Emotional privacy: a philosophical investigation of emotion recognition technology

Prégent, A.P.

4. Emotional expressions in technosocial contexts: Rethinking functions and modes of recognition[75]

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 Prégent (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 06b9). 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, *in*dependent of dominant norms, values, or entrenched social causes that would have supported their development.

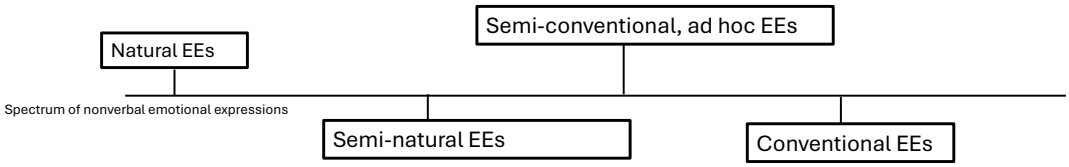
Finally, at the opposite end of the spectrum, there would be the "conventional EEs", which would be conventional nonverbal emotional expressions. They are also typically considered intentional and voluntary but contrary to semi-conventional EEs they are *de*pendent of dominant norms, values, or entrenched 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-targeted[86] 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.

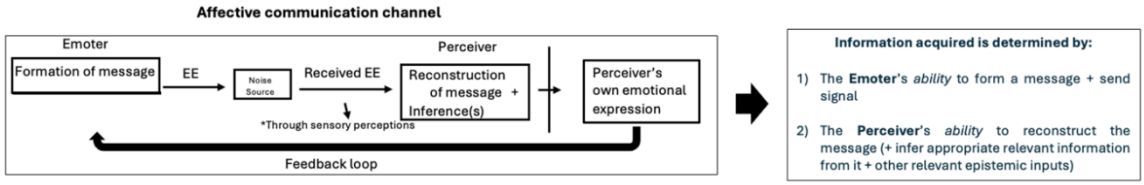
Here I move beyond the emoter and their EEs and enter the social realm of communication and interaction. To analyse EEs' communication, I adapt Shannon's (1948) model of information transmission, which originally focused on the encoding, transmission, and decoding

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* veridical[95] 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 them[97] 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.