
- Ryberg, J. (2008). Moral rights and the problem of privacy in public: A reply to Lever and Goold. *Res Publica*, 14(1): 49-56.
- Ryberg, J. (2017). Neuroethics and brain privacy: Setting the stage. *Res Publica*, 23(2): 153-158.
- de Sant'Anna, M., Morat, B., & Rigaud, A. S. (2012). Adaptabilité du robot Paro dans la prise en charge de la maladie d'Alzheimer sévère de patients institutionnalisés. *NPG Neurologie-Psychiatrie-Gériatrie*, 12(67): 43-48.
- Sánchez-Monedero, J., & Dencik, L. (2022). The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl. *Information, Communication & Society*, 25(3): 413-430.
- Sandel, M. J. (1982). *Liberalism and the limits of justice*. Cambridge University Press.
- Scarantino, A. & Piccinini, G. (2010). Information without truth. *Metaphilosophy*, 41(3): 313-330.
- Scarantino, A. (2017). How to do things with emotional expressions: The theory of affective pragmatics. *Psychological Inquiry*, 28(2-3): 165-185.
- Scarantino, A. (2019). Affective pragmatics extended: From natural to overt expressions of emotions. In Hess, U. & Hareli, S. (eds.) *The social nature of emotion expression: What emotions can tell us about the world* (pp. 49-81). Springer.
- Scarantino, A. (2020) Exploring the roles of emotions in self-control. In Mele, A. R. (ed.) *Surrounding self-control* (chap. 7). Oxford University Press.

- Scarantino, A. & de Sousa, R. (2018). Emotion. In Zalta, E. N. (ed.), *The Stanford encyclopedia of philosophy* (Fall 2018 ed.). Stanford University.
- Schlenker, P. (2016). The semantics-pragmatics interface. In Aloni, M. & Dekker, P. (eds.), *Formal Semantics* (pp. 664-727). Cambridge University Press.
- Scheler, M. (1923). *The nature of sympathy* (Heath, P. trans.). Archon Books.
- Scherer, K. R. (1984). On the nature and function of emotion: a component process approach. In Scherer K.R. & Ekman P. (eds.), *Approaches to Emotion* (pp. 293-317). Lawrence Erlbaum Associates.
- Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In Scherer, K. R., Schorr, A., & Johnstone, T. (eds.), *Appraisal Processes in Emotion: Theory, Method, Research* (pp. 92-120). Oxford University Press.
- Scherer, K. R. (2005). What are emotions? And how can they be measured?. *Social science information*, 44: 695–729.
- Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, 23(7): 1307-1351.
- Schoeffer, J., De-Arteaga, M., & Elmer, J. (2025). Perils of label indeterminacy: A case study on prediction of neurological recovery after cardiac arrest. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1080-1094).
- Schoeman, F. (1984). Privacy: Philosophical dimensions. *American Philosophical Quarterly*, 21(3): 199–213.
- Schwartz, B. (1968). The social psychology of privacy. *American Journal of Sociology*, 73(6): 741-752.
- Schwartz, O. (2019). Don't look now: Why you should be worried about machines reading your emotions. *The Guardian*. <https://www.theguardian.com/technology/2019/mar/06/facial-recognition-software-emotional-science> (Accessed on 15th of August 2024).
- Searle, J. R. (1979). *Expression and meaning*. Cambridge University Press.
- Semin, G. R., & Manstead, A. S. (1982). The social implications of embarrassment displays and restitution behaviour. *European Journal of Social Psychology*, 12(4): 367-377.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379-423.
- Shen, Z., J. Cheng, X. Hu and Q. Dong (2019). Emotion recognition based on multi-view body gestures. *IEEE International Conference on Image Processing* (pp. 3317-3321).
- Shoumy, N. J., Ang, L-M., Seng, K. P. (2020). *Journal of Network and Computer Applications*, 149.
- Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., et al. (2018). A review of emotion recognition using physiological signals. *Sensors*, 18(7), 2074.
- Shuzhen Y., Androsoy, A., Yan, H., & Chen, Y. (2024). Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments. *Computers & Education*, 220.
- Silent Talker System. (2003). Intelligent lie-detector. <https://www.mmu.ac.uk/news-and-events/news/story/intelligent-lie-detector> (Accessed on 15th of January 2023).
- Simmel, G. ([1908]1950). *The Sociology of Georg Simmel* (Kurt H. Wolff, ed. & trans.). The Free Press.
- Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review*, 74(1): 29–39.
- Singer, T. & Klimecki, O. M. (2014). Empathy and compassion. *Current biology*, 24(18): 875-878.
- Smarteye. (n.d.). *Smarteye: Develop the advertising strategies precisely*. <https://www.smarteye.com/smarteye/> (Accessed on 29th of July 2025).
- Smith, C. A. & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4), 813.

