PASSION

The Journal of the European Philosophical Society for the Study of Emotion

Volume: **01**
year: **2025**
Publisher: **Open Press TiU**
ISSN: **2773-1715**

03

Table of contents

Received: 13 August 2025
Accepted: 13 August 2025
Published: 28 August 2025

Introduction: Emotions—More Like Stars or Constellations?

Heidy Meriste and Bruno Mölder – University of Tartu, Estonia

1. Stars and Constellations

What kind of things are emotions? To use a compelling metaphor, we may ask: are emotions more like stars, or more like constellations? A rough-and-ready characterisation treats stars as a metaphor for natural kinds, unified by some objective feature in the world. Constellations, on the other hand, are not natural kinds; they result from the ways in which people group stars together. In the field of emotion research, the metaphor of stars versus constellations has been used by James A. Russell. He writes:

> Are discrete emotions like the stars, long recognized as fundamental astronomical entities, or are they more like the Big Dipper and other constellations? Constellations were once thought fundamental entities (indeed, powerful forces) but are now seen as merely happenstance configurations, seen from an arbitrary perspective and with no deep role to play in astronomy. Different cultures historically recognized somewhat different constellations. (2003, 152)

Russell's quote introduces an important additional consideration. Constellations, while once conceived of as causally efficacious entities that influenced the course of human lives, are no longer regarded as useful categories in genuine scientific research. If emotion categories—that is, our folk-psychological groupings of emotions—are like constellations, then they would have minimal value, if any, for the scientific study of emotions.

Of course, the metaphor of stars and constellations goes back to the philosopher Nelson Goodman, who also used the example of the Big Dipper (which, in fact, is not a constellation in itself, but a prominent asterism within the constellation Ursa Major). Goodman famously maintained that both constellations and stars are made by us, insofar as both involve conventional elements of categorisation and boundary-drawing. "We make a star as we make a constellation, by putting its parts together and marking off its boundaries" (1984, 42; see also McCormick 1996). Hilary Putnam (1992, 114), however, pointed out that there is a difference between stars and the Big Dipper. "Big Dipper" is a proper name that applies to a particular group of stars solely due to our naming conventions, whereas "star" is a general term whose application is not just a matter of linguistic convention. Even though we created the concept of a star, we do not thereby bring it about that any particular object counts as a star.

But since Putnam also highlighted that we likewise do not bring it about that a particular person counts as a bachelor, his point does not allow us to explain how stars differ from bachelors. When we want to talk about stars as natural kinds, the sense in which they do not depend on human conventions is a stronger one. While nature does not seem to suggest any clear-cut reason why we should distinguish bachelors as special units (just as it does not suggest a reason why the night sky should be divided into one set of constellations rather than another), it does seem to suggest a reason to distinguish stars as separate units. There is a special sense in which natural kinds are independent of human classificatory practices.

Exactly what sort of independence is at issue here? Sam Page (2006) makes a useful distinction between different types of dependence and independence, which are sometimes conflated in discussions about how reality depends on the mind. The first type is *ontological* dependence: entities that are ontologically dependent on us would no longer exist if we did not exist. Page's examples include money, speed limits, and other "social realities." The second type is *causal* dependence: these entities exist because they were caused by us. All human-made artifacts fall into this category. They are ontologically independent of us, since, once created, they continue to exist even in the absence of humans. The third type is *structural* dependence: structurally dependent entities are structured by us. Otherwise, they would be completely structureless blobs. Examples of structurally dependent entities are hard to find. In fact, Page concedes that, in the natural world, nothing might be structurally dependent on us, not even seemingly structureless phenomena like clouds or an aurora borealis. The fourth type of dependence, and the one relevant to the present context, is *individuative*: individuatively dependent entities are determined by how we delineate their boundaries. It is important to note that things can be individuatively dependent, yet structurally independent, of us. If something is individuated by our conventions, it does not follow that it is also structured by us. Page elucidates the notion of individuative independence with the familiar example of stars and constellations:

> We individuate the night sky into constellations. We, or more specifically our ancestors, determined which stars comprise which constellations. ... Though it is *prima facie* plausible that reality is individuated intrinsically into stars, reality is not individuated intrinsically into constellations, since it is people who divide the night sky into constellations. (2006, 328)

Page notes that constellations remain structurally independent of us. Apparently, this is because they are composed of stars, which are not contingent upon human categorisation. By contrast, views that espouse natural kinds (including stars) take them to be independent of us not only in the ontological, causal, and structural senses, but also in the individuative sense.

Having introduced the notion of individuation-dependence, we can now precisify the question posed in the title as follows: are emotions individuatively independent of, or dependent on, human classifications? Given that belonging to a natural kind is one way for something to be individuatively independent of our classifications, making progress on this question requires a closer look at natural kinds.

2. Natural Kinds

What is a natural kind? A natural kind is a group of things in nature that belong together by virtue of some objective set of properties. Usually, natural kinds are thought to have uniform members, and the boundaries between kinds are fairly discrete. This grouping is mind-independent in the sense that natural kinds are individuatively independent of us. It is important to distinguish between our classification of something as a natural kind and the kind that this classification purports to represent (cf. e.g. Mölder 2024, 145). While people sometimes loosely use "natural kind" to refer both to the category in nature and to our scientific classification that tracks it, only the kind in nature is a proper natural kind. That these are separate is evident from the fact that natural kinds can be used to make sense of scientific practice (Boyd 1991; Griffiths 1997; Samuels 2009). The aim of developing scientific classifications is for them to eventually match natural kinds. Furthermore, the existence of mind-independent natural kinds explains why our inductive generalisations work, and wherein the difference lies between discovery and invention.

There are two main approaches to natural kinds: one is quite demanding, the other sets more lenient requirements. *Essentialism* about natural kinds states that members belong to their kinds in virtue of their common essence (cf. e.g. Ellis 2002). This means that, necessarily, every member of the kind has to have certain essential properties in order to belong to that kind. These properties are intrinsic to the entities that make up the kind: they are possessed independently of anything else, including context. Ellis developed his essentialism in relation to chemical elements, and, within the field of chemistry, his claim that natural kinds must be categorically distinct is plausible. However, it is much more difficult to defend in the case of biological and psychological kinds.

A more relaxed view on natural kinds, the *homeostatic property cluster* (HPC) theory, was developed by Richard Boyd (1989; 1991). On this view, a natural kind consists of members that possess a characteristic cluster of properties. These property clusters are contingent, as the properties do not necessarily co-occur in virtue of any essence. The property clusters are sustained by homeostatic mechanisms that cause the properties to occur together. Such mechanisms regulate the kind by keeping specific properties together while filtering out others. The HPC account does not require strict essences that are both necessary and sufficient for something to count as a member of a kind. It allows variation among kind members and dispenses with the idea that natural kinds must have discrete boundaries. The paradigmatic examples of HPC kinds are biological species. Their members share similar characteristics because of causal homeostatic mechanisms that consist of adaptive forces such as descent from a common ancestor, gene flow, selection, and developmental canalisation (Griffiths 1997, 189).

The HPC account might be too lenient, however, as any kind that involves some common mechanisms, and supports better-than-chance predictions about its members, may qualify as a natural kind. Boyd himself argues for a very broad account of natural kinds that could also extend to different economic systems and philosophical positions (1999; see also Zachar 2022, 8). Arguably, this leads to an overproliferation of natural kinds, and loses the sense in which talk about natural kinds was originally supposed to carve *nature* at its joints. But even if the presence of just any causal homeostatic mechanism is not sufficient, it has nevertheless been widely accepted that, at least when it comes to the natural kind status of biological categories, causal homeostatic mechanisms play a crucial role. And insofar as the natural kind status of emotions rests on viewing them as biological traits, the broadness concerns of the HPC account need not shed doubt on talk of emotions as natural kinds.

The fact that natural kinds can be conceived of in different ways adds complexity to the question of whether emotions are natural kinds. It may well be that emotions qualify under one notion of natural kinds, but not under another. It is also possible that not all emotion types stand or fall together with respect to their natural kind status. It may be that only a select few kinds of emotions qualify as natural (Scarantino 2012a, 360). In fact, in this very issue, Charlie Kurth argues that, while shame is a natural kind, guilt is not.

Among the more biologically-minded researchers who are optimistic about the existence of natural kinds within the sphere of emotions, natural kinds are usually identified at the level of specific emotion types, rather than with the broad category of emotion as such. According to Paul Griffiths (1997; 2004), the broad, vernacular category of emotion is not a natural kind, as this category fragments into diverse types of states—namely, basic emotions, "higher cognitive" or complex emotions, and culturally influenced "socially sustained pretences." He argues that these types are too different from one another to form a single natural kind. In line with the HPC model, he notes that there is no common homeostatic mechanism present in these cases. Thus, there is no guarantee that generalisations made about one subcategory are also valid for others (1997, 242). But, in his view, it is also incorrect to identify the category of emotion with any one of its subcategories. This leaves open the possibility that the subcategories could latch on to natural kinds. Indeed, this is what Griffiths (230) has maintained about basic emotions—emotions that are supposed to be relatively universal across cultures and are usually taken to include, though not necessarily be exhausted by, (certain subsets of our vernacular categories of) fear, anger, sadness, disgust, and happiness (Ekman 1992; Ekman and Cordaro 2011).

Basic emotions are commonly understood as phenomenally salient and relatively short-lived patterns of characteristic physiological, cognitive and behavioural changes, which are undergirded by evolutionarily hardwired neural affect programmes (Tomkins 1962; Ekman 1992; see also Panksepp 1998 on emotion-specific neural circuits). They qualify as natural kinds according to the HPC conception (Griffiths 1997; Scarantino and Griffiths 2011; Scarantino 2012c; Kurth 2018). At the physiological level, it is the neural affect programme that functions as the relevant causal homeostatic mechanism that holds the emotion kind together. So, even though the exact components of the emotion-specific response profile may vary, they nevertheless tend to co-occur, because they are coordinated by the same underlying mechanism. At a general ecological level, however, the relevant causal homeostatic mechanism is the set of adaptive forces that have shaped each emotion type (Griffiths 1997, 238). This explains the origin of basic emotions and the relevant affect programmes themselves. They have developed as hardwired solutions to ancient fitness challenges and will also have homologues in other species in our lineage.

3. Constructions

Proponents of constructionism, however, argue that emotions are not natural kinds. Even though some of the constructionist critique is targeted against the more simplistic idea of emotions as essentialist natural kinds, much of their critique is meant to extend to emotions as HPC natural kinds as well, because constructionists argue that there is insufficient empirical evidence to support the existence of the pancultural affect programmes that basic emotion theorists posit (Barrett 2006; 2017; Barrett and Lida 2025).

Psychological constructionists in particular suggest that, instead of positing that the components of the emotion-specific response profile are coordinated by a single hardwired affect programme, the more plausible explanation is that it is simply an amalgam of many independently occurring, and more basic, psychological processes. As explained by Zachar, using the example of fear:

> The fearful expression, specifically the widening of the eyes that is part of it, may result from information gathering in time of uncertainty, yelling out may be an automatic reaction that was selected because it potentially warns others of danger, the running away may stem from the perception of danger that evoked a quick plan to save oneself, and the physiological arousal may be generated to support the execution of this plan. (2022, 4)

The core idea of psychological constructionism is that we construct emotions from a varying set of psychological components. This is also why the metaphor of constellations is especially apt: we make up emotions like we make up constellations from the stars. Though psychological constructionists agree that emotions consist of a plethora of elements, they highlight some as especially important. According to Russell (2003), the key role is played by "core affect"—an integral blend of feeling that varies with respect to the dimensions of valence (feeling good or bad) and activation (feeling energised or lethargic). Core affect is important because this is what makes the emotion "hot", or "emotional" (148). Barrett also requires that emotions involve a situated conceptualisation of core affect (Barrett 2017; Barrett and Lida 2025). According to her, our brain is always involved in predictive categorisation, preparing us for what comes next. This is an automatic background process that does not require any conscious reasoning or explicit labelling. Our brains just monitor the signals that we receive about the state of the world and the state of our bodies, and reassemble past experience to guide action and give meaning to what is going on. Each emotional episode, too, starts with such predictive categorisation. We create an emotion category on the fly that groups the current constellation of components together with a set of our past experiences, and this act of categorisation is itself a constitutive part of emotion. Since Russell does not require such situated conceptualisations, his and Barrett's views have been described as not only divergent but competing research programmes (Zachar 2022, 12).

There is also a more traditional strand of constructionism—social constructionism. The difference between psychological and social constructionism is sometimes described in terms of the former constructing emotions out of psychological ingredients, and the latter focusing on social and cultural ingredients (Barrett and Lida 2025, 352). This may make it seem that both theories suggest the same process of construction, simply out of different source materials, but this would be misleading. As noted in Zachar's contribution to this issue, psychological and social constructionist theories of emotion evoke the term "construction" in different senses. Psychological constructionism (as coined by Russell in his 2003) is a more recent development of constructionism, where the sense of construction is narrowed down to how we group an emotional episode together out of various psychological components. Social constructionists do not necessarily insist that we literally group emotions together out of social and cultural components. They are more interested in putting emotions into a wider context—showing how culture has shaped our emotions and how emotions are deeply embedded in social dynamics. They need not object to a psychological constructionist view of emotion, but they highlight that, at a more fundamental level, emotions are not just a matter of individual psychological processes, but a result of social dynamics (Mesquita and Parkinson 2025). This may leave the concept of construction somewhat broad and vague, though. But, in this issue, Charlie Kurth will seek to clarify at least one sense of what it means to culturally construct an emotion by providing us with a genealogical account of how cultures have developed guilt as a social technology.

Even though the nuances of different forms of constructionism vary, they all share the idea that emotions involve a configuration of elements that is not so much held together by some Darwinian modules like affect programmes, but by our own categorisations and sociocultural practices. In other words, they view emotions as human constructions rather than natural kinds.

4. Folk and Scientific Categories

However, one thing that is widely shared among both more biologically-minded and constructionist emotion theorists is that we should not expect folk-psychological categories to directly pick out scientific emotion kinds. Constructionists tend to be sceptical about the existence of natural kinds within the sphere of emotions in general. But even biological theorists who do argue for emotions as natural kinds do not think as if folk emotion categories required absolutely no refinement or revision for scientific purposes. As already mentioned earlier, basic emotions like fear or anger are only meant to pick out certain subsets of the relevant folk categories. The biological theorist Jaak Panksepp (2008, 402) even uses capital letters to refer to emotion-specific brain circuits, so that the difference between scientific technical terms and similar-sounding folk terms would be especially clear. All this raises the question: what relevance, if any, do folk-psychological emotion concepts have for emotion research?

Psychological constructionist James Russell has been especially critical about the import of folk emotion categories into scientific research. This takes us back to his metaphor of stars and constellations. Comparing emotions with the latter, he says:

> Stars in the heavens can be grouped into an uncountable number of different constellations. Doing so was useful in a preliminary way in astronomy and remains useful in navigation. Still, no grouping of stars into constellations proved useful in advancing astronomy, to the development of a deep scientific understanding, for the simple reason that constellations are not causal entities. (2008, 424–25)

Russell's point is that, just as constellations have outlived their usefulness for the science of astronomy, so too have vernacular emotion categories like "anger," "fear," "happiness," and even "emotion" itself outlived their usefulness for scientific research on emotions. He draws attention to the fact that basic emotion theory still rests on folk categories of emotion, the coherence of which does not withstand closer scrutiny. These categories are deeply rooted in Western culture and may be as arbitrary as constellations. Studying such categories only tells us something about people's beliefs about them, not about the real causal powers, which—for Russell as a constructionist—lie in the components from which emotional episodes are constructed. Following a distinction drawn by Bickle (2012), Zachar (2022, 8) characterises Russell's position as "little e" eliminativism about emotion. Let us first explain "big E" eliminativism, with which it is contrasted. "Big E" eliminativism, that is, eliminative materialism (Churchland 1981), makes an *ontological* point. On this view, folk psychology is a theory with ontological commitments, but because the folk theory of the mind is false, those commitments are not met, and folk categories (such as beliefs and desires) are to be replaced by those of future neuroscience. "Small e" eliminativism, in contrast, can allow that our folk psychological concepts are approximate characterisations of the underlying physiological reality. It just refers to eliminativism as a *methodological* principle. To the extent that emotions are constructions rather than natural kinds, they do not function as units that could figure in scientific explanations. Viewed from the natural kinds perspective, they are not caused by a common mechanism that would allow us to make projectable generalisations about them, and neither do they function as causally efficacious units that could themselves have any influence (even though their parts may well do so). As such, they are superfluous and distracting elements in the scientific discourse.

Andrea Scarantino (2012a), who is more enthusiastic about the search for natural kinds within the realm of emotions, however, has made a prominent proposal that we draw a sharp distinction between two projects with different goals: the descriptive "Folk Emotion Project," which aims to describe how people actually use folk-psychological categories like "emotion," "fear," and "anger," and the prescriptive "Scientific Emotion Project," which provides definitions of emotion categories as natural kinds. Scarantino (365) holds that these projects have different adequacy conditions. While the folk project should strive to accommodate all empirical findings related to phenomena classified in folk terms, the scientific project should aim to capture natural kinds (in the HPC sense). These two projects do not, and are not supposed to, converge on the same definitions of emotions. On his view, the natural kinds discovered by scientists are not folk-psychological kinds. This yields an important methodological constraint: one should not criticise scientific definitions of emotions on the basis of ordinary language use (Scarantino 2012b, 392).

Instead, what happens is that the folk categories are transformed into categories that are usable in scientific research. These transformed categories are given prescriptive definitions that should specify the causal mechanisms that constitute their natural kindhood, while also bearing some similarity to the original folk kind. In this way, the folk categories remain linked to the newly transformed scientific categories. However, Scarantino (2012b, 392) claims that the projected new natural kind categories are not coextensive with folk categories, so the old vernacular terms cannot be used for them. This leads to the question of how we should label these natural kind categories. One option is to use neologisms, but this may obscure similarities to the folk terms; another option is to modify the old term by using capitals, subscripts, or qualifiers, but this may lead to confusion with the original term (Scarantino 2012a, 366).

There is also an alternative outlook. Cecilea Mun (2016) has offered an interesting taxonomy of theories of emotion that draws attention to the fact that, in practice, Scarantino's two projects are intertwined. Her taxonomy can be seen as a criticism of Scarantino's proposal to separate the scientific search for natural kinds from the study of folk concepts. She distinguishes theories of emotion along the metaphysical axis, which concerns positions on what kind of kinds emotions are, and the metasemantic axis, which concerns the meaning of emotion terms as used in scientific and ordinary language.

On the metaphysical axis, emotions could be either objective kinds, unified by properties that do not depend on human concepts, or subjective kinds, whose unification necessarily depends on human concepts. (The account in Mun 2021 is more complex.) Mun (2016, 249–50) does not frame the metaphysical issue in terms of natural kinds, as she regards the notion as unhelpful. Namely, there are various conceptions of a natural kind, which yield different answers to the question of whether emotions are natural kinds. In addition, natural kinds were originally contrasted with social constructions, but on some views, things that counted as natural kinds could also be socially constructed.

On the metasemantic axis, accounts of emotion are positioned according to their stance on the scientific value of folk emotion terms. Mun (253) uses Putnam's notion of a "trans-theoretical term" to frame the issue. Trans-theoretical terms retain the same reference across different theories. Mun divides theorists of emotion into optimists and pessimists about ordinary language: the former regard folk emotion terms as trans-theoretical terms, while the latter believe that folk emotion terms are not trans-theoretical terms and, consequently, that folk and scientific discourse refer to different things when talking about emotions. Russell's and Scarantino's views mentioned above represent the pessimistic side of Mun's taxonomy.

Mun points out that the metaphysical and metasemantic axes are conceptually distinct from each other, and that positions on these dimensions can be combined. This results in a matrix of four kinds of positions, which differ in their views on the trans-theoretical status of folk emotion terms and on whether folk and scientific emotion terms refer to objective or subjective kinds (for more details on these positions, see Mun 2016, 257–60, or her revisitation of the taxonomy in this special issue).

As argued by Mun, Scarantino's proposal fails to take into account that there are also many emotion theorists on the optimistic side of her matrix (namely, realists such as Ekman, Goldie, and Scherer, and instrumentalists such as Averill, Solomon, and Nussbaum), who regard the study of folk emotions as "an integral part" of the scientific study of emotions (261). However, the force of the argument from Mun's taxonomy may be somewhat limited, insofar as her taxonomy aims to map the logical space of positions in the existing field of emotion research, and locates actual positions within this space, whereas Scarantino's proposal can be read as a revisionary one, claiming that this is how research on emotions should proceed, even if it is not currently conducted that way.

In this special issue, the optimistic stance is explored by Juan R. Loaiza, who focuses on building bridges between the Folk Emotion Project and the Scientific Emotion Project. Among other things, his contribution shows how the study of folk emotion concepts can benefit science even if not all folk concepts map well onto natural kinds. We are inevitably influenced by the categorisations in our language, but once we learn more about emotion concepts in different cultures, this may help to shake us out of old categorisations and inspire research in new and more promising directions. Viewed from this perspective, the very opposite of what eliminativists suggest might be true: to break the ties of prejudice, more rather than less research on folk emotion categories may be called for.

5. Overview of Contributions

What is it that our vernacular categories of emotion capture—do they roughly still map onto natural kinds, or do they capture individually and culturally variable constellations of biologically fragmented elements? And how much weight should we put on the study of folk emotion concepts at all? From one angle or another, all of these questions are addressed in the current special issue "Emotions—More Like Stars of Constellations?" The contributions originate from the EPSSE pre-conference workshop of the same title, organised by Heidy Meriste, Bruno Mölder, and Uku Tooming in Tartu in June 2023. In the remainder of this introduction, we offer brief overviews of the contributions.

The opening article is by Charlie Kurth, who goes against the common idea that emotions as a class are either natural kinds or social constructions. Using the example of guilt and shame, he argues that even emotions that may often be considered as belonging to the same family (the so-called self-conscious emotions) can diverge with respect to their status as natural kinds or constructions. Reviewing the empirical work on both emotions, he shows that while there is strong evidence in favour of shame being a natural kind (a biologically hard-wired adaptation), the case for guilt is relatively weak. Maintaining that guilt is better viewed as a social construction, Kurth goes on to add further detail to this idea by developing an account of guilt as "a type of emotional technology: a culturally-driven innovation that helped our ancestors address particular, recurrent challenges of social life." As such, he suggests that we view guilt as akin to other culturally-driven phenomena like promises, currencies, and running *amok*. This is an important contribution because the nature of non-basic, socially constructed emotions has remained relatively elusive (Griffiths 2004), and Kurth's account of guilt as a technology allows us to appreciate how guilt is both similar to, and different from, more basic emotions—while it has not come about as a hard-wired biological adaptation, we get a detailed account of how it has developed as a result of social pressures.

Peter Zachar's paper presents a nuanced perspective on understanding the constitution of emotions, drawing on the philosophy of Ernst Mach. According to psychological constructionist approaches, emotions do not exist independently of our ways of classifying them as expressions of mind-independent affect programmes, but are assembled from other psychological components (e.g. core affect, cognitive appraisal, categorisation). Zachar proposes that we should conceive of psychological construction in terms of the selection of features from a larger set. As a feature of measurement, the act of selecting partly constitutes the emotions studied by scientists. This selection also involves a conventionalist element, as it results from non-arbitrary but contingent processes and is not entirely determined by the facts. He points out that different selections of features can lead to different conceptualisations of emotions.

Zachar also comments on Russell's analogy between astronomical constellations and emotions. Russell holds that both are coincidental arrangements of components, whereas Zachar argues, following Scherer, that the components of an emotion need not be independently occurring events, but can enter into causal relationships with each other. Zachar points out that, with a more pluralistic model of causation, even constructed emotions qualify as kinds, since they can play a causal role and support generalisations. Zachar's empiricist-selectionist stance occupies a middle ground between Goodmanian worldmaking and passive perspective-taking, a position he describes as "engineering."

In her contribution, Cecilea Mun places the major theories of emotion within her metasemantic taxonomy. As already mentioned, in this taxonomy, theories are classified based on their stances toward the emotion words in ordinary language and their metaphysical views on whether emotions belong to objective or subjective kinds. This results in a fourfold structure, in which certain pairs of theories (e.g. realism and eliminativism, or instrumentalism and eliminative-realism) are contradictories and thus cannot both be true or false at the same time, whereas other pairs (e.g. realism and instrumentalism, or eliminativism and eliminative-realism) are contraries, and thus can both be false, but not true simultaneously. She treats these relationships as logical constraints on emotion research, and argues that this shows that not all theories of emotion can be unified or fully integrated.

The final paper, by Juan R. Loaiza, takes up the question of why and how we should investigate folk emotion concepts. Loaiza's article emphasises how the so-called Folk Emotion Project is also relevant for the Scientific Emotion Project. To the extent that scientists have usually opted to revise rather than altogether abandon folk emotion labels, it is important to grant that scientific emotion concepts share sufficient similarity with their folk psychological counterparts. Otherwise, they would not merit being called by the same name. This invites a clarification of the extension of folk emotion categories. But how should we go about this task? While Mun (2021) suggests that our shared starting point, the fundamental base of emotion science, should also include first-personal emotional experiences, Loaiza argues that the latter does not provide us with an intersubjectively accessible starting point that would be fit for anchoring scientific emotion concepts. Instead, we should stick to other sources of information. While those other sources have commonly been taken to include emotion attribution and recognition studies (Mun 2021), Loaiza also stresses the need to look into cross-cultural linguistics and research on emotion scripts and norms. This is not just important for justifying the current use of scientific emotion terms, but also has the potential to shape scientific research in new and fruitful directions: seeing how the sphere of emotions may be carved up in many alternative ways might also facilitate a more open-minded search for emotions as natural kinds.

References

Barrett, L. F. 2006. "Are Emotions Natural Kinds?" *Perspectives on Psychological Science* 1 (1): 28–58.

———. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Barrett, L. F. and T. Lida 2025. "Constructionist Theories of Emotions in Psychology and Neuroscience." In *Emotion Theory: The Routledge Comprehensive Guide, Volume I. History, Contemporary Theories, and Key Elements*, edited by A. Scarantino, 350–87. Routledge.

Bickle, J. 2012. "Lessons for Affective Science from a Metascience of 'Molecular and Cellular Cognition'." In *Categorical Versus Dimensional Models of Affect: A Seminar of the Theories of Panksepp and Russell*, edited by P. Zachar and R. D. Ellis, 175–87. John Benjamins.

Boyd, R. 1989. "What Realism Implies and What It Does Not." *Dialectica* 43 (1/2): 5–29.

———. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61 (1/2): 127–48.

———. 1999. "Homeostasis, Species, and Higher Taxa." In *Species: New Interdisciplinary Essays*, edited by R. A. Wilson, 141–85. The MIT Press.

Churchland, P. 1981. "Eliminative Materialism and Propositional Attitudes." *Journal of Philosophy* 78 (2): 67–90.

Ekman, P. 1992. "An Argument for Basic Emotions." *Cognition and Emotion* 6 (3–4): 169– 200.

Ekman, P. and D. Cordaro. 2011. "What Is Meant by Calling Emotions Basic." *Emotion Review* 3 (4): 364–70.

Ellis, B. 2002. *The Philosophy of Nature*. Acumen.

Goodman, N. 1984. *Of Mind and Other Matters*. Harvard University Press.

Griffiths, P. E. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press.

———. 2004. "Is Emotion a Natural Kind?" In *Thinking about Feeling: Contemporary Philosophers on Emotions*, edited by R. C. Solomon, 233–49. Oxford University Press.

Kurth, C. 2018. *The Anxious Mind: An Investigation into the Varieties and Virtues of Anxiety*. The MIT Press.

McCormick, P. J. (ed.) 1996. *Starmaking: Realism, Anti-Realism, and Irrealism*. The MIT Press.

Mesquita, B. and B. Parkinson 2025. "Social Constructionist Theories of Emotions." In *Emotion Theory: The Routledge Comprehensive Guide, Volume I. History, Contemporary Theories, and Key Elements*, edited by A. Scarantino, 350–87. Routledge.

Mölder, B. 2024. "Mental Kinds and Practical Realism." In *Practical Realist Philosophy of Science: Reflecting on Rein Vihalemm's Ideas*, edited by A. Mets, E. Lõhkivi, P. Müürsepp, and J. Eigi-Watkin, 143–59. Rowman & Littlefield.

Mun, C. 2016. "Natural Kinds, Social Constructions, and Ordinary Language: Clarifying the Crisis in the Science of Emotion." *Journal of Social Ontology* 2 (2): 247–69.

———. 2021. *Interdisciplinary Foundations for the Science of Emotion: Unification Without Consilience*. Palgrave Macmillan.

Page, S. 2006. "Mind-Independence Disambiguated: Separating the Meat from the Straw in the Realism/Anti-Realism Debate." *Ratio* 19 (3): 321–35.

Panksepp, J. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press

———. 2008. "Carving 'Natural' Emotions: 'Kindly' from Bottom-Up but Not Top-Down." *Journal of Theoretical and Philosophical Psychology* 28(2): 395–422.

Putnam, H. 1992. *Renewing Philosophy*. Harvard University Press.

Russell, J. A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110 (1): 145–72.

———. 2008. "In Defense of a Psychological Constructionist Account of Emotion: Reply to Zachar." *Journal of Theoretical and Philosophical Psychology* 28 (2): 423–29.

Samuels, R. 2009. "Delusion as a Natural Kind." In *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, edited by M. Broome and L. Bortolotti, 49–79. Oxford University Press.

Scarantino, A. 2012a. "How to Define Emotions Scientifically." *Emotion Review* 4 (4): 358–68.

———. 2012b. "Some Further Thoughts on Emotions and Natural Kinds." *Emotion Review*, 4 (4): 391–93.

———. 2012c. "Discrete Emotions: From Folk Psychology to Causal Mechanisms." In *Categorical Versus Dimensional Models of Affect*, edited by P. Zachar and R. D. Ellis, 135–54. John Benjamins.

Scarantino, A. and P. Griffiths 2011. “Don’t Give Up on Basic Emotions.” *Emotion Review*, 3 (4): 444–54.

Tomkins, S. S. 1962. *Affect, Imagery, Consciousness, Vol. 1. The Positive Affects.* Springer.

Zachar, P. 2022. “The Psychological Construction of Emotion—A Non-Essentialist Philosophy of Science.” *Emotion Review* 14 (1): 3–14.

Received: 19 February 2024
Accepted: 08 July 2025
Published: 28 August 2025

Emotion, Adaptation, and Natural Kinds: A Look at Shame and Guilt

Charlie Kurth – Clemson University, USA - ckurth@clemson.edu

Helsinki Collegium for Advanced Studies, Finland

Abstract

Much of the work on the evolutionary origins of human emotions views emotions as standing or falling together: either all of our emotions are natural kinds or none of them are. In this paper, I challenge the orthodoxy. Taking shame and guilt as case studies, I argue that while we have good reason to see shame as a biological adaptation, and so a kind, the case for guilt is much less impressive. But this conclusion raises an important question: if guilt isn't an adaptation, then what is it? In response, I argue that guilt might be best understood as a type of emotional technology: a culturally-driven innovation that helped our ancestors address particular, recurrent challenges of social life. The resulting picture not only offers a revisionary account of the nature and origins of shame and guilt, but also reshapes our thinking about whether emotions—as a class—are natural kinds.

Keywords: Emotion, Evolution, Natural kind, Emotional technology, Shame, Guilt

It's a perennial question in emotion theory: are emotions natural kinds or social constructions? On this, I say it depends. It depends because although the term "emotion" does not pick out a natural kind, some of the things that we call emotions are kinds. In what follows, I defend this claim by taking shame and guilt as case studies, looking at them through an evolutionary lens. As a pair of self-conscious emotions, shame and guilt are generally thought to be aligned in the sense that either both are adaptations or neither is (see e.g. Ramsey and Deem 2022; Frank 1988; James 2011; D'Arms and Jacobson 2023; Krebs 2011; Griffiths 1997, Prinz 2004; c.f., Ortony 1987, Elison 2005). Moreover, evidence that an emotion is an adaptation invites the further conclusion that it's also a kind, for being an adaptation is generally seen as the mark of kindhood in biology and psychology (Griffiths 1994, 1996; Kurth 2018). But while widely held, the idea that shame and guilt stand and fall together is mistaken. In what follows, I'll argue that while we have good reason to see shame as a biological adaptation—and so a kind—the parallel case for guilt is much less impressive. I'll also show how appreciating this more complicated picture of the origins of shame and guilt has revisionary implications for broader debates about whether emotions should be understood as natural kinds.

To make my case, I begin by specifying a set of criteria that tell for when an emotion is plausibly understood as an adaptation (§1). Drawing on research in philosophy as well as the social and cognitive sciences, I then use these criteria to defend my claim: shame is an adaptation, but guilt is not (§§2–3). Yet this conclusion leaves a big question unanswered: if guilt is not an adaptation, then why do we have the ability to feel it? Here I argue that guilt might be best understood as a type of emotional technology: a culturally-driven innovation that helped our ancestors address particular, recurrent challenges of social life (§4). Not only does this proposal enrich our understanding of the nature and origins of guilt, but it adds needed detail to the vague talk among emotion theorists about emotions being "social constructions." I then conclude by returning to the question of kinds, showing how these conclusions about shame and guilt should reshape our thinking.