- Smith, J. (2017). The perceptibility of emotion. In Naar, H. & Teroni, F. (eds), *The Ontology of Emotions* (1st ed.), (pp. 130–48). CUP.
- Snowdon, C. T. (2003). Expression of emotion in nonhuman animals. In Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (eds.), *Handbook of affective sciences* (pp. 457-480). Oxford University Press.
- Snowdon, P. F. (1992). How to interpret ‘direct perception’. In Crane, T. (ed.), *The Contents of Experience* (pp.48-78). Cambridge University Press.
- Softbank Robotics. (n.d.). *Softbank Robotics: Empowering the future with robotics and AI*. <https://www.softbankrobotics.com/> (Accessed on 30th of May 2025).
- Solomon, R. C. (1976). *The passions: The myth and nature of human emotion*. Anchor.
- Solomon, Robert C. (1977). The logic of emotion. *Noûs*, 11(1): 41-49.
- Solomon, R. C. (2003). *Not passion’s slaves: Emotions and choice*. Oxford University Press.
- Solomon, R. C. (2008). The philosophy of emotions. In Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. (eds.), *Handbook of Emotions* (3rd ed.) (pp. 3-16). Guilford.
- Solove, Daniel J., *Understanding Privacy*. Harvard University Press.
- Sound Intelligence. (n.d.). *Aggression Detection*. <https://www.soundintel.com/products/overview/aggression/> (Accessed on 1st of December 2025).
- de Sousa, R. (1987). *The Rationality of Emotion*. MIT Press.
- Standaert, M. (2021). Smile for the camera: the dark side of China’s emotion-recognition tech. *The Guardian*. <https://www.theguardian.com/global-development/2021/mar/03/china-positive-energy-emotion-surveillance-recognition-tech> (Accessed on 19th December 2024).
- Stanley, J. (2019). Bogus “aggression detectors” are audio-recording people in public. *ACLU*. <https://www.aclu.org/news/privacy-technology/bogus-aggression-detectors-are-audio-recording> (Accessed on 16th of April 2025).
- Stark, S. (2001). Virtue and emotion. *Noûs*, 35(3): 440-455.
- Stein, E. ([1917]1964) On the Problem of Empathy (trans. Stein, W.). *Springer Science & Business Media, BV*.
- Stöckli, S., Schulte-Mecklenbeck, M., Borer, S. et al. (2018). Facial expression analysis with AFFDEX and FACET: A validation study. *Behavioural Research*, 50, 1446-1460.
- Stote, K. (2015). *An act of genocide: Colonialism and the sterilization of Aboriginal women*. Fernwood Publishing.
- Stryker, S., & Burke, P. J. (2000). The past, present, and future of an identity theory. *Social psychology quarterly*, 284-297.
- Suhaimi, N. S., Mountstephens, J., & Teo, J. (2020). EEG-Based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational intelligence and neuroscience*.
- Sundby, S. E. (1997). Capital jury and absolution: the intersection of trial strategy remorse and the death penalty. *Cornell L Rev*, 83:1557.
- Susser, D., & Cabrera, L. Y. (2024). Brain data in context: Are new rights the way to mental and brain privacy?. *AJOB neuroscience*, 15(2): 122-133.
- Swierstra, T. & Rip, A. (2007). Nano-ethics as NEST-ethics: Patterns of moral argumentation about new and emerging science and technology.” *NanoEthics* 1 (1): 3-20.
- Tappolet, C. (2000). *Émotions et valeurs*. Presses Universitaires de France.
- Tappolet, C. (2017). *Emotions, Values, and Agency*. Oxford University Press.
- Tappolet, C. (2022). *Philosophy of emotion: A contemporary introduction*. Routledge.