1. Features that Tell for an Emotion Being an Adaptation

To say an emotion is an adaptation is to say there's some distinct, heritable trait that was selected for by Darwinian forces because that trait provided a fitness advantage to those who possessed it. What I will call the "telling features" ("TFs," for short) are things that, when found, provide evidence that a trait is an adaptation. The five TFs that I focus on should be familiar—they're found across a range of disciplines in discussions about the evolutionary origins of human emotions (e.g. Sznycer and Cohen 2021; Ekman 1999; Griffiths 1997; Scarantino and Griffiths 2011; Maibom 2010; Kurth 2016; 2018; Boehm 2012; Fessler 2004; 2007; Frank 1988). Importantly, the five TFs should not be understood as specifying a set of necessary and sufficient conditions on what it is for an emotion to be an adaptation. Rather, they're (defeasible) pieces of evidence. So, generally speaking, the more TFs a given emotion displays, the stronger the case that it's an adaptation. Additionally, while other TFs have been proposed (see, e.g., the discussion of "quick onset" in Ekman 1999), the five that I focus on dominate the literature. Since I take the points that follow to be familiar, I will keep my discussion brief.

> *TF1: Adaptive scenario*. If an emotion is an adaptation, then there should be an account of why it was selected for. Here we need more than a mere "just so" story. We need an account of both (i) the *adaptive challenge/opportunity* that our ancestors faced and (ii) the *distinctive functional role* of the emotion in question. But we also need (iii) evidence that an emotion with this functional role *could have helped* address the challenge/opportunity that was faced.
>
> *TF2: Proto-versions of the emotion*. If an emotion is an adaptation, then it's an adaptation of something else. So we should see evidence of that something else—precursors or proto-versions of the emotion—in our evolutionary ancestors.
>
> *TF3: Distinctive mechanisms*. To say that an emotion is an adaptation is to say that there's a heritable trait that was selected for. So we should see evidence of the mechanisms—neural, biochemical, social-psychological, etc.—that undergird that trait. In saying this, I'm not presuming that emotions (traits) should be identified with these mechanisms. Rather, my point is more modest: if we have a trait, then we have a characteristic set of features or behaviours. So there should be an identifiable mechanism (or set of mechanisms) that underwrites this patterning.[1]
>
> *TF4: Distinctive expressive routine*. Accounts of the distinctive functional roles that emotions play often point to their role as signals. So if signalling is part of an emotion's function, then we should see evidence of a corresponding, distinctive expressive routine. In saying this, I am not endorsing Ekman's famously controversial idea that the presence of a *unique* facial expression is *definitive* of an emotion

[1] Importantly for what's to come, in talking about mechanisms, I follow the philosophical mainstream in thinking that biopsychological traits are most plausibly understood on an anti-essentialist model (e.g. the homeostatic property clusters of Boyd 1999 or the stable property clusters of Slater 2015), rather than as things that have some unique underlying essence (for discussion, see Scarantino and Griffiths 2011; Kurth 2018, ch. 2).

type (1999). Rather, I'm making the more pedestrian claim that if an emotion has a signalling function, then there should be some characteristic signal—some distinctive pattern of vocalisation, body posture, or facial configuration—that tends to accompany tokens of that emotion. For instance, on the standard picture, anger signals aggression in a way that allows a dominant individual to preserve their standing without the risk of a costly fight. Hence the characteristic puffed-up body posture of anger. By contrast, while hunger comes with pain, that feeling is presumably part of an internal, motivational alarm, and not a signal to others of one's hunger. So here there's no expectation of a distinctive, observable expressive routine.

TF5: Heritability. If an emotion is an adaptation, then it has a genetic basis. So we should see some evidence of this heritability—e.g., that the emotion is pan-cultural or that it emerges very early in development.

With this account of the TFs in hand, we can ask whether they're found in shame and guilt. Moreover, while these emotions are familiar, the terms "shame" and "guilt" are often used interchangeably (at least among native English speakers). So an initial gloss on the two emotions will be helpful, even if it's one we'll refine in the discussion that follows.[2] At first pass, I'll understand shame as a painful experience that one has when one, for instance, fails to meet others' expectations, gives a clumsy public talk, harms another person (accidentally or on purpose), or looks a certain way. When ashamed, one feels that one is lower, degraded, or inadequate as a result of what one has done or who one is. These feelings also tend to be accompanied by distinctive motivations: in some situations, one may try to hide (oneself or the source of the shame); in others, one may try to make up for the harm done. Like shame, guilt is also a painful feeling that we experience in a similarly wide range of situations. We feel guilt if we've, e.g., hurt others (intentionally or not), failed to heed our diet or exercise regime, cheated on our partner, or performed poorly. We can also experience guilt in situations where we haven't done harm (e.g. being the only survivor of an accident). In contrast to shame, when we feel guilt, we're more likely to feel a sense of responsibility for the harm or bad outcome. Our attention shifts to the harm or damage that has occurred, and we tend to ruminate on how things could have turned out better. In some cases, we feel motivated to make up for the guilt-producing event, but in others we may try to hide or cover up the harm done.

2. The Case for Shame as an Adaptation

To defend the claim that shame is an adaptation, I draw on a significant body of work in philosophy and the social/cognitive sciences that highlights that shame displays the telling features. While there are limitations in these findings, I argue that they make for a compelling, cumulative case. To bolster this conclusion, I end by considering what I see as the strongest objection to the argument I've made.

We can begin with TF1 (adaptive scenario). Here I take it to border on platitude that an individual's fitness is promoted by stable cooperative arrangements. But it's also near platitude both that humans (and so presumably our ancestors) are imperfect rule followers and that our failure to meet group norms and expectations tends to undermine other's willingness to engage with us—especially as the severity and frequency of our failures

[2] Though these sketches are not uncontroversial (see, e.g., Ortony 1987; Elison 2005), they emerge from a wide range of cross-cultural work in philosophy and the cognitive sciences (e.g. Keltner and Buswell 1996; Fessler 2004; 2007; Tangney et al. 1996; Maibom 2019; Fontaine et al. 2006).

increase. This then is the adaptive challenge: how does an individual maintain their status as a cooperative partner in the wake of a (serious or persistent) violation of group norms, expectations, or ideals? Shame is widely thought to be part of the solution. More specifically, in these discussions, shame is understood as an emotion that sensitises one to occasions where one has failed to meet group norms and expectations. In so doing, shame works as a signal and source of motivation. Heidi Maibom explains:

> The person who is ashamed shows to others—through the shame display—not just a *recognition* that they have *failed to live up to public expectations*, but also that they have an adverse emotional reaction to it. … Her shame indicates that she can be *counted on* to live a life with others within the constraints set by the community. (2010, 587–88; emphasis added; see also Beall & Tracey 2020; Sznycer et al. 2018; Fessler 2007; Boehm 2012; Kurth 2025)

But here one might balk, for this account of shame's function seems unable to explain instances of shame that don't concern norm transgressions: shame about one's big ears, say, or being seen naked. But while these are not transgressions *per se*, they are cases where one nonetheless feels that one has failed to live up to (implicit) social expectations or ideals (concerning modesty or how one should look). Moreover, there's an adaptive advantage to being ashamed in "non-transgressive" cases like these—after all, a sensitivity to these social expectations signals one's suitability as a cooperative partner (Greenwald and Harder 1998; Fessler 2004; Maibom 2010).

Moving on, this account of shame's function and adaptive scenario gets enriched by game theoretic modelling indicating that a behavioural response of this sort—namely, behaviour where the transgressor indicates their recognition of the transgression and their intention to do better going forward—is an evolutionarily stable strategy. More specifically, these models indicate that it pays to express shame after a transgression, to the extent that doing so reduces the severity or likelihood of the punishment one's transgression would have otherwise brought. In fact, shame can bring these benefits even if it's costly (e.g., it's painful to experience; it reveals one to have transgressed and so be a legitimate target for punishment). In short, if shame works as researchers like Maibom propose, then these game theoretic models indicate that it's a response that evolution could have selected for as a response to the proffered adaptive challenge (O'Connor 2016; Rosenstock and O'Connor 2018; Shen 2018).[3]

Turn to TF2, where we look for evidence of a proto form of shame. Here researchers note significant parallels between human shame displays and the appeasement responses of chimpanzees and bonobos (Maibom 2010; Boehm 2012; Heesen et al. 2022; Fessler 2007; Keltner and Buswell 1996). For all three primates, we find that aggression by, or even the mere presence of, a dominant group member tends to bring a common pattern of behaviour in subordinates: gaze avoidance, a downward-turned head, and slumped body posture. Moreover, this pattern of behaviour appears to have a distinctive function: it's a signal of submission aimed at forestalling aggression from the dominant, thus helping the individual preserve their standing within the group. Importantly, because this affiliation-seeking submissive behaviour is seen not just in humans but

[3] In much of this game theoretic work, the pattern of behaviour modelled gets called "guilt" and not "shame." This might seem to undermine the point made in the text. But it does not, for these researchers are quick to note that their labels are *mere placeholders* for whatever psychological state brings about the behaviour being modelled. It's the behaviour, not the label, that matters—and per the proposal, shame brings the modelled behaviour. But given that guilt also brings this behaviour, why should these findings be taken as evidence for shame's evolution rather than guilt's (as some have argued—e.g., Ramsey & Deem 2022)? The answer, as we'll see, is that independent research suggesting that shame arrived on the evolutionary scene *before* guilt. So it's more plausible to take the game theoretic evolutionary models as evidence for shame. For further discussion, see Kurth 2023.

also in chimps and bonobos, it provides evidence about what the psychological capacities of our common ancestor, *Pan*, may have looked like. Specifically, *Pan* would likely have had a tendency to engage in affiliation-seeking submissive behaviours toward dominant group members (Maibom 2010; Boehm 2012; Fessler 2007; Gilbert and McGuire 1998; Clark 2008). But to say that is just to say that we have evidence of a proto form of shame—namely, the standing-protecting psychological capacity from which both human shame and chimp/bonobo appeasement behaviour both descended (Figure 1). While this argument for proto-shame is somewhat speculative, we have independent evidence of its plausibility. For instance, both humans and apes show similar patterns of activity in their immune systems and stress-management mechanisms (i.e., the HPA-axis) when engaging in affiliation-seeking submission (see e.g. Gruenewald et al. 2007; Kemeny et al 2004). These findings suggest that the shame/appeasement responses are underwritten by common biochemical mechanisms.

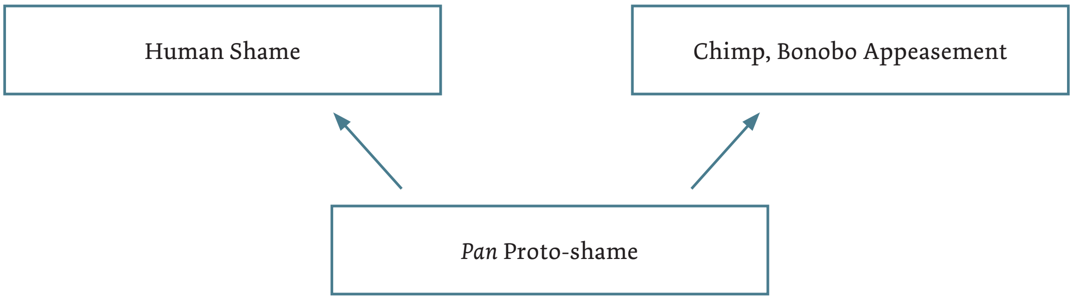

Figure 1. Shame, Appeasement, and Proto-Shame

Moving to TF3 (underlying mechanisms), we've just seen evidence that shame is underwritten by biochemical mechanisms, though it's less clear how unique these mechanisms are. Here work in neuroscience does better, suggesting that shame is supported by distinct neural architecture. More specifically, meta-studies of imaging results suggest that shame and guilt are associated with distinct, but overlapping, neural networks. For shame, we have a network that's centred on *inter alia* the insula and the sensorimotor/premotor cortex, while guilt's network is more associated with the insula and temporoparietal junction (Piretti et al. 2023; Bastin et al. 2016). But we also get an interesting twist from a pair of studies examining cross-cultural differences in the neural patterning of shame and guilt. Petra Michl and colleagues (2014) compared their imaging work on German individuals to the findings of Hidehiko Takahashi et al. (2004) on Japanese participants. The results not only found that there are more cross-cultural differences in the patterning associated with guilt than there are for shame, but also that these differences are disproportionately located in brain regions associated with perspective taking (e.g. the medial frontal gyrus). In an effort to explain this, Michl and colleagues suggest that the differences "indicat[e] that shame manifests itself similarly across cultures, whereas guilt is based more on specific social standards"—a conclusion they take to be grounded in the hypothesis that shame "is linked more directly to biophysiological processes," while guilt is more "culture dependent and learned" (155–56).

We should, of course, be cautious in drawing conclusions from just a few studies. That said, these findings provide interesting data about the origins of shame and guilt. Moreover, Michl and colleagues' conclusion—that the neural mechanisms associated with shame suggest it's biologically hardwired in a way that guilt is not—gain some independent support from a large-scale modelling study done by Dacher Keltner and his team (Keltner et al. 2023). Aggregating findings on the emotion-related experiences of thousands of individuals from around the world, they found cross-cultural support for 21 distinct emotion kinds. While shame was one of these emotions, guilt was not.

Regarding TF4 (expressive routine), shame does not appear to have a distinctive facial expression akin to, say, the gape face of disgust. However, there's widespread agreement that shame comes with a characteristic expressive routine that includes a downward turned head, gaze aversion, and slumped body posture (Beall and Tracey 2020; Maibom 2010; Fessler 2004). In support of this, we've already noted that human shame displays have significant physiological and functional affinities with the appeasement displays of other primates, pointing to their apparent common descent from a form of proto-shame. There is also cross-cultural evidence that these displays are identified as displays of shame at rates significantly greater than chance in Western and non-Western populations, including small-scale societies in Fiji and Burkina Faso (Babcock and Sabini 1990; Tracy and Robins 2008). This is what we'd expect to find given that shame's adaptive scenario (TF1) ascribes a signalling function to the emotion.

Turning to TF5, we have developmental evidence of shame's heritability. For instance, twin studies indicate not just the unsurprising result that shame-proneness is a product of genetic and environmental factors, but also something more significant: while heritable factors are a more significant driver of shame-proneness, environmental factors are a more significant driver of guilt-proneness (Zahn-Waxler and Robinson 1995). Bolstering the conclusion that shame is not a wholly learned or culture-specific behaviour, other work reveals that congenitally blind athletes from both Western and non-Western cultures show the characteristic shame display after being defeated in a competition (Tracy and Matsumoto 2008). That athletes from a range of cultures show shame displays after defeat—despite being blind from birth—strongly tells for it being a pan-cultural, robustly heritable behaviour.

All told, we have evidence that shame exhibits all five telling features. Moreover, the strength of this support is on par with what we find for other emotions that are widely thought to be adaptations, including fear (Öhman 2008; Griffiths 1997), disgust (Kelly 2011; Kurth 2021; Tybur et al. 2013), and anxiety (Kurth 2016; 2018; Marks and Nesse 1994). Yet one might resist this conclusion. After all, we see cultural differences in both what elicits shame and how individuals express the shame they feel. For instance, talk about shame is significantly more prevalent in non-Western cultures such as Indonesia than it is in places like the US (Fessler 2004; see also Kollareth et al. 2018). Moreover, outside of the West, shame is elicited by a wider range of phenomena. The Chinese, for instance, are more likely to feel shame about the misdeeds of a relative than are individuals from the US (Stipek 1998). Similarly, outside the West, shame tends to bring more pro-social behaviour (Bedford 2004; Breugelmans and Poortinga 2006). Findings like these, especially when combined with general scepticism about "basic emotions" (e.g. Barrett 2017; Russell 2004), might lead one to reject the above case for shame as an adaptation.

But that would be too quick.[4] For starters, while we should acknowledge that there are cross-cultural differences in when and how individuals feel shame, we should also recognise that there's a significant degree of similarity. For instance, in a comparative study of shame and guilt in three cultures—Hungary, Belgium, and Peru—the authors noted that "one of the most important findings of this study was the *remarkable similarity* between the three cultural groups'" in how they experienced these emotions (Fontaine et al. 2006). Related work reveals high degrees of similarity between the shame (and guilt) experiences of individuals in Indonesia, Mexico, and the Netherlands (Breugelmans and Poortinga 2006). Another set of large-scale studies found robust patterns of similarity in shame elicitors across three WEIRD nations (Sznycer et al. 2016) and fifteen small-scale societies

[4] Here we should flag two things. First, not only is the general scepticism about "basic emotions" voiced by psychological constructivists highly controversial (see, e.g., Scarantino and Griffiths 2011; Kurth 2022, ch. 3), but there's independent reason to doubt their alternative accounts of what emotions are (see, e.g., Scarantino 2015; Kurth 2019).

(Sznycer et al. 2018). So, taken as a whole, it seems that the most plausible conclusion to draw is this: human shame has a biological core that was selected for by evolutionary forces, though the underlying mechanisms also allow for (sometimes significant) cultural refinement with regard to when and how shame is experienced. In this way, shame appears to parallel emotions like anger (Samore and Fessler 2024), disgust (Kelly 2011), and anxiety (Kurth 2018).

3. The Case Against Guilt as an Adaptation

Shifting from shame to guilt, I'll deploy the same strategy: examining the extent to which we see evidence of the telling features. Starting with TF1 (adaptive scenario), the first thing to note is that, in contrast to shame, we have less work in investigating guilt's evolutionary origins (McGee and Giner-Sorolla 2019; Fessler 2004). That said, within the work we have, advocates tend to adopt one of two strategies to defend the guilt-as-adaptation claim. While the first—one that looks to the adaptive advantages of guilt felt in *anticipation* of transgressions—is more historically prominent, it is also less plausible. So I'll focus on the second strategy, where the evolutionary benefits are seen as being brought by *post-transgression* occasions of guilt. But before turning to that, a brief discussion of the first option is in order.

According to the anticipatory guilt strategy, guilt's adaptive value lies in helping individuals sustain cooperative arrangements in the face of opportunities to defect. Anticipatory guilt can do this because the aversive nature of the emotional experience changes one's decision calculus: it makes defection less appealing (Frank 1988; Joyce 2006; James 2011; Krebs 2011).[5] Various issues have been raised about the adequacy of this line of argument. For one, because these models focus on *anticipatory* guilt, and because most of the guilt we experience is *post*-transgression, without more of a story—which defenders have not provided—it's unclear how the benefits of anticipatory guilt could have been selected for (Stich 2008). These models also say little to explain why the benefits of anticipatory guilt outweigh the cost of being guilt-prone (e.g., expressing guilt invites exploitation; guilt is also linked with various psychopathologies) (Ramsey and Deem 2022).

But in addition to these issues, there's a more damning, but largely unnoticed, problem: these arguments do nothing to distinguish feelings of anticipatory guilt from superficially similar—but distinct—emotions. For instance, to my knowledge, nobody who takes guilt to be an adaptation notes the affinities between anticipatory guilt and fear of punishment, much less offers an explanation for why, for the adaptive scenarios proffered, we should think we're talking about guilt rather than fear.[6] This matters because it seems that fear of punishment would actually be a *better* candidate for the mechanism that changes one's decision calculus. After all, fear is generally conceptualised as a forward-looking emotion and so (in comparison to the predominantly backward-looking nature of guilt; c.f. Meriste 2019) would be better equipped to fulfil the role of countering temptations to defect (Kurth 2018). This is all the more so given that fear (of punishment) is generally thought to have arrived on the evolutionary scene before guilt (Kitcher 2011; see also §4 below), thus raising questions about why evolution would have selected for another tool to do the job that fear already does (Kurth 2023).

[5] Nichols and Barlassina (n.d.) develop a sophisticated version of this strategy, one that seeks to capture the adaptive advantage of anticipatory guilt, not in terms of its benefits to the guilt-prone *individual*, but rather the groups that the guilt-prone are part of. While this is an interesting twist, the proposal remains vulnerable to the problems discussed in what follows.

[6] Elison (2005) notes the affinities between guilt and fear of punishment. But unlike the guilt-as-adaptation advocates, he *denies* that guilt is a distinct emotion that evolutionary forces have selected for. We will return to discuss Elison's view below.

The second strategy focuses not on the role of anticipatory guilt, but of post-transgression guilt. Here the most worked-out accounts of guilt's adaptive scenario focus on its ability to help individuals preserve their standing after transgressing group norms or expectations (see, e.g., Ramsey and Deem 2022; Fowers 2019). More specifically, these accounts focus on guilt as a tool that helps individuals change the motivations of conspecifics, thereby facilitating forgiveness, reducing punishment, and speeding up reintegration.[7] Developing this, defenders note that guilt is a painful emotion that we experience when we feel we're responsible for a transgression. As such, it motivates us to engage in various efforts to make up for the harm we've done: by acknowledging that one has transgressed and by showing an interest in making amends, one makes progress in regaining the trust of one's peers.

But defenders add that if this were the end of the story, it wouldn't be enough—it wouldn't explain why the transgressor's peers opt to reincorporate them into community life rather than just taking advantage of them (Ramsey and Deem 2022; see also Fowers 2019). To close this gap, defenders argue that we need to appreciate the work that our tendency to feel *empathy* in response to seeing the distress of another can do. More specifically, empathy is a psychological mechanism that not only allows individuals to experience a psychological state whose positive or negative valence mirrors the valence of the affective states that they take others to be experiencing, but also motivates individuals to act so as to either preserve positively valenced states brought on by empathy or to eliminate the negatively valenced ones. Moreover, the negative experiences that empathy brings when one sees another in distress are associated with both efforts to alleviate the distress (in oneself or the person in distress) and diminished feelings of anger and aggression toward others. Combining all this then delivers the payoff. In expressing one's feelings of guilt, one communicates the pain that one is experiencing. Because of this, such an expression would have had some tendency to engage the empathy of community members. These community members would then have also tended to experience negatively valenced affective states, and so have been motivated to act in ways that would eliminate the negative affect they were feeling. The result would have been some tendency for community members to restore the social status of those who express post-transgression guilt.

Importantly, unlike proponents of the anticipatory guilt proposal, defenders of this strategy typically acknowledge that because guilt is similar to other emotions, more needs to be said about why, for this adaptive challenge, it's plausible to think that it was guilt, and not some other emotion, that evolutionary forces were selecting for. Here they focus on the most likely competitor for a post-transgression emotional response: shame. For instance, drawing on empirical findings, both Fowers (2019) and Ramsey and Deem (2022) present guilt as an emotion that is focused on *transgressive acts*, that engages beliefs that one is *responsible* for what happened, and that motivates *pro-social efforts* to repair or apologise for the harm done. They then contrast this profile of guilt with shame, which they understand as an emotion that's focused on *one's self*, engages beliefs that one's self is *defective or flawed*, and that motivates *anti-social* behaviours like hiding, withdrawal, or even violence.

The central issue with this proposal is its vagueness. Advocates' contentions to the contrary, we do not have an account of guilt's function that distinguishes it from shame and so we do not have an account that explains why guilt—rather than shame—was selected for by evolution. Turning to the details, it's true that the above account

[7] A third strategy sees guilt's adaptive advantage not as facilitating reintegration, but rather as promoting altruism: guilt brings an altruistic concern against harming others that evolved out of the empathetic/sympathetic mechanism undergirding the parent/child care system (see, e.g., Gilbert 2003; Tangney and Dearing 2002). But as others have noted, defenders of this proposal offer little by way of empirical support for their view (McGee and Giner-Sorolla 2019)—it seems little more than a just-so story. Given this, I won't discuss it further.

of the differences between guilt and shame has some empirical support, especially from early work on these emotions—and that's the problem. More specifically, the vast majority of this early work builds from survey results that use the Test of Self-Conscious Affect (TOSCA) developed by June Tangney (1990). But the TOSCA measure employs a biased conceptualisation of the two emotions: questions assessing guilt-proneness are framed in terms of prosocial tendencies (e.g. taking responsibility, making amends), whereas questions assessing shame-proneness are framed in terms of dysfunctional behaviours (e.g. avoidance, negative self-assessment). So no surprise then that guilt is pro-socially oriented, but shame isn't—the emotions are *defined* so as to deliver this result (Maibom 2019; Elison 2005; Luyten et al. 2002). Moreover, when non-TOSCA-based studies of the differences between guilt and shame are used, the findings suggest that there's little—if any—difference between the two emotions with regard to (a) perceptions of whether a moral standard was violated, (b) the extent to which the emotions engage thoughts of responsibility, or (c) a focus on one's actions rather than one's self (see, e.g., Tangney et al. 1996; Keltner and Buswell 1996). Add to this that (d) there is a significant body of work challenging the contention that guilt prompts prosocial motivations while shame prompts dysfunctional ones (e.g. Gausel et al. 2016; de Hooge et al. 2008; Pivetti et al. 2016). In fact, some of this research suggests that shame—not guilt—is the more pro-socially oriented emotion (e.g. Allpress et al. 2014; de Hooge et al. 2011).

These findings are particularly important for assessing guilt with regard to TF1. After all, (a)–(d) correspond to the aspects of the guilt models that do the heavy lifting in the above explanations of guilt's adaptive scenario. So if there is little difference between guilt and shame along these dimensions, then why think we have an evolutionary account of guilt specifically? Moreover, the argument does not fare better if we set (a)–(d) aside and focus on areas where guilt and shame actually do differ (according to non-TOSCA measures). This work indicates that, in comparison to shame, guilt has stronger associations with both thoughts that one has harmed another person and tendencies to ruminate on what happened (see, e.g., Fontaine et al. 2006; Breugelmans and Poortinga 2006). But it's hard to see how these internal cognitive tendencies would have provided an adaptive advantage with regard to securing post-transgression reincorporation. So we have little support for TF1.[8]

Turn then to TF2 (proto-guilt). Here defenders of the guilt-as-adaptation proposal are divided. Some are sceptical, noting that "the available data [from primatology and anthropology] are inconclusive as to whether other primates experience guilt or some form of proto-guilt" (Deem and Ramsey 2016, 575). This group speculates that guilt may be best understood as an adaptation that is distinctive of humans. Others are more optimistic, offering a range of proposals about which emotion guilt is thought to have emerged from—e.g. sympathy (Frank 1988), empathy (Gilbert 2003), shame (Boehm 2012; Joyce 2006), as well as regret/sadness (Fessler 2004). The fact that we don't have agreement on guilt's precursor is not, on its own, an issue—these are difficult questions that we are only beginning to understand. That said, it's worth noting that these suggestions are notably thin on details. For instance, Robert Frank appreciates that there should be "some emotional precursor to guilt," and then provides the following as his account: "Sympathy is a natural candidate. In order for an act that harms another person to summon guilt, it is necessary that we feel at least some sympathy" (1988, 65). Even if we accept this as an account of where guilt came from, the thing to note is how hand-wavey it is, especially in comparison to either the more detailed account that we saw for shame and appeasement, or the accounts that we have for other emotions that are thought to be adaptations (e.g. fear and anxiety as emerging from predator defence mechanisms (Öhman 2008; Kurth 2018; 2016)). All told, the charitable conclusion to draw is that the case for guilt meeting (or explaining away) TF2 is questionable.

[8] For additional worries with this way of defending guilt's adaptive scenario, see Kurth 2023 and Nichols and Barlassina n.d.

Our discussion of guilt and the remaining telling features can proceed more quickly given the findings discussed in §2. On TF3 (underlying mechanisms) we've seen evidence that guilt exhibits a distinctive pattern of neural activation. But, as we also noted, it is unclear whether this patterning is evidence of a biologically grounded mechanism or a learned one. We've also seen that both neuroimaging and large-scale modelling work tell against guilt's mechanisms being biologically-grounded. So—at best—we have qualified support for TF3. Turning to TF4, there's widespread consensus that guilt lacks a distinctive expressive routine—even among those who argue for guilt as an adaptation (e.g. Ramsey and Deem 2022). As for TF5 (heritability), we find little support. For instance, while some research suggests there are guilt displays in young children (e.g. Barrett 1998), this work does little to distinguish these purported instances of guilt from confounds like empathetic distress. Additionally, and as noted above, twin studies suggest that environmental factors, not genetic ones, are the primary driver of guilt-proneness. So, unlike shame, there is little support here either.

Stepping back, Table 1 sums up the evidence for shame and guilt exhibiting the telling features. From this, we can see that while there's good reason to think that shame is an adaptation, there is little reason to say the same about guilt.

	Shame	**Guilt**
TF1. Adaptive Scenario	✓	X
TF2. Other Primates	✓	?
TF3. Mechanisms	✓	✓*
TF4. Expressive Signal	✓	X
TF5. Heritable	✓	?

Table 1. Shame, Guilt, and the Telling Features

Importantly, this conclusion challenges the received view about the evolutionary origins of these emotions. With a few exceptions (Ortony 1987; Elison 2005), shame and guilt are standardly thought to stand and fall together: if one is (not) an adaptation then so (neither) is the other (see, e.g., Ramsey and Deem 2022; Frank 1988; James 2011; D'Arms and Jacobson 2023; Krebs 2011; Griffiths 1997; Prinz 2004).[9] Moreover, if the received view about the origins of shame and guilt is mistaken, then we should also rethink the more general (and equally common) contention that self-conscious emotions—shame, guilt, pride, embarrassment, etc.—are of a piece with regard to their evolutionary origins and cognitive architecture.

4. Why Do We Feel Guilt?

Suppose the argument so far is convincing. It leaves an important question unanswered: if guilt is not an adaptation, then why do humans feel it? The answer, I suggest, is that guilt is a kind of social technology—a culturally-driven innovation that helped our ancestors address particular, recurrent challenges of social life. So understood, guilt is akin to other social technologies that humans have developed to facilitate cooperation and social interaction. To develop this, I begin with some familiar examples of these types of (non-emotional) social technologies. Seeing how they differ from things like fads will help us see what a corresponding emotional technology would look like. With this done, I will make my cases for guilt as a technology.

[9] Ekman is an interesting case insofar as he has argued both that guilt and shame are adaptations (see e.g. 1999) and that neither is one (e.g. 1984). Fessler (2007) flirts with the idea that while shame is an adaptation, guilt is not. But he ultimately concludes that we lack sufficient evidence to decide the matter.

4.1 Preliminaries

Let's begin with some examples of non-emotional social technologies. First, there are promises. As Hume's example of the two farmers reveals, temporally-extended *quid pro quo* exchanges present a distinctive challenge: why should I agree to help you with your harvest today in exchange for your help with mine tomorrow given that, come tomorrow—and your corn already safely in the silo—you'll have no reason to help me? As Hume sees it, promises are "a certain form of words *invented* for" addressing this problem: your promise gives me comfort about taking the risk of helping you today because it allows you to "bind [yourself] to the performance of" helping me tomorrow (1978, 3.2.5; emphasis added; see also Khan 2024). Similarly, currency is a technology introduced to facilitate commercial exchange in situations where bartering fails (because, say, one party doesn't have goods or services that other party wants) (van der Spek and van Leeuwen 2018).

For present purposes, two aspects of these technologies are of note: (i) they're purposeful in that they were developed in order to address distinctive coordination problems, and (ii) the technologies are sustained by the combination of internalised models of appropriate behaviour and the social structures that regulate them (e.g. blaming practices for broken promises; institutions of trust that sustain currencies). These features distinguish technologies from fads and similarly short-lived, cultural novelties. While fads like the Macarena or pet rock collecting might bring a brief burst of attention and social affiliation, they lack both the coordination-problem-addressing purposefulness of (i) and the practice-sustaining internalisation and regulation of (ii).

Importantly, we can extend this picture to emotional phenomena. Consider *amok*, a negatively valenced affective experience that brings an intense episode of violent behaviour that typically culminates in the violence-doer's own death. Building on historical analyses, we can see *amok* as an emotional technology that emerged within the Malayan honour cultures of Southeast Asian. Briefly, the term *amok* originates from the Malay *meng-âmuk*—meaning, roughly, "to make a furious and desperate charge" (Saint Martin 1999). Add to this that the earliest written accounts of *amok* suggest that it's an emotional innovation: a combination of anger, frustration, and shame that afforded individuals a way to preserve their reputation in contexts—such as defeat in war or impending enslavement—where they had no other way to protect their honour (Imai et al. 2019). This work also suggests that the violence characteristic of one who runs *amok* is something that one "learn[s] … is an appropriate response to certain unbearable social pressures" (Griffiths 1997, 141). So understood, *amok* is both purposeful in the sense of (i), and robustly internalised and sustained in the manner of (ii). Like promises and currencies, it has the requisite marks of a technology.