- Tarantola, Andrew (2017) "Robot caregivers are saving the elderly from lives of loneliness", Engadget. (Accessed on 14th of August 2024)
- Taslitz, A. E. (2002). The Fourth Amendment in the twenty-first century: Technology, privacy, and human emotions. *Law & Contemporary Problems*, 65, 125.
- Tavani, H. T. & Moor, J. H. (2001). Privacy protection, control of information, and privacy-enhancing technologies. *ACM Sigcas Computers and Society*, 31(1): 6-11.
- Taylor, C. (1980). Theories of Meaning. *Man and World*, 13: 281-302.
- Taylor, C. (1992). *Multiculturalism and "the politics of recognition"*. Princeton University Press.
- Tcherkassof, A. (2018). *Le sens dessus dessous des expressions faciales des émotions: vers un nouveau tournant paradigmatique*. Psychologie. Université Grenoble Alpes.
- Tcherkassof, A. & Dupré, D. (2021). The emotion-facial expression link: evidence from human and automatic expression recognition, *Psychological research*, 85(8): 2954-2969.
- Teroni, F. (2007). Emotions and formal objects. *Dialectica*, 61(3): 395-415.
- Thaler, R. H. & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Thaler, R. H. & Sunstein, C. R. (2021). *Nudge: The final edition*. Penguin.
- Thomas, D. (2018). The cameras that know if you're happy - or a threat. *BBC*. <https://www.bbc.com/news/business-44799239> (Accessed on 17th of August 2024).
- Thomson, Judith Jarvis (1975). The right to privacy. *Philosophy and Public Affairs*, 4(4):295-314.
- Tomkins, S. S. (1962). *Affect, imagery, consciousness: The positive affects*. Springer.
- Tormey, A. (1971). *The concept of expression: A study in philosophical psychology and aesthetics*. Princeton University Press.
- Truog, R. D., & Miller, F. G. (2014). Changing the conversation about brain death. *The American Journal of Bioethics*, 14(8): 9-14.
- Truong, K. P., van Leeuwen, D. A., Neerincx, M. A. (2007). Unobtrusive multimodal emotion detection in adaptive interfaces: Speech and facial expressions. In Schmorow, D. D. & Reeves, L. M. (eds.), *Foundations of Augmented Cognition: Lecture Notes in Computer Science*. Springer.
- Udinmwen, E. (2025). This AI platform claims to 'understand' human emotions and sense stress and anxiety: here's what you need to know. *Techradar*. <https://www.techradar.com/pro/security/this-ai-platform-claims-to-understand-human-emotions-and-sense-stress-and-anxiety-heres-what-you-need-to-know> (Accessed on 8th of July 2025).
- Ugail, H., Yap, M. H., & Rajoub, B. (2007). Face reading technology for lie detection. *University of Bradford, School of Computing, Informatics, and Media*.
- United Nations General Assembly. (1948). *Universal Declaration of Human Rights*. United Nations. <https://www.un.org/en/about-us/universal-declaration-of-human-rights> (Accessed on 2nd of December 2025).
- Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
- van der Schalk, J., Kuppens, T., Bruder, M., & Manstead, A. S. (2015). The social power of regret: the effect of social appraisal and anticipated emotions on fair and unfair allocations in resource dilemmas. *Journal of Experimental Psychology: General*, 144(1), 151.
- Van Doorn, E. A., Heerdink, M. W., & Van Kleef, G. A. (2012). Emotion and the construal of social situations: Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment. *Cognition and Emotion*, 26(3): 442-461.
- Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science*, 18(3): 184-188.

- Van Kleef, G. A. (2016). *The interpersonal dynamics of emotion: Towards an integrative theory of emotions as social information*. Cambridge University Press.
- Van Kleef, G. A., & Côté, S. (2007). Expressing anger in conflict: when it helps and when it hurts. *Journal of Applied Psychology*, 92(6), 1557.
- Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004a). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, 86: 57–76.
- Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004b). The interpersonal effects of emotions in negotiations: A motivated information processing approach. *Journal of Personality and Social Psychology*, 87: 510–528.
- Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. In Zanna, M. P. (ed.), *Advances in experimental social psychology* (pp. 45-96). Academic Press.
- van de Poel, I. (2013). Why new technologies should be conceived as social experiments.” *Ethics, Policy & Environment*, 16 (3): 352-55.
- van de Poel, I. (2022). Socially disruptive technologies, contextual integrity, and conservatism about moral change. *Philosophy & technology*, 35(3): 82.
- van de Poel, I., Hermann, J., Hopster, J., Lenzi, D., Nyholm, S., et al. (2023). *Ethics of socially disruptive technologies: An introduction*. Open Book Publishers.
- Veale, M. & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4): 97-112.
- Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Towards a cognitive lie detection approach. *Current Directions in Psychological Science*, 20(1): 28-32.
- Wagner, H. (1997). Methods for the study of facial behavior. In Russell, J. A. & J. S. Fernández-Dols (eds.), *The Psychology of Facial Expression* (pp. 31-54). Cambridge University Press.
- Wakefield, J. (2021). AI emotion-detection software tested on Uyghurs. *BBC*. <https://www.bbc.com/news/technology-57101248> (Accessed on 14th of August 2024).
- Walzer, M. (1984). *Spheres of justice : a defense of pluralism and equality*. Basic books.
- Wang, Q., Wang, M., Yang, Y., & Zhang, X. (2022). Multi-modal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine*, 149.
- Wang, X., Shazeda, W., & Leufer, D. (2023). Bodily harms: how AI and biometrics curtail human rights bodily harms. *Access Now*. <https://www.accessnow.org/wp-content/uploads/2023/10/Bodily-harms-mapping-the-risks-of-emerging-biometric-tech.pdf> (Accessed on 2nd of February 2024).
- Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., et al. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, 83: 19-52.
- Warren, S. D. & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, 4(5): 193-220.
- Westin, A. (1967). *Privacy and Freedom*. Ig Publishing.
- Wieser, M. J., & Brosch, T. (2012). Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in psychology*, 3, 471.
- Weisman, R. (2014). *Showing remorse: Law and the social control of emotion*. Ashgate Press.
- Weizenbaum, J. (1976). *Computer power and human reason, from judgement to calculation*. WH. Freeman.
- Widacki, J. (2007). Polygraph examinations in Poland. *European Polygraph*, 1(1): 24-34.
- Widacki, J. (2020). John Augustus Larson (1892-1965). *European Polygraph*, 14(51): 9-10.

- Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press.
- Wittgenstein, L. (1953). *Philosophical investigations*. (trans. G. E. M. Anscombe). Blackwell.
- Wittgenstein, L. (1980). *Remarks on the philosophy of psychology*, Vol. I. Basil Blackwell.
- Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies*, 61(2): 341-355.
- Wilms, R., Lanwehr, R., & Kastenmüller, A. (2020). Emotion regulation in everyday life: The role of goals and situational factors. *Frontiers in Psychology*, 11, 877.
- Wilson, J. P. & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, 26: 1325-1331.
- Wistrich, A. J., Rachlinski, J. J., & Guthrie, C. (2015). Heart versus head: Do judges follow the law or follow their feelings. *Texas Law Review*, 93: 855-923.
- Wright, J. (2021). Suspect AI: Vibraimage, emotion recognition technology and algorithmic opacity. *Science, Technology and Society*, 28(3): 468-487.
- Xu, Y. W., & Fan, L. (2023). Emotional labor and job satisfaction among nurses: The mediating effect of nurse-patient relationship. *Frontiers in psychology*, 14, 1094358.
- Yin, Z., Zhao, M., Wang, Y., Yang, J., & Zhang, J. (2017). Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, 140: 93-110.
- Yoon, Y. C. (2022). Can we exploit all datasets? Multimodal emotion recognition using cross-modal translation. In *IEEE Access*, 10.
- Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmenta, J. M., et al. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, 551(7679): 159-163.
- Zaborowski, R. (2020). Revisiting mixed feelings. *Axiomathes*, 30: 201-226.
- Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2): 151-175.
- Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, 39(2): 117-123.
- Zebrowitz, L. A. (1997). *Reading faces: Window to the soul?*. Westview Press.
- van der Zee, S., Poppe, R., Taylor, P. J., & Anderson, R. (2019). To freeze or not to freeze: A culture-sensitive motion capture approach to detecting deceit. *PloS One*, 14(4): 1-18.
- Zeitchik, S. (2022). Former Google scientist says the computers that run our lives exploit us – and he has a way to stop them. *Washington Post*. <https://www.washingtonpost.com/technology/2022/01/17/artificial-intelligence-ai-empathy-emotions/> (Accessed on 17th of August 2024).
- Zhang, J., Yin, Z., Chen, P., Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. *Information Fusion*, 59: 103-126.
- Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., & Cichocki, A. (2018). Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49(3): 1110-1122.
- Zheng, W. L., Zhu, J. Y., & Lu, B. L. (2017). Identifying stable patterns over time for emotion recognition from EEG. *IEEE transactions on affective computing*, 10(3): 417-429.