The *amok* example also highlights how emotional technologies, as I'm understanding them, differ from more linguistically-focused refinements of emotions that have been discussed by others. For instance, Shaver and colleagues (1992) suggest that certain emotions like guilt and embarrassment are not distinct emotion types, but rather occasions where we've introduced a new word in order to circumscribe some variation in the intensity or content of a more "basic emotion." Thus, "embarrassment" might be the label used to specify mild forms of shame, and "guilt" is the label that might mark off instances of sadness where concerns about one's responsibility for harm are the focus. In this way, Shaver et al.'s account of guilt and embarrassment is akin to what Justin D'Arms and Dan Jacobson (2003) call "cognitive sharpenings"—(quasi-)stipulative truncations of (basic) emotions. A different linguistically-focused proposal focuses specifically on guilt, understanding our talk of it as a kind of metaphor.[10] More specifically, when I say that I feel guilty, I'm speaking in an "as-if"

[10] Proposals of this sort have found traction among both psychologists (e.g., Ortony 1987, Elison 2005) and philosophers (e.g., Maley & Harman 2019 and, perhaps, Greenspan 1992).

way: I am (elliptically) indicating both (a) that I believe that I am guilty (in a socio-legal understanding of the term) and (b) that I am feeling the emotions that one would typically feel when one believes and cares that one is in this condition of being guilty (e.g. shame, sadness, fear). Importantly, on this picture, guilt is not an emotion. Rather, it's an affective-cognitive hybrid anchored in (metaphorical) associations to socio-legal notions of guilt.

With these examples in hand, we can compare them to the idea of emotions as social technologies. For starters, both of the linguistically-focused proposals are similar to the above emotional technology account of *amok* insofar as they seek to locate guilt (and embarrassment) within a broader typology of human affective states: as technologies, cognitive sharpenings, or affective-cognitive hybrids. But the emotional technology proposal is distinctive insofar as it explicitly takes up the question that we're focused on here: if an emotion like guilt or *amok* is not an adaptation, then why do we feel it? Moreover, notice that to understand an emotion like *amok* or guilt as a technology (rather than, say, a sharpening or metaphor) is to see it as an *inherently* purposeful and socially-sustained practice on par with things like promises and currencies. The result is a richer account of the nature and bio-cultural underpinnings of these emotions.

With this foundation laid, I'll now turn to make my case for understanding guilt as an emotional technology. To do this, I draw on findings from cognitive science to locate a coordination problem that guilt—but not shame—is particularly well-suited to address. I then show how my guilt-as-technology proposal makes predictions that find support in the empirical record.

4.2 The Limits of Shame and the Gap Guilt Fills

The overlap in the functional profiles of shame and guilt suggests that guilt is a culturally created tool—a technology—that works to supplement and extend the work that shame evolved to do. To develop this idea, we can begin by considering what the social life of our hunter-gatherer ancestor is thought to have been like and why life in such small-scale communities facilitated the evolutionary emergence of shame. This will help us see why, with the emergence of large-scale civilisation, shame wasn't enough and, more specifically, which gap guilt may have been developed to help fill.

To begin, we've seen that shame is thought to have been selected for because of its ability to help individuals retain their status as cooperative partners in the wake of violations of group expectations. To better appreciate the adaptive advantage of having such a mechanism, recall that we're thinking about our early hunter-gatherer ancestors, individuals whose lives and livelihood depended on group cohesion. It's not just that the tasks of daily life—child-rearing, big game hunting, food sharing, and the like—were collaborative exercises, but also that, given the small size of these communities, their success turned crucially on the group's ability to secure stable coordination and consensus about what to do. Unsurprisingly, then, studies of the social organisation of contemporary nomadic peoples (whose lifestyle is thought to approximate that of our hunter-gatherer ancestors) reveal structures that facilitate this cooperation: flat social organisation, democratic decision-making, and practices where community members quickly address norm violations (Boehm 1999; Grossmann 2023; Kitcher 2011; Kurth 2016; Sterelny 2012).

Against this backdrop, related research indicates that the presence of bullies and psychopaths likely posed a decidedly pernicious threat to small-scale social life. Because these individuals so persistently and deliberately break the rules, and because they're unfazed by the standard social sanctions that return ordinary transgressors to the fold, their free-riding can quickly undermine cooperation. In fact, this threat was so significant that those with bullying and psychopathic tendencies were likely to be killed if they didn't express an intention to reform their ways (Boehm 2012; Fessler 2007). In this context, post-transgression shame is thought to have been particularly important because it provided the needed signal: when ashamed, one shows that one (now) accepts the prevailing community norms (§2). Being shame-prone was then a way to stay alive by demonstrating that one is *not* a bully or psychopath.

But let's now jump forward to the large-scale civilisation of Mesopotamia and Egypt. Here the lives of our ancestors were very different. For one, group size had grown significantly—no longer was it possible to recognise everyone you lived with, much less know much about them. Social organisation had also changed. There was greater variety in, for instance, the civil and economic roles that individuals occupied (jobs as, e.g., farmers, metal workers, accountants). As a result, social, political, and economic structures became more hierarchical and less democratic. More importantly, specialisation and the concentration of knowledge, skills, and resources in the hands of those who specialised, not only brought a shift away from flat, democratic political structures focused on group consensus, but also shepherded in new forms of social organisation that were centred on *prestige* and fostering relationships with those who have it (Kitcher 2011; Sterelny 2012; Fessler 2004; Henrich and Gil-White 2001).

For our purposes, there are three noteworthy aspects of this transition. First, not only did the transition bring significant growth and change in the norms governing social life, but these elaborations were also surely a source of trouble: more uncertainty about what was forbidden, less clarity on what the sanctions for norm violations were, and so on (Kitcher 2011; Kurth 2016). Second, the trouble that these changes brought was also likely to have weakened traditional mechanisms of norm enforcement. Gossip loses its force as a corrective when it's about someone you don't know and aren't likely to engage with. Similarly, shame—i.e., a tool that evolved to motivate and signal concern for *group* norms and expectations—seems less significant, even *problematic*. After all, throughout this transition, we see a *waning* in concern for group cohesion and a *rise* in attention to maintaining prestige and relationships with individuals who have it. Given this, shame's distinctive—and visible—expressive signature would have started to carry costs. After all, while broadly signalling one's intention to conform is advantageous when one violates *group* expectations, it becomes a liability when one's actions merely harm an individual or when public exposure of one's misdeeds comes at a cost to one's prestige (Fessler 2004; Greenwald and Harder 1998). Finally, despite these significant alterations in social organisation and orientation, *cooperation remains stable* (Kitcher 2011; Kurth 2016; Sterelny 2013). So what explains this?

As a first step toward an answer, notice that the above discussion of the transition from hunter-gatherer living to large-scale civilisation reveals not just that life became more complicated, but that power shifted. Those who had control over resources and specialised knowledge—that is, those with prestige—came to have outsized influence over political, social, and economic decision making. This means that violations of the expectations of individuals with prestige were likely to be particularly costly. Moreover, while shame, as an emotion sensitive to *group expectations*, would have been of some help in rectifying harms done to the prestigious, its automatically engaged expressive signature would, as we just saw, also bring significant costs. Thus, there would be benefits in mechanisms that allowed one to make up for damage done to *particular (prestigious) relationships*, but that didn't come with shame's liabilities.

Enter guilt, here understood as a culturally-fashioned emotional technology that some groups developed in order to supplement the work that shame does. In comparison to shame, guilt is distinctive, I suggest, because it serves as a repair-motivation booster. In comparison to shame, guilt is more concerned with harm done to *individual relationships* than violations of *group-level* expectations and ideals. And unlike shame, guilt *lacks* a publicly broadcasted signal. So while both shame and guilt are repair-oriented emotions, guilt has features that allow it to do the work that, as human social structures shifted, shame was poorly equipped to do. Importantly, we can say all this *without* also assuming that guilt is an evolutionary adaptation. Rather, like promises, currency, or *amok*, guilt is better understood as a culturally-created, socially-sustained technology.

4.3 The Evidence

The above proposal fits with the lessons we've learned in examining the telling features (§§2–3). For instance, the idea that guilt is an innovation, not an adaptation, accords with the neuroimaging and heritability findings suggesting that guilt is a more culture-dependent and learned response than shame. But the guilt-as-emotional-technology proposal also makes novel predictions that find support in the empirical record.

Prediction 1: Shame preceded guilt. On the above proposal, humans developed the capacity to experience guilt in order to fill a gap left by shame. This entails that shame preceded guilt in evolutionary time. Is there evidence that supports this? Yes. Two converging lines of research suggest that guilt arrived later. First, guilt is generally seen as cognitively complicated in the sense that feeling guilt requires the ability to form complex representations of oneself and others, to distinguish oneself from others, and to see oneself as responsible for a particular happening (Olthof et al. 2000; Lagattuta and Thompson 2007). This final capacity—the ability to make responsibility assessments—is particularly noteworthy here because it is not a capacity that, as we've seen, seems essential to the ability to feel shame. Thus, guilt is more cognitively complicated than shame and so is likely a capacity that appeared later. Second, work in child development indicates that the ability to experience and recognise guilt emerges in children well after they are able to experience and recognise other, less cognitively complicated responses like fear, sympathy, *and shame*. While these observations are about the time sequence in which emotions come online for individuals, it's also generally taken as evidence about the emergence of emotions in evolutionary time (Ramsey and Deem 2022; Barrett 1998; Harris 1989; Zahn-Waxler and Kochanska 1990). Thus, we have another piece of evidence that guilt arrived after shame.

Prediction 2: Guilt is more oriented toward relationship repair—especially relationships with the prestigious. If guilt functions as a repair-motivation booster oriented toward damage done to particular relationships, then we should see evidence that, in comparison to shame, it's more concerned with damage done to individual relationships. Four lines of evidence provide support for this. The first set of results concerns the types of situations that elicit guilt. Here we find, for instance, that when individuals are asked to recall the details of their emotional experiences, their guilt descriptions are, in comparison to shame, not only more often about occasions involving individuals they esteem or have close relationships with (73% for guilt, 61% for shame), but also less often about strangers and casual acquaintances (21% for guilt, 30% for shame) (Tangney et al. 1996; see also Baumeister, Stillwell, and Heatherton 1995). Additionally, Darren McGee and Roger Giner-Sorolla (2019) directly tested the hypothesis that, while shame is more focused on group-level reputational issues, guilt is more oriented toward damage done to individual relationships. To do this, they asked individuals to imagine how they would induce guilt (or shame) in another person. The results affirmed their hypothesis: for shame, participants chose to use a *public exposure* strategy—"Look how others see you"—significantly more often; but for guilt, they were significantly more likely to appeal to *interpersonal harm*—"Look at how you've harmed me"—especially when asked to imagine interacting with a friend (versus a stranger).

Second, with regard to relationship repair tendencies, we see that guilt-proneness better predicts whether one will apologise to an individual they've transgressed than does shame-proneness (Chrdileli and Kasser 2018; see also Ruckstaetter et al. 2017). We also see that guilt-driven efforts toward relationship repair are moderated by one's beliefs about whether the harmed individual will know that the transgressor tried to make up for the damage done (Cryder et al. 2012). That guilt inclines you to apologise only to the extent that you think the apology will be recognised as such by the person you harmed strongly suggests that it's a strategy for relationship repair. Of course, shame also motivates effort to repair. Here a third line of research helps us understand how guilt and shame differ on this front. In contrast with the above findings indicating that guilt has a comparatively strong connection to relationship repair, other work highlights shame's greater connection to group-level concerns. More specifically, we see that in situations where members of one's *group* have done wrong, shame—not guilt—is a better predictor of one's motivation to make up for the harm done (de Groot et al. 2021; Allpress et al. 2014; Rees et al. 2013).

Finally, we have results affirming guilt's orientation toward the prestigious, understood here as individuals who are admired because of their superior knowledge/skills and their willingness to share what they know with others (Henrich and Gil-White 2001). For instance, we see that people feel more guilt when they've harmed someone they esteem than when they harm someone that they have little regard for (Baumeister et al. 1995; see also Berndsen et al. 2004, Vangelisti et al. 1991). We also find that individuals tend to experience stronger feelings of guilt when they believe that a wrong they've done has been observed by someone they esteem than when they think the misdeed has been seen by a stranger (Oda and Sawada 2021). All told, while both guilt and shame prompt relationship repair efforts (§3), they systematically differ in the damage they're concerned to fix: guilt is more orientated toward damage done to particular relationships, while shame is more focused on group-level concerns.

Prediction 3: In small-scale groups, terms for "shame" pervade but terms for "guilt" do not. The guilt-as-technology proposal maintains that, in small-scale societies, shame plays a more prominent role in structuring social interactions than does guilt. Now add an independent premise: the function of emotion terms lies, in part, in their role in regulating the associated emotions—that is, we have the terms "anger" and "pride" in part because these labels help us regulate anger and pride, respectively (D'Arms 2005; Lindquist et al. 2006). Together, this predicts that, since small-scale communities have less need to regulate guilt, they will be less likely to have a term for it; but the same will not be true of shame. Again, this is what we see. While there is no research suggesting that there are cultures lacking a term for shame, we have a range of anthropological studies indicating that many small-scale societies do not have terms for guilt (Boehm 2012; Breugelmans and Poortinga 2006; Fessler 2004). Moreover, it's not just that many small-scale communities lack a word for guilt, but that they also have trouble understanding the kinds of "guilt" situations we're considering—situations where one feels bad as a result of harming another individual—as situations where one should feel bad. For instance, Daniel Fessler found that among the Bangkulu villagers he was studying, these paradigmatic "guilt" events were *puzzling*. As he explains, "people seemed hesitant or confused" when asked about cases like these and often "simply remarked on the wrongness of harming others, *making no reference to emotions*" (2004, 223; emphasis added).

Prediction 4: Cross-cultural differences. On the guilt-as-technology proposal, guilt is an emotion that some groups developed to address challenges that emerged as human civilisations became bigger, less democratic, and more prestige oriented. This invites the following cross-cultural predictions. First, given shame's greater focus on group-level concerns, it should, in comparison to guilt, play a more significant role in collectivist cultures that have a greater orientation toward the group. By contrast, given guilt's comparatively greater focus on individual relationships, it should be more prominent in individualist cultures where an orientation toward independence and prestige are emphasised.

There is some support for these predictions. For instance, Fessler's anthropological studies of individuals in Southern California and Bengkulu suggest that guilt is more pronounced than shame among Southern Californians (2004; 2007; see also Wallbott and Scherer 1995). More provocatively, Fessler's work also reveals that among Southern Californians, features common to these emotions—i.e., a concern about doing harm, efforts to make amends, reduced concern for the opinions of uninvolved observers—are more often attributed to guilt than shame. In short, the more individualistic Southern Californians are not only more guilt-oriented, but their guilt functions as the guilt-as-technology proposal suggests it will: it's a mechanism that helps individuals address damage done to particular (prestigious) relationships. Similarly, Millie Creighton's (1990) anthropological work found that while individuals in the US and Japan experience both guilt and shame, guilt is the more prominent emotion in the US; but the reverse is true in Japan. Elaborating on this finding, Creighton notes how these contrasting emotional emphases cohere with these cultures' broader individualistic (US) and collectivistic (Japan) approaches to things like child rearing. More recently, Lina Liw and colleagues (2022) found that broadly collectivistic values predicted shame-proneness, while individualistic values predicted guilt-proneness (see also Rozin 2003). Together, findings like these support the above predictions.