English Summary

In recent years artificial intelligence has moved decisively into the domain of affect. Technological breakthroughs, together with commercial enthusiasm and political interest, have accelerated this expansion. Affective monitoring cameras are now present in public spaces. Emotion sensitive tutoring systems are being introduced in classrooms. Workplaces experiment with analytics that track employees' positivity (or lack thereof). And border control authorities explore systems that allegedly detect deception through micro-expressions. Although many of these systems are scientifically contested and face significant technical limitations, emotion recognition technologies (ERTs) are making impressive leaps forward, and this continuous progress raises an urgent philosophical question: what would happen to social life, to our expressive practices, and to our normative expectations if ERTs were to function as intended and reliably track our affective lives in real time?

This is the question guiding the dissertation. Rather than evaluating current (and flawed) systems, it adopts an anticipatory ethical approach anchored in real sociotechnical developments to map the potential disruptions that would follow from reliable, continuous, and ubiquitous affective legibility. The prospect of fully functional ERTs allows us to anticipate this world and the new kinds of harms that would arise from their development. After investigating the ways in which emotional expressions currently function in social interactions, the dissertation proposes and articulates a novel conception of emotional privacy, which has for purpose to provide an additional (normative) conceptual tool to better grasp the stakes at play in the deployment and use of ERTs for both individuals and societies.

The main body of the thesis is structured into three parts across five chapters. **Chapter 2** develops the first part by reconstructing the technosocial trajectory of ERTs. It follows their evolution from early face-based systems that relied on simplistic emotion models and static images to context-sensitive systems that integrate multimodal inputs to detect individuals' emotional states. In the process, the chapter identifies persistent technical challenges that currently significantly limit these technologies, even as ERTs continue to be deployed across many domains of social life, including retail, education, policing, border control, healthcare, human resources, self-tracking applications, and robotics. The chapter concludes with a discussion of regulatory initiatives, with a particular focus on the EU AI Act and its treatment

of ERTs. While these initiatives explain the pressing need to strictly regulate the use of current (and flawed) ERTs, they do not provide principled reasons to constrain future systems that might work as intended. This exposes a conceptual and normative gap that the rest of the dissertation is designed to address.

The second part of the thesis, developed in **Chapters 3** and **4**, turns to emotional life itself to clarify what ERTs aim to track and what might be at stake in making emotional life fully legible.

Chapter 3 focuses on emotions and adopts an integrative perspective that identifies their central functions in social and moral life. It revisits historical debates about the place of emotion in morality through the lens of recent reinterpretations, emphasising emotions' normative significance for moral perception, deliberation, and motivation. It then surveys biological and evolutionary, cognitivist, and social constructionist perspectives, drawing out their shared functional commitments. On this basis, the chapter characterises emotions as evaluative, dynamic, and socially embedded phenomena that orient attention, shape practical and moral reasoning, and support interpersonal understanding. Tracking one's emotional states thus also implies tracking one's values, priorities, and unfolding moral commitments.

Chapter 4 shifts the focus from emotions as intrapersonal states to emotional expressions as the primary medium through which affective life becomes socially legible. It positions them as integral components of the informational infrastructure of social life, following two claims: (1) nonverbal affective communication constitutes a central channel of everyday social interaction, and (2) emotional expressions play this role because they function as defeasible yet generally reliable sources of information on which emoters and perceivers rely to coordinate action, manage expectations, and regulate social relations. The chapter then turns to emotional expressions in technosocial contexts. It argues that emotional expressions exhibit a bifunctionality under conditions of technological mediation. For human perceivers, emotional expressions function as communicative instruments through which social coordination, negotiation, and norm regulation are achieved. For ERTs, by contrast, the same expressions are treated as diagnostic data points used to infer, classify, and predict inner affective states. The chapter sketches a revised design-based account which conceives emotional expressions as relatively stable patterns of behavioural and physiological change that are designed, biologically, culturally, or conventionally, to render affective and motivational states recognisable rather than merely perceptible. This framework preserves a descriptive and function-first account of

emotional expressions while acknowledging that human and technological ‘perceivers’ take distinct epistemic routes to construct their own interpretation. From these distinct functional orientations emerges the idea that technology-based recognition may collapse the socially constructed opacity that currently sustains social interactions, limiting human inferences.