Summing up, though more work is certainly needed, we have converging lines of support for the idea that guilt is a piece of emotional technology: an innovation that some cultures developed in order to help address challenges distinctive of life in large-scale communities.[11]

4.4 A Refined Understanding of Shame and Guilt

In addition to providing support for the idea that guilt is an emotional technology, the above discussion also provides a refined understanding of what shame and guilt are. On this picture, the functional profiles of these emotions *significantly overlap*—more so than is generally appreciated (especially in TOSCA-based work). As we've seen, both are concerned with protecting one's standing as a cooperative partner post-transgression, both prompt combinations of pro- and anti-social behaviour, and both are often concerned with matters of (moral) responsibility. But there are also subtle differences. For instance, while shame is more oriented toward

[11] Taken as a whole, the discussion in the text suggests that the guilt-as-technology account is more explanatorily powerful than are the language-based proposals discussed in §4.1. Though space doesn't allow me to say much here, I'll briefly flag three issues. First, if guilt were merely a linguistic variant of sadness, as Shaver et al. suggest, or a metaphorical extension of shame (or some other basic emotion) as Ortony, Elison, and others maintain, then we would expect instances of guilt to retain parts of the distinctive expressive routines of the emotions they're associated with (sadness, shame, etc.). But guilt's lack of an expressive signature (§3) suggests this is not so—and that tells against these views. Second, on the Shaver et al. proposal, guilt is just a subset of sadness. So the functional roles of the two emotions should be similarly structured. But, again, that's not what we see. For instance, the dominant action tendency of sadness is withdrawal. By contrast, the action tendency for guilt is more complex. While it can at times prompt withdrawal, it also (and more often) tends to bring efforts to *engage*—to repair or make amends for the harm done. Finally, on the account of guilt defended by Ortony, Elison, and Maley and Harman, believing that one is guilty is *essential* to feeling guilty—this element, after all, captures the sense in which they are affective-*cognitive* hybrids (see, e.g., Ortony 1987, 285; Elison 2005, 11; Maley and Harman 2019, 30, 28). But research on the guilt that individuals feel in response to accidents suggests beliefs aren't actually essential. For instance, individuals who accidentally kill or harm another person will report feeling guilty while acknowledging that—given that it was an accident—they are not guilty of anything (see Zhao 2020 for discussion). While defenders might reply that cases like these still involve an attenuated belief about being guilty, this move is less plausible when we look at phenomena like survivor guilt. While survivor guilt is a complicated phenomenon, clinical work points to a range of cases where individuals report feeling guilt despite knowing they were not causally or morally responsible for (say) the fatal plane crash. What seems essential to these guilt experiences is that the individuals view the outcome as *unfair*, not that they believe themselves to be guilty for what happened, however attenuated that belief might be (see, e.g., Murray et al. 2021; also, Lebra 1983).

group-level concerns (e.g. violations of group expectations, damage to group image), guilt is more tuned to protecting individual relationships (Prediction 2). Moreover, this difference in the concerns of these emotions helps explain other differences between them (§1, §2). If shame is more group-focused, then it makes sense both that it has stronger associations with the feeling that one is being looked at by others and that it carries a greater concern for damage done to one's public reputation. And if guilt is more concerned with damage done to individual relationships, it helps explain why it has stronger associations with thoughts of having harmed another person, and is more likely to bring rumination about what happened. The functional profiles of these emotions also shed light on why shame, but not guilt, has a distinctive expressive signature, as well as why we're ashamed not just of transgressions, but also things like our looks, family history, and financial status. Taking these in turn, the slumped body posture and gaze aversion characteristic of shame signals to others that one knows one has violated group expectations. By contrast, given guilt's focus on damage to individual relationships, not only is there less need for a widely broadcast signal, but as we've seen, having one may be a liability insofar as it could undermine one's standing in the eye of individuals one has *not* harmed. Rather, what matters when one damages a relationship is that one is able to convey *just* to the harmed individual that one is sincerely concerned to repair the damage done. For this, words and actions seem better tools. Shifting gears, our tendency to be ashamed of things like our appearance is just an extension of shame's orientation toward failures to meet public expectations—be they explicit rules not to harm others, or basic standards of social acceptability (Fessler 2004; Maibom 2010; Kurth 2025).

5. Conclusion: Are Emotions Kinds?

Let's return to our original question: are emotions best understood as natural kinds or social constructions? Though our focus has been on a different question—are shame and guilt adaptations?—what we've learned is telling. After all, the received opinion maintains that being an adaptation is the quintessential mark of the natural kinds that are characteristic of biology and psychology. So combining this with what we've learned about shame—namely, that a strong case can be made for it being an adaptation—brings the conclusion that shame is a kind. Guilt, by contrast, looks a lot more like a social construction. Not only is there a comparatively weak case for viewing it as an adaptation, but on the account developed here, it seems more like a culturally-driven innovation—a technology—akin to how we understand things like promises, currencies, and *amok*. If this is right, then there's also a larger lesson. The common idea that emotions—as a class—must be understood as either kinds or constructions is mistaken: for whether an emotion is a kind depends on what emotion we're looking at.

Acknowledgment
Earlier versions of this paper were presented at the University of Helsinki and the 2023 EPSSE meeting in Tartu. Thanks to those audiences for fruitful conversations. Hat tips as well to: Heidi Maibom, Antti Kauppinen, Anni Kajanus, Ritva Palmén, Fredrik Westerlund, and two anonymous referees. A special thanks is also due to Heidy Meriste for the invitation to contribute to this special issue and for her work bringing the collection into print.

References

Allpress, J., R. Brown, R. Giner-Sorolla, J. Deonna, and F. Teroni. 2014. “Two Faces of Group-Based Shame.” *Personality and Social Psychology Bulletin* 40 (10): 1270–84.

Babcock, M. K. and J. Sabini. 1990. “On Differentiating Embarrassment from Shame.” *European Journal of Social Psychology* 20: 151–69.

Barrett, K. 1998. “The Origins of Guilt in Early Childhood.” In *Guilt and Children*, edited by J. Bybee, 75–90. Academic Press.

Barrett, L. F. 2017. *How Emotions Are Made*. Houghton Mifflin Harcourt.

Bastin, C., B. Harrison, C. Davey, J. Moll, and S. Whittle. 2016. “Feelings of Shame, Embarrassment and Guilt and their Neural Correlates.” *Neuroscience & Biobehavioral Reviews* 71: 455–71.

Baumeister, R., A. Stillwell, and T. Heatherton. 1995. “Guilt as Interpersonal Phenomenon.” In *Self-Conscious Emotions*, edited by J. P. Tangney and K. W. Fischer, 255–73. Guilford.

Beall, A. and J. Tracy. 2020. “The Evolution of Pride and Shame.” In *The Cambridge Handbook of Evolutionary Perspectives on Human Behavior*, edited by L. Workman, W. Reader, and J. Barkow, 179–93. Cambridge University Press.

Bedford, O. 2004. “The Individual Experience of Guilt and Shame in Chinese Culture.” *Culture & Psychology* 10: 29–52.

Berndsen, M., J. van der Pligt, B. Doosje, and A. Manstead. 2004. “Guilt and Regret: The Determining Role of Interpersonal and Intrapersonal Harm.” *Cognition and Emotion* 18: 55–70.

Boehm, C. 1999. *Hierarchy in the Forest*. Harvard University Press.

———. 2012. *Moral Origins*. Basic Books.

Boyd, R. 1999. “Homeostasis, Species, and Higher Taxa.” In *Species: New Interdisciplinary Essays*, edited by R. A. Wilson, 141–85. MIT Press.

Breugelmans, S. and Y. Poortinga. 2006. “Emotion without a Word.” *Journal of Personality and Social Psychology* 91 (6): 1111–22.

Chrdileli, M. and T. Kasser. 2018. “Guilt, Shame, and Apologizing Behaviour: A Laboratory Study.” *Personality and Individual Differences* 135: 304–306.

Clark, J. A., 2008. “Relations of Homology between Higher Cognitive Emotions and Basic Emotions.” *Biology & Philosophy* 25: 75–94.

Creighton, M. R. 1990. “Revisiting Shame and Guilt Cultures: A Forty- Year Pilgrimage.” *Ethos* 18 (3): 279–307.

Cryder, C.E., Springer, S., and C. K. Morewedge. 2012. “Guilty Feelings, Targeted Actions.” *Personality and Social Psychology Bulletin* 38: 607–618.

D’Arms, J., 2005. “Two Arguments for Sentimentalism.” *Philosophical Issues* 15: 1–21.

D’Arms, J. and D. Jacobson. 2003. “The Significance of Recalcitrant Emotion.” *Royal Institute of Philosophy Supplements* 52: 127–45.

———. 2023. *Rational Sentimentalism*. Oxford University Press.

de Groot, M., J. Schaafsma, T. Castelain, K. Malinowska, L. Mann, Y. Ohtsubo, M. T. A. Wulandari, R. F. Bataineh, D. P. Fry, M. Goudbeek, and A. Suryani. 2021. “Group- Based Shame, Guilt, and Regret Across Cultures.” *European Journal of Social Psychology* 51: 1198–212.

de Hooge, I. E., S. M. Breugelmans, and M. Zeelenberg. 2008. “Not So Ugly After All: When Shame Acts as a Commitment Device.” *Journal of Personality and Social Psychology* 95: 933–43.

de Hooge, I., R. Nelissen, S. Breugelmans, and M. Zeelenberg. 2011. “What Is Moral about Guilt? Acting ‘Prosocially’ at the Disadvantage of Others.” *Journal of Personality and Social Psychology* 100: 462–73.

Deem, M. and G. Ramsey. 2016. “Guilt by Association?” *Philosophical Psychology* 29 (4): 570–85.

Ekman, P. 1984. “Expression and the Nature of Emotion.” In *Approaches to Emotion*, edited by K. Scherer and P. Ekman, 314–44. Erlbaum.

———. 1999. “Basic Emotions.” In *Handbook of Cognition and Emotion* edited by T. Dalgleish and M. Power, 45–60. Wiley.

Elison, J. 2005. “Shame and Guilt: A Hundred Years of Apples and Oranges.” *New Ideas in Psychology* 23: 5–32.

Fessler, D. 2004. “Shame in Two Cultures: Implications for Evolutionary Approaches.” *Journal of Cognition and Culture* 4 (2): 207–62.

———. 2007. “From Appeasement to Conformity.” In *The Self-Conscious Emotions*, edited by J. Tracy, R. Robins, and J. Tangney, 174–93. Guilford Press.

Fontaine, J. R. J., P. Luyten, P. de Boeck, J. Corveleyn, M. Fernandez, D. Herrera, A. Ittzés, and T. Tomcsányi. 2006. “Untying the Gordian Knot of Guilt and Shame.” *Journal of Cross-Cultural Psychology* 37: 273–92.

Fowers, B. 2019. “The Evolution of Guilt and Its Non-instrumental Enactments.” In *The Moral Psychology of Guilt*, edited by B. Cokelet and C. Maley, 113–30. London: Rowman & Littlefield.

Frank, R. 1988. *Passions within Reason*. W. W. Norton.

Gausel, N., V. Vignoles, and C. Leach. 2016. “Resolving the Paradox of Shame: Differentiating Among Specific Appraisal-Feeling Combinations Explains Pro-social and Self-Defensive Motivation.” *Motivation and Emotion* 40: 118–39.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Harvard University Press.

Gilbert, P. 2003. “Evolution, Social Roles, and the Difference in Shame and Guilt.” *Social Research* 70: 1205-1230.

Gilbert, P. and M. McGuire. 1998. “Shame, Status, and Social Roles: Psychobiology and Evolution.” In *Shame*, edited by P. Gilbert and B. Andrews, 99–125. Oxford University Press.

Greenspan, P. 1992. “Subjective Guilt and Responsibility.” *Mind* 101: 287–303.

Greenwald, D. and D. W. Harder. 1998. “Domains of Shame: Evolutionary, Cultural, and Psychotherapeutic Aspects.” In *Shame: Interpersonal Behaviour, Psychopathology, and Culture*, edited by P. Gilbert and B. Andrews, 225–24. Oxford University Press.

Griffiths, P. 1994. “Darwinism, Process Structuralism, and Natural Kinds.” *Philosophy of Science* 63: S1–S9.

———. 1996. “Cladistic Classification and Functional Explanation.” *Philosophy of Science* 61: 206–227.

———. 1997. *What Emotions Really Are*. University of Chicago Press.

Grossmann, T. 2023. “The Human Fear Paradox.” *Behavioral and Brain Sciences* 46: e52.

Gruenewald, T., S. Dickerson, and M. E. Kemeny. 2007. “A Social Function for Self-Conscious Emotions.” In *The Self-Conscious Emotions*, edited by J. Tracy, R. Robins, and J. Tangney, 68–90. Guilford.

Harris, P. L. 1989. *Children and Emotion*. Blackwell.

Heesen, R., D. A. Austry, Z. Upton, and Z. Clay. 2022. “Flexible Signalling Strategies by Victims Mediate Post-Conflict Interactions in Bonobos.” *Philosophical Transactions of the Royal Society B* 377: 20210310.

Henrich, J. and F. Gil-White. 2001. “The Evolution of Prestige.” *Evolution and Human Behavior* 22: 165–96.

Hume, D. 1978. *A Treatise of Human Nature*. Edited by L. Selby-Bigge and P. Nidditch. Oxford University Press.

Imai, H., Y. Ogawa, K. Okumiya, and K. Matsubayashi. 2019. “Amok: A Mirror of Time and People.” *History of Psychiatry*, 30 (1): 38–57.

James, S. 2011. *An Introduction to Evolutionary Ethics*. Wiley-Blackwell.

Joyce, R. 2006. *The Evolution of Morality*. MIT Press.

Kelly, D. 2011. *Yuck!* MIT Press.

Keltner, D. and B. Buswell. 1996. “Embarrassment: Its Distinct Form and Appeasement Functions.” *Psychological Bulletin* 122 (3): 250–70.

Keltner, D., J. A. Brooks, and A. Cowen. 2023. “Semantic Space Theory: Data-Driven Insights into Basic Emotions.” *Current Directions in Psychological Science* 32 (3): 242–49.

Kemeny, M., T. Gruenewald, and S. Dickerson. 2004. “Shame as the Emotional Response to Threat to the Social Self: Implications for Behavior, Physiology, and Health.” *Psychological Inquiry* 15: 153–60.

Khan, S. 2024. “Commitment: From Hunting to Promising.” *Biology & Philosophy* 39: 5.

Kitcher, P. 2011. *The Ethical Project*. Harvard University Press.

Kollareth, D., J. Fernandez-Dols, and J. A. Russell. 2018. “Shame as a Culture-Specific Emotion Concept.” *Journal of Cognition and Culture* 18: 274–92.

Krebs, D. 2011. *The Origins of Morality: An Evolutionary Account*. Oxford University Press.

Kurth, C. 2016. “Anxiety, Normative Uncertainty, and Social Regulation.” *Biology and Philosophy* 31: 1–21.

———. 2018. *The Anxious Mind*. MIT Press.

———. 2019. “Are Emotions Psychological Constructions?” *Philosophy of Science* 86: 1227–38.

———. 2021. “Cultivating Disgust: Prospects and Moral Implications.” *Emotion Review* 13: 101–112.

———. 2022. *Emotion*. Routledge.

———. 2023. “An Evolutionary Account of Guilt?” *Philosophy of Science* 92: 510–18

———. 2025. “Sames and Selves: On the Origins and Foundations of a Moral Emotion.” *British Journal for the Philosophy of Science*.

Lagattuta, K. H. and R. A. Thompson. 2007. “The Development of Self-Conscious Emotions.” In *The Self-Conscious Emotions*, edited by J. L. Tracy, R. W. Robins, and J. P. Tangney, 91–113. Guilford.

Lebra, T. 1983. “Shame and Guilt: A Psychocultural View of the Japanese Self.” *Ethos* 11: 192–209.

Lindquist, K. A., L. F. Barrett, E. Bliss-Moreau, and J. A. Russell. 2006. “Language and the Perception of Emotion.” *Emotion*, 6 (1): 125.

Liw, L., A. Ciftci, and T. Kim. 2022. “Cultural Values, Shame and Guilt, and Expressive Suppression as Predictors of Depression.” *International Journal of Intercultural Relations* 89: 90–99.

Luyten, P., J. Fontaine, and J. Corveleyn. 2002. “Does the Test of Self-Conscious Affect (TOSCA) Measure Maladaptive Aspects of Guilt and Adaptive Aspects of Shame?” *Personality and Individual Differences* 33: 1373–87.

Maibom, H. 2010. “The Descent of Shame.” *Philosophy and Phenomenological Research* 80 (3): 566–94.

———. 2019. “On the Distinction between Shame and Guilt.” In *The Moral Psychology of Guilt*, edited by B. Cokelet and C. Maley, 37–51. Rowman and Littlefield.

Maley, C. and G. Harman. 2019. “The Feeling of Guilt.” In *The Moral Psychology of Guilt*, edited by B. Cokelet and C. Maley, 13–36. Rowman & Littlefield.

Marks, I. and R. Nesse. 1994. “Fear and Fitness.” *Ethology and Sociobiology* 15: 247–61.

McGee, D. and R. Giner-Sorolla. 2019. “How Guilt Serves Social Functions from Within.” In *The Moral Psychology of Guilt*, edited by B. Cokelet and C. Maley, 149–70. Rowman & Littlefield.

Meriste, H. 2019. “Against Exclusively Retrospective Guilt.” In *The Moral Psychology of Guilt*, edited by B. Cokelet and C. Maley, 71–94. Rowman & Littlefield.

Michl, P., T. Meindl, F. Meister, C. Born, R. Engel, M. Reiser, and K. Hennig-Fast. 2014. “Neurobiological Underpinnings of Shame and Guilt: A Pilot fMRI Study.” *Social Cognition and Affective Neuroscience* 9 (2): 150–57.

Murray, H., Y. Pethania, and E. Medin. 2021. “Survivor Guilt: A Cognitive Approach.” *The Cognitive Behaviour Therapist* 14: e28.

Nichols, S. and L. Barlassina. n.d “Not for Me: On the External Function of Guilt.” Unpublished manuscript.

O’Connor, C. 2016. “The Evolution of Guilt: A Model-Based Approach.” *Philosophy of Science* 83: 897–908.

Oda, R. and K. Sawada. 2021. “Do Social Relationships with Those Who Witness Moral Transgression Affect the Sense of Guilt?” *Evolutionary Psychology* 19.