The third part of the dissertation develops the core (normative) contribution: the concept of emotional privacy and its place within informational privacy theory. **Chapter 5** begins by reviewing the evolution of privacy theory from early individualistic accounts through relational and social accounts to contemporary informational approaches. It shows how privacy concepts have repeatedly been adapted in response to new technologies and uses the recent emergence of mental privacy in the context of neurotechnology as a precedent for the kind of tech-induced gap that ERTs seem to create in our (normative) conception of privacy. To fill this gap, the chapter builds on the first two parts of the dissertation and on recent work in privacy theory to introduce a novel account of emotional privacy, treated as a subconcept of informational privacy. This conception of emotional privacy is presented as a dual-level structure. Actual Emotional Privacy concerns particular situations in which an individual can manage how legible their emotional states are to different audiences. Standing Emotional Privacy, in contrast, concerns the ability to maintain expressive discretion across time and contexts. The thesis argues that it is this standing dimension that is crucial for understanding what is at stake with fully functional ERTs. The analysis then demonstrates that emotional privacy protects expressive discretion and that such discretion is a constitutive condition for self-presentation, for rule-bound norms of expressiveness, and for the regulative and cohesive functions of emotional expression.

The final chapter of the thesis asks whether emotional privacy should be protected by a moral right and what difference this would make for the ethical evaluation of ERTs. **Chapter 6** defends the view that emotional privacy warrants recognition as a distinct moral right, demonstrating that it cannot be simply reduced to Control-based or Access-based moral rights to informational privacy or fully captured by Contextual Integrity. The argument demonstrates that technology-based emotion recognition is not a simple extension of ordinary mindreading but a different kind of recognition. When faced with fully functional ERTs, individuals would not only lose Actual Emotional Privacy, but they would risk losing the ability to exercise expressive discretion altogether. The chapter distinguishes acceptable losses of emotional privacy from wrongful ones through the notion of defeat, which marks the point at which

Standing Emotional Privacy is lost. To make these stakes concrete, the thesis develops three technomoral scenarios set at the individual, interpersonal, and societal levels. Across these scenarios, fully functional ERTs undermine practices of self-presentation, reduce the opacity that is necessary for the negotiated management of social roles, and disrupt the social functions of expressive norms in public life. The harms identified arise not only from illegitimate data flows or the exposure of affective content, but from a structural reorganisation of the expressive and epistemic fabric of social life. The chapter concludes by embedding the proposed moral right to emotional privacy into established ethical assessment frameworks. It shows that recognising emotional privacy alters the weight assigned to privacy in certain trade-offs within regulatory decision-making and provides firmer grounds for limiting or conditioning the deployment of affective technologies.

Through the chapters, the thesis offers a framework for understanding and evaluating the disruptive transformations that fully functional affective technologies could bring. Whether fully functional ERTs ultimately emerge or not, the questions raised in this thesis speak to a broader challenge of our time: how to preserve the spaces in which persons can shape who they are, how they are seen, and how they relate to others within increasingly mediated worlds. Its central conclusion is that safeguarding those spaces is not only a matter of protecting information from malicious use but of protecting the social fabric in which meaning, agency, and shared life are formed.

Curriculum Vitae

Alexandra Prigent was born on December 14, 1996, in Longueuil, Canada. She completed her secondary education at Polyvalente Armand-Racicot (2014), and her Cégep (2016), both in Saint-Jean-Sur-Richelieu, Québec. From 2016 to 2019, she pursued her undergraduate studies at the Université de Montréal, and obtained her degree in Littératures de langue française et Philosophie. In 2021, under the supervision of Prof. dr. Jocelyn Maclure, she obtained her Master's degree in Philosophy at Université Laval. In December 2021, she began working on her PhD on the philosophical investigation of emotion recognition technology at Leiden University. In Spring 2025, during her PhD, she was a visiting research fellow at the London AI and Humanity Project (LAIHP), affiliated at the University of London. Her PhD was funded by the Social Sciences and Humanities Research Council of Canada (SSHRC), and led, jointly with the thesis itself, to two peer-reviewed publications at the time of the formal submission of her dissertation.