Öhman, A. 2008. “Fear and Anxiety.” In *Handbook of Emotions*, edited by M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, 127–56. Guilford.

Olthof, T., A. Schouten, H. Kuiper, H. Stegge, and A. Jennekens- Schinkel. 2000. “Shame and Guilt in Children: Differential Situational Antecedents and Experiential Correlates.” *British Journal of Developmental Psychology* 18: 51–64.

Ortony, A. 1987. “Is Guilt an Emotion?” *Cognition and Emotion* 1: 283–98

Piretti, L., E. Pappaianni, C. Garbin, R. I. Rumiati, R. Job, and A. Grecucci. 2023. “The Neural Signatures of Shame, Embarrassment, and Guilt: A Voxel-Based Meta-Analysis on Functional Neuroimaging Studies.” *Brain Sciences* 13 (4): 559.

Pivetti, M., M. Camodeca, and M. Rapino. 2016. “Shame, Guilt, and Anger: Their Cognitive, Physiological, and Behavioral Correlates.” *Current Psychology* 35: 690–99.

Prinz, J. 2004. *Gut Reactions*. Oxford University Press.

Ramsey, G. and M. Deem. 2022. “Empathy and the Evolutionary Emergence of Guilt.” *Philosophy of Science* 89 (3): 434–53.

Rees, J. H., J. A. Allpress, and R. Brown. 2013. “Nie Wieder: Group- Based Emotions for In- Group Wrongdoing Affect Attitudes toward Unrelated Minorities.” *Political Psychology* 34 (3): 387-407.

Rosenstock, S. and C. O'Connor. 2018. "When It's Good to Feel Bad: An Evolutionary Model of Guilt and Apology." *Frontiers in Robotics and AI* 5: 9.

Rozin, P. 2003. "Five Potential Principles for Understanding Cultural Differences in Relation to Individual Differences." *Journal of Research in Personality* 37 (4): 273–83

Ruckstaetter, J., J. Sells, M. D. Newmeyer, and D. Zink. 2017. "Parental Apologies, Empathy, Shame, Guilt, and Attachment: A Path Analysis." *Journal of Counseling & Development* 95: 389–400.

Russell, J. A. 2004. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110: 145–72.

Saint Martin, M. 1999. "Running Amok: A Modern Perspective on a Culture-Bound Syndrome." *Primary Care Companion to the Journal of Clinical Psychiatry* 1: 66–70.

Samore, T. and D. Fessler. 2024. "Steps Toward an Interdisciplinary Anthropology of Mind and Emotion: Intersections Between Evolutionary and Psychological Anthropologies." In *The Cambridge Handbook of Psychological Anthropology*, edited by E. Lowe. Cambridge University Press.

Scarantino, A. 2015. "Basic Emotions, Psychological Construction, and the Problem of Variability." In *The Psychological Construction of Emotion*, edited by L. F. Barrett and J. A. Russell, 334–76. Guilford.

Scarantino, A. and P. Griffiths. 2011. "Don't Give up On Basic Emotions." *Emotion Review* 3: 1–11.

Shaver, P., S. Wu, and J. Schwartz. 1992. "Cross-Cultural Similarities and Differences in Emotion and Its Representation." In *Review of Personality and Social Psychology*, vol. 13, edited by M. S. Clark. 175–212. Sage.

Shen, L. 2018. "The Evolution of Shame and Guilt." *PLoS ONE* 13: e0199448.

Slater, M. 2015. "Natural Kindness." *British Journal for the Philosophy of Science* 66: 375–411.

Sterelny, K. 2012. *The Evolved Apprentice*. MIT Press.

———. 2013. "Life in Interesting Times." In *Cooperation and Its Evolution*, edited by K. Sterelny, R. Joyce, B. Calcott, and B. Fraser, 89–108. MIT Press.

Stich, S. 2008. "Some Questions about 'The Evolution of Morality'." *Philosophy and Phenomenological Research* 77: 228–36.

Stipek, D. 1998. "Differences between Americans and Chinese in Circumstances Evoking Pride, Shame, and Guilt." *Journal of Cross-Cultural Psychology* 79: 616–29.

Sznycer, D. and A. S. Cohen. 2021. "Are Emotions Natural Kinds After All? Rethinking the Issue of Response Coherence." *Evolutionary Psychology* 19.

Sznycer, D., D. Xygalatas, E. Agey, S. Alami, X.-F. An, K. I. Ananyeva, Q. D. Atkinson, B. R. Broitman, T. J. Conte, C. Flores, S. Fukushima, H. Hitokoto, A. N. Kharitonov, C. N. Onyishi, I. E. Onyishi, P. P. Romero, J. M. Schrock, J. J. Snodgrass, L. S. Sugiyama, K. Takemura, C. Townsend, J.-Y. Zhuang, C. A. Aktipis, L. Cronk, L. Cosmides, and J. Tooby. 2018. "Cross-Cultural Invariances in the Architecture of Shame." *Proceedings of the National Academy of Sciences* 115: 9702–707.

Sznycer, D., J. Tooby, L. Cosmides, R. Porat, S. Shalvi, and E. Halperin. 2016. "Shame Closely Tracks the Threat of Devaluation by Others, Even Across Cultures." *Proceedings of the National Academy of Sciences* 113: 2625–30.

Takahashi, H., N. Yahata, M. Koeda, T. Matsuda, K. Asai, and Y. Okubo. 2004. "Brain Activation Associated with Evaluative Processes of Guilt and Embarrassment." *Neuroimage* 23 (3): 967–74.

Tangney, J. 1990. "Assessing Individual Differences in Proneness to Shame and Guilt: Development of the Self-Conscious Affect and Attribution Inventory." *Journal of Personality and Social Psychology* 59 (1): 102–111.

Tangney, J. and R. Dearing. 2002. *Shame and Guilt*. Guilford.

Tangney, J., R. Miller, L. Flicker, and D. Barlow. 1996. "Are Shame, Guilt, and Embarrassment Distinct Emotions?" *Journal of Personality and Social Psychology* 70 (6): 1256–69.

Tracy, J. and D. Matsumoto. 2008. "The Spontaneous Display of Pride and Shame." *Proceedings of the National Academy of Sciences* 105: 11655–60.

Tracy, J. and R. Robins. 2008. "The Automaticity of Emotion Recognition." *Emotion* 7: 789–801.

Tybur, J. M., D. Lieberman, R. Kurzban, and P. DeScioli. 2013. "Disgust: Evolved Function and Structure." *Psychological Review* 120: 65–84.

van der Spek, R.J. and B. van Leeuwen, eds. 2018. *Money, Currency and Crisis*. Routledge.

Vangelisti, A. L., J. A. Daly, and J. R. Rudnick. 1991. "Making People Feel Guilty in Conversations: Techniques and Correlates." *Human Communication Research* 18: 3–39.

Wallbott, H. G. and K. R. Scherer. 1995. "Cultural Determinants in Experiencing Shame and Guilt." In *Self-Conscious Emotions*, edited by J. P. Tangney and K. W. Fischer, 465–87. Guilford.

Zahn-Waxler, C. and G. Kochanska. 1990. "The Origins of Guilt." In *Socioemotional Development: Nebraska Symposium on Motivation*, edited by R. A. Thompson, 183–258. University of Nebraska Press.

Zahn-Waxler, C. and J. Robinson. 1995. "Empathy and Guilt: Early Origins of Feelings of Responsibility." In *Self-Conscious Emotions*, edited by J. P. Tangney and K. W. Fischer, 143–73. Guilford.

Zhao, M. 2020. "Guilt Without Perceived Wrongdoing." *Philosophy and Public Affairs* 48: 285–314.

Received: 26 March 2024
Accepted: 10 June 2025
Published: 28 August 2025

Constituting Emotional Phenomena — A Mach-Influenced Empiricist Perspective

Peter Zachar – Auburn University Montgomery, USA, pzachar@aum.edu

Abstract

Using the philosophical writings of Ernst Mach as a backdrop, I explore how concepts and classifications partly constitute the phenomena studied in the science of emotion by selecting features from a larger population of features. This process of selection is a matter of decision and is not inevitable, but it promotes populating concepts with empirical content. The openness of empirical concepts suggests that this selectionist constituting does not characterise only the early stages in the development of a science because background and foreground shifts are potentially ongoing. The theory of psychological construction, which contends that emotional episodes are constructed on the fly out of shifting sets of components, exemplifies this selectionist sense of constituting to the extent that it advocates for a resemblance nominalism, similar to that of Locke, in which selection is involved in naming kinds. Examples of constituting can be seen in changing definitions of whether animals experience emotion and in the choice of causal models.

Keywords: essentialism, nominalism, open concepts, operational definition, psychological construction, scientific conventionalism

1. Introduction

Those perspectives that adopt a more or less constructionist account of emotional phenomena have several features in common. First, according to constructionism, the way in which phenomena are conceptualised and classified is not inevitable. A second feature, following from the first, is that alternative descriptions of phenomena are possible. A third feature, inspired by the first two, is that some choice, decision, or selection is involved in describing phenomena.

Although constructionist analyses are usefully contrasted with views that construe emotional phenomena as mind-independent, natural kinds, there are accounts of natural kinds that can accommodate all three features just enumerated. This has somewhat deflated the constructed kind versus natural kind contrast. With respect to mind-independence, a more contemporary question focuses on the extent to which our concepts and classifications in some way constitute emotional phenomena.

Stronger versions of constituting, often labelled neo-Kantian, hold that concepts and classifications actively structure phenomena—imparting structure that is not already inherently there. A milder version of constituting, typically more aligned with empiricist perspectives, takes constituting to be a matter of actively selecting relevant features from a wider population of features.

In this article, I will elaborate on an empiricist-selectionist approach to constituting, drawing on the work of Ernst Mach. Mach is interesting because, alongside the pragmatists, he articulated one of the first post-Darwinian approaches to empiricism, and did so in opposition to neo-Kantian philosophies. He anticipated many features of the scientific conventionalism that was central to the empiricism of the logical positivists, but did not go so far as to claim that scientific conventions are definitions disguised as descriptions of facts. Finally, of interest to psychological scientists, Mach pondered the constituting role of measurement, thus anticipating operationalism.

2. The Philosophy of Ernst Mach

The physicist and psychologist Ernst Mach's (1914) philosophical writings are not as well known as they once were, but his perspective was a basic assumption of most of the early work in the philosophy of science. The spirit of Mach's views was ably summarised by Einstein ([1916] 1996, 142):

> Concepts that have proven useful in ordering things can easily attain an authority over us such that we forget their [worldly][1] origin and take them as immutably given. … Therefore, it is not at all idle play when we are trained to analyze the entrenched concepts, and point out the circumstances that promoted their justification and usefulness and how they evolved from the experience at hand. This breaks their all too powerful authority.

Mach is frequently described as having incorrectly disputed the reality of atoms, construing them as theoretical posits not supported by sensory evidence. He rejected atoms, but his reasons for doing so were more nuanced than claiming that atoms are not real because we cannot see them. His doubts seemed to derive from his view that believing in atoms involved being too literal about abstract and simplifying scientific models. His scepticism about atoms, however, by itself, does not justify relegating Mach to a footnote in the philosophy of science. He remains relevant for articulating an empiricist alternative to trendy neo-Kantian ideas. This type of empiricism lives on in both pragmatism and various deflationary views on metaphysics.

Mach's abiding concern was that scientists took notions drawn from what Einstein called experience at hand and converted them into absolute realties. Especially disagreeable to Mach's "anti-metaphysical" view would be to make a distinction between appearance (understood as what we can experience on the surface) versus reality (understood as what is behind those appearances). Like William James, Mach rejected the claim that, to understand experience, we have look at what is *behind* experience, or *underlies* it, or is transcendentally *prior* to it. We can apply metaphors to distinguish between shallow and deeper features of experience and between more or less penetrating accounts of experience, but such metaphors do not get us outside of experience to an absolute thing-in-itself.

Another important feature of Mach's philosophy is his claim that knowledge generation is analogous to a Darwinian process of selection in which potential experience contains a wide variety of features. We selectively attend to some of these variations and ignore or minimise others—always working to put things into an orderly arrangement under the guidance of our current goals and purposes. Mach refers to this as the principle of economy. Economy is partly psychological. We cannot represent all the details potentially available to us

[1] The English translation had this as wordly not worldly. Worldly is more consistent with the original German. Thanks to Emma Bolton for tracking this down.

and so need to work with summaries, rules of thumb, and generalisations. More recent philosophers refer to these summaries as “models” (van Fraassen 2002; Giere 1999).

3. Scientific Conventionalism

The selection process discussed by Mach, which can be more or less conscious, involves decisions. In the early 20th century, the importance of decisions in the generation of scientific knowledge was placed at the forefront of the philosophy of science, under the rubric of scientific conventionalism. Three features of the conventionalist analysis can be highlighted.

First, scientific conventions result from choices that are made and which are not inevitable, but once in place they may become taken for granted and assumed. For example, it is not inevitable that longitude and latitude are oriented west to east and north to south, but that orientation is so standardised that it seems inevitable.

Second, the choice of a scientific convention has an extra-empirical dimension. Such choices are not arbitrary, but neither are they necessitated by the facts. In the philosophy of science this feature is associated with *underdetermination*, which holds that a plurality of concepts, theories, and classifications can be adequate to the facts (or to experience). For example, the metre is a convention. However useful it may be, no spatial facts make the metre the true measure of length any more than they make the 12-inch foot the true measure of length.

Third, scientific conventions are chosen because they promote the discovery of facts, or, in the language of the logical positivists, because they are bridges that help populate theoretical concepts with empirical content (van Loo and Romejin 2015). Thus, scientific conventions are decisions that facilitate the discovery of facts—or, in Mach’s terms, that facilitate the adaptation of thoughts to facts.

For example, Paul Ekman initially proposed that basic emotions such as fear last only seconds because that was how long emotional facial expressions last. He later altered his view and stipulated that basic emotions can last longer than seconds but maintained that the duration of an emotion is still relatively brief—not more than a few minutes (Ekman and Cordaro 2011; Ekman 1993).

Once either of these conventions about duration is adopted, they can guide scientists to populate a basic emotion concept with empirical content. For example, physiological processes that occur on a time scale longer than seconds could not be used to populate a short duration basic emotion concept—in measuring such processes we would be measuring something in addition to basic emotions. Once a convention about duration is adopted, the process of discovering facts about a basic emotion so defined can proceed, but shorter and longer durations will limit the basic emotion concept differently and thus facilitate the discovery of different assortments of facts.

4. Operational Definitions and Open Concepts

Conventionalism in the philosophy of science is closely aligned with operationalism in scientific methodology. In the context of measurement, both conventionalism and operationalism hold that one’s choice of a measure makes a non-trivial contribution to what is measured (Tal 2020). One meaning of non-trivial is that the measure (or the concept) in some way constitutes what is measured.

As noted, the extra-empirical considerations utilised in the choice of a scientific convention imply that other choices were possible. In one version of scientific conventionalism, advocated for by Poincaré ([1905] 2001), a convention is not subject to correction by the facts. According to this interpretation, conventions are definitions, like "A bachelor is an unmarried man." This definition prescribes the meaning of the term "bachelor." We could decide to change the meaning of the term, but there is no fact that we can discover that would force us to change the meaning of the term.

The "not subject to correction by the discovery of facts" feature of scientific conventionalism is called into question by the notion of open concepts. This notion undermined the view of scientific conventions as a priori, pre-empirical assertions that are true by definition only.

As a scientific convention, a duration criterion such as "Basic emotions last less than 10 seconds" is definitional of basic emotions—or what Poincare called a definition disguised as a description. For Poincare, definitions in disguise are neither true nor false, rather, they stipulate the boundaries and limits of our concepts. In contrast, Pap (1953) argued that there is a malleable relationship between the defining features and the more contingent features of a concept. In Pap's view, new information can lead us to revise what we take to be the defining features of a concept.

With open concepts there can potentially be foreground–background shifts, in which something that was relegated to the background is brought forward and given prominence, or something formerly in the foreground is backgrounded and seen as more contingent. Described in this way, when Ekman altered the duration criteria for a basic emotion, he moved facial expressions more into the background and brought appraisals a bit more into the foreground.

This process is more than a reshuffling of already known features. The importance of openness is found in encounters with something new and unexpected that suggests possibilities for taxonomic change. This is a more active form of constituting, similar to the way that selecting a sample based on specific features (e.g., college-educated Asian females) more actively constitutes the sample than does random selection from a population.

In psychology, Paul Meehl used open concepts to criticise operationalism as a theory about the meaning of concepts (MacCorquodale and Meehl 1948; Cronbach and Meehl 1955). As a theory of meaning, operationalism holds that the concept is synonymous with the measurement instrument. In Bridgman's (1927) terms: "If we have more than one set of operations, we have more than one concept," and "strictly there should be a separate name to correspond to each different set of operations" (10). Bridgman's (1945) views were more nuanced than these quotes indicate, but they do express how psychologists understood operationalism in the 1930s and 1940s, often illustrated by referring to Boring's (1923) definition of intelligence as the capacity to do well on intelligence tests.[2]

Meehl thought that this was simply incorrect. For example, one way to begin measuring intelligence is to rely on teachers' evaluations of children's cognitive abilities. A teacher's concepts, however, would doubtlessly include implicit meanings such as "brighter kids retain more information" and "learn this information quicker." In philosophical terms, these implicit meanings are surplus meanings of the teacher's concepts.

[2] One of Boring's points was that we should not confuse what we know about the measure of intelligence (a narrow concept) with an everyday common sense notion of intelligence.

By systematically elaborating on surplus meanings and explicitly incorporating them into one's measure of intelligence, Meehl claimed that we can, over time, iterate our way to a more valid measure/concept of intelligence. Eventually, as very few improvements can be made, the concept can even be closed.

With respect to openness, a related perspective was articulated by Waismann (1945), who asserted that scientific conventions and operational definitions are inescapable in science. Nor are they limited to the early stages of scientific research programmes. Rather than open concepts, Waismann wrote about open textures, by which he meant that no description of a phenomenon can cover all the possible facts and circumstances that may be relevant to characterising that phenomenon (Makovec and Shapiro 2019; Makovec 2019). What he means is illustrated by the following quote.

> We can never be quite sure that we have included in our definition everything that should be included, and thus the process of defining and refining an idea will go on without ever reaching a final stage. In other words, every definition stretches into an open horizon. (Mackinnon, Waismann, and Kneale 1945, 125)

As noted by Makovec (2025), one implication of this view is that future meanings, concepts, and classifications are indeterminate and, in some cases, not predictable from past meanings.

5. *Construction* in psychological construction is different than *construction* in social construction

Psychological constructionist perspectives are useful for illustrating how conventions can constitute phenomena, but illustrating this requires elaborating on what is meant by psychological "construction." For example, a psychological constructionist would not dispute that how we understand and experience emotions is influenced by cultural and social processes, thus psychological construction can readily incorporate social constructionist insights. Does this mean that psychological and social construction are examples of the same process of construction? An article by Kurth (2019) articulates such a view. According to Kurth, the psychological constructionists James Russell and Lisa Barrett claim that an essential feature of emotion is a state of core affect labelled with a folk emotion concept, i.e., that distinguishing between emotions involves projecting culturally-fashioned concepts onto felt affective episodes.

This claim is a bit off target. Rather than claiming that fear is psychologically constructed by projecting an emotion concept onto a state of core affect, the theory of psychological construction is about the construction of complex psychological states out of a variety of components. It is a different kind of "construction" than what is meant by "social construction." In psychological construction, emotions are not manifestations of ready-made dispositions, they are psychological episodes put together on the fly—out of components. For Russell (2012), these components include core affect, cognitive appraisal, and self-categorisation.

6. Abundant Variation

Mach's principle of economy is contingent on variation being ample and abundant. The notion of abundant variation, as used in both biology and psychometrics, refers to the proliferation of individual differences, many of which lie outside the boundaries of our concepts for biological species, emotions, personality traits, and so on. Concepts remain open because different groupings of abundant variation can be foregrounded.

Likewise, the theory of psychological construction is predicated on viewing an emotion, such as fear, as a population of episodes replete with variation, and which has no core essence that is equally present across all of those variants. Any specific kind of emotion has the structure of a radial category composed of prototypical instances, variations from the prototype, and boundary cases (Russell 1991).

One useful way to understand this feature of psychological construction is to view it as a form of resemblance nominalism similar to that of Locke (Zachar 2022). For Locke, naming something is the workmanship of human understanding, which guides us to recognise the variety of features co-occurring across episodes as going together to form a *kind*. One of Locke's ([1689] 1997) examples of nominalism was the naming of different kinds of killing, e.g., genocide, herbicide, infanticide, parricide, and suicide. Locke says we have specific names for certain kinds of killing, but not all. For example, parricide refers to the killing of one's parent, but we do not have a specific name for the killing of a second female cousin. Likewise, in psychological construction, if the patterning of components resembles a pattern encoded in some emotion concept (e.g., fear), we will identify that pattern as being that *kind* of emotion. However, many patterns of components occur that are not classified and named (as kinds).

One way to elaborate on resemblance nominalism is to contrast psychological construction with essentialist forms of basic emotion theory. According to an essentialist perspective on basic emotion, affect programmes are innate features that are hard wired into the brain during development. When activated, affect programmes are said to automatically produce the coordinated affective, cognitive, and behavioural responses that characterise specific basic emotions such as fear. Thus, all valid episodes of basic fear are supposed to share the affect programme for fear. In contrast, the nominalist theory of psychological construction claims that different episodes of fear share many things in common, but the only one thing they all share in common is the name "fear."

7. Constituting Emotion Concepts (and Phenomena)

Scientific conventions and operational definitions specify (more-or-less) what is to be included within the boundaries of a concept and what is to be excluded. For a definition such as "A bachelor is an unmarried man," this is relatively straightforward. A 30-year-old unmarried man is a bachelor, a 13-year-old boy and a 30-year-old woman are not. But what about a widower, or someone who is living with a romantic partner but not married? Certain definitions of bachelor exclude them—so there is potentially some selection. When it occurs, selecting plays a constituting role.

Let's look at some examples of constituting in the domain of emotion. When my little dog retreats to the safety of the laundry room during a thunderstorm or meets me at the door when I come home if it is storming, does he experience fear? Joseph LeDoux, who did pioneering work on studying the role of the amygdala in the classical conditioning of emotional responses, would say no, he does not.

LeDoux's groundbreaking work consisted in mapping the pathway from the pairing of an unconditioned stimulus (e.g., a light shock) and a conditioned stimulus (e.g., a tone) to a conditioned "fear response" involving the amygdala (LeDoux 1996). Several years later, after realising that this system works independently of any conscious awareness, LeDoux (2014) decided that what he was actually mapping was the behavioural and physiological responses to threat. LeDoux argued that the defining feature of fear for humans, namely the self-conscious experience of fear that occurs when one is in danger, is generated by cortical machinery that is not conserved across species, as the threat response circuitry is. In this redefinition, the conditioned response to threat, rather than being described as fear, was shifted into a broader conceptual background for fear.

Importantly, this definitional shift was more than an arbitrary stipulation. According to LeDoux and Pine (2016), treatments developed for anxiety disorders on the basis of animal models of fear and anxiety have had limited success with humans. This, they believe, is because those treatments target behaviour and physiological responses to threat, not conscious fear and anxiety. Indeed, this new definition of fear potentially supports some empirical progress regarding treatment targets, differentiating between when to target the process of attention to threat and when to target cognitive appraisal. For example, one might target attention to threat when treating phobias in children and target cognitive appraisal when treating bereavement-related depression in adults.

A more ontologically weighted perspective on constituting is advocated in Lisa Barrett's (2017) constructed theory of emotion. According to the constructed theory, an instance of an emotion such as fear is constructed when conceptual knowledge about fear is actively brought to bear on an occurrence of unpleasant affect being changed in response to an event in the environment. This process, however, is more involved than a projection of a concept onto an episode of core affect.

An important background theory to Barrett's notion of constituting is that conscious experience does not contain a literal representation of the world; rather it is a simulation of a world, based on how incoming information is interpreted. The incoming information lacks structure. According to the constructed theory of emotion, many concepts are available to structure incoming information, but when the incoming information is actively augmented by an emotion concept such as fear, the brain changes its own pattern of activity (its simulation) and *generates* fear on the spot.

This view is more neo-Kantian, arguing that emotion concepts constitute emotional phenomena by structuring them. In Barrett's view, emotions are not ready-made phenomena that we recognise and classify, but rather in some sense emotions are generated when they are classified.

For psychological construction in an empiricist-selectionist framework, the psyche is like a chess board with many pieces that can be arranged in multiple ways. Some of these arrangements are given specific names, such as the Sicilian Defence. For the more neo-Kantian-framed constructed theory, emotions are generated in being conceptualised. In the chess example, this would be more like the act of bringing conceptual knowledge to bear resulting in the pieces then moving into the pattern named. The moving of the pieces sounds like magic in the chess example, but less so when the example is the neural simulation of the body–world relationship.

The constructed theory includes a selectionist form of constituting as well because it makes a "conceptual act" a defining feature of emotions, thus limiting it in specific ways. It is still a nominalist perspective because the emotion of fear refers to a family of fear episodes (or population of specific varieties of fear) with no shared essence. These variations are embedded in a linguistic network of other concepts, such as existential loss and death, none of which are available to my dog, and so his brain cannot construct an instance of fear. Whatever he experiences when he retreats to the laundry room when it is thundering has been relegated by Barrett to a background feature of fear.

These revised definitions are scientific conventions. They are freely chosen and not forced on us by the facts themselves. Each one is a possible way of framing an emotion such as fear—guiding us on how to populate the concept with empirical content—but neither corresponds to what fear really is in some privileged sense. If we adopted a different convention, my dog could likely be described as experiencing fear. For instance, Scarantino and Griffiths (2011) demarcate three kinds of basicness: conceptual, biological, and psychological. Biologically basic emotions are homologous across species and evolved to rapidly adapt to fundamental life tasks—and do not incorporate everything that a folk concept would classify as fear. Under their description of "biologically basic," my dog could reasonably be said to experience fear.

8. Mind-Framed, Not Mind-Controlled

An informative perspective on the constituting role of conceptualising and naming is articulated by Hasok Chang (2022). In Chang's terms, phenomena are not pre-figured. Put another way, rather than already being specified independent of conceptualisation, phenomena are non-specific, i.e., phenomena do not exist already classified in some correct, privileged way. A classification cannot be pre-correct, independent of one's purposes for classifying. For example, a map of the London underground accurately shows one how to travel from Paddington Station to Baker Street Station via the subway, but if your goal is to walk from Paddington Station to Baker Street, that "tube" map would be a terrible guide (Kitcher 2001). You need a street map. The correctness of any map depends on what you want to use it for.

As an alternative to Locke's workmanship of human understanding, Chang labels the act of identifying and naming as mind-framing. The notion of "mind-framed" rather than "pre-figured" is nicely illustrated by Putnam's (1996) claim that language cannot be divided into a part that describes our conceptual contributions and a part that describes the world as it is anyway. Putnam's views shared important similarities with Carnap's ([1956] 1991) claim that metaphysical assertions can be legitimate internal to a conceptual system once some assumptions are in place, but not legitimate external to all conceptual systems in an absolute sense.

It is important to clarify that we cannot frame and fit things together in whatever way we want. Wanting or preferring something to be true or correct does not make it so. There are constraints. Chang's terms are helpful here as well—what we name is mind-framed but not mind-controlled. Whether it be atoms or patricide or fear, the patterns that occur are more than things we make up.

9. Causation, Convention, and Kinds

One of the most counter-intuitive features of Mach's philosophy was his view that causation is a metaphysical notion that can be minimised in science. In Mach's view, if one understands the functional relationships between variables (e.g., f = ma), labelling one variable the cause (f, or force) and the other the effect (a, or acceleration), does not tell us anything new about these relationships. Mach would supplant the simplified analyses of one cause, one effect, with an analysis of functional relationships.

The psychological constructionist criticism of the essentialist version of basic emotions is grounded in a particular model of causation, namely a mechanistic model based on decomposition and localisation. In this model, to causally explain a psychological process involves decomposing it into parts and understanding how those parts are organised to produce or sustain an outcome (Craver and Darden 2013).

For instance, a complex psychological state such as fear could be decomposed into a rapid heartbeat, pupil dilation, and attention to threat. Attention to threat can be further decomposed, and so on, until the parts are simple enough to be explained mechanistically. In Russell's (2003) view, once scientists demarcate causal mechanisms for the components, there will be no role for a causal mechanism of fear in addition to these component mechanisms.

Ironically, this mechanistic model of causation is much the same as the model of causation utilised by basic emotion theories in which affect programmes are identity-determining causal mechanisms. In the affect programme model, the affect programmes mediate between external events and the coordination of subjective feeling, cognitive appraisals, action tendencies, etc. Psychological constructionists such as Russell, however, argue that there are no emotion-specific mechanisms that can be described as affect programmes.

Russell (2003) illustrates his alternative to the affect programme model of emotions by making an analogy with constellations and emotions. In astronomy, stars are points of light in the night sky that, unlike planets, do not change their positions relative to each other (because they are very far away). Just as people can see faces in clouds, they can see patterns in the relative positions of the stars. Historically, the most important of these patterns were the constellations of the zodiac. Examples include the constellations of Leo and Pisces.

The point of the analogy is that just as constellations are happenstance configurations of stars, for Russell discrete emotions like fear are happenstance configurations of components. Stars are important phenomena in the science of astronomy, but astrological constellations are not. Likewise, according to Russell, components such as core affect, cognitive appraisal, and action tendencies are important phenomena in the science of psychology, but configurations of those components that happen to resemble folk concepts such as fear are not.

For scientific research programmes, however, a mechanistic model is not the only causal model available. Other causal models include interventionist models, causal cascade models, pathway models, and the causal network models used in systems theory (Ross and Woodward 2023). Mach's functionalist perspectives shared some similarities to the network model. Any one of these models could also be applied to the scientific study of emotions. The choice of a causal model would foreground the selection of specific causal features and would thus partly constitute the phenomena that scientists study.

For example, using a cascade model of causation, ruminating on negative emotions such as fear and anger can increase the intensity of those emotions, increasing the probability of behaviours such as avoidance and aggression. These behaviours influence the environment and how it is experienced. Such behaviour and environment pairings can initiate a positive feedback loop that amplifies the emotions, and—if ongoing—potentially alters the developmental trajectories of self-concept formation and relationship patterns in the long term (Selby and Joiner 2009). With this alternative model, emotions would have a role to play in causation-based science of psychology.

One difference between basic emotions and constellations is that the constellations are fully constituted as kinds by projections of concepts onto a random pattern in the night sky. Emotional episodes, however, may possess more internal coherence than constellations.

The theory of psychological construction can be taken to imply that the components independently co-occur, and some of those clusters of co-occurring events resemble folk emotion concepts. However, the relationships between the components are not necessarily arbitrary and random, even if the patterning of components is not the outcome of an identity-determining causal mechanism (or essence).

For example, Scherer's (2009) component process model is also constructionist. It describes emotions as being put together out of components and acknowledges that a near infinite number of arrangements of components are possible across individuals. However, for Scherer, these components, such as subjective feelings, cognitive appraisals, and action tendencies are more than independent events that occasionally co-occur. In his view, somewhat like in causal network models, the components can be synchronised as kinds by entering into recursive causal relationships with each other.

This type of causal coherence among components does not mean there are basic emotions in an essentialist sense, but the patterns may be coherent enough to form a kind that supports some generalisations. Within an emotion category such patterns could also vary and have fuzzy boundaries. In this version of a causation-based science of psychology, basic emotions would not be defined by affect programmes that are homologous across species, but they could be redefined as common variants as opposed to rare variants. That is to say, fear is a basic emotion because it is a frequently occurring pattern that is recognised as highly relevant for understanding ourselves and others.

10. Conclusions: Worldmaking, Mind-Framing, and Constituting

A philosopher who introduced his own notion of the workmanship of human understanding, Nelson Goodman (1978), wrote about "worldmaking." According to Goodman, the world is always going to contain more than can be represented by any of our concepts, thus we should not talk about some neutral world as viewed from what Putnam (1990) called a God's-eye view perspective; rather, what is available to us are multiple world versions.

Less metaphysically, we might say with Goodman that the world versions discovered by scientists are going to involve decisions about how to lump, split, order, and weight things. The decisions need not be arbitrary, but they are decisions, nevertheless.

Goodman's notion of world-making could be taken to imply making-up or inventing, whereas the simple naming of a pattern could imply passive recognition. Constituting, in its Machian selectionist form, potentially splits the difference between world-making and simple naming. It is neither as creative as making-up nor as passive as recognition. A better metaphor for a selectionist approach to constituting might be *engineering*.

References

Barrett, L. F. 2017. *How Emotions Are Made*. Houghton Mifflin.

Boring, E. G. 1923. “Intelligence as the Tests Test It.” *New Republic* 36: 35–37.

Bridgman, P. W. 1927. *The Logic of Modern Physics*. Macmillan.

———. 1945. “Some General Principles of Operational Analysis.” *Psychological Review* 52 (5): 246–49.

Carnap, R. (1956) 1991. “Empiricism, Semantics, and Ontology.” In *The Philosophy of Science*, edited by R. Boyd, P. Gasper, and J. D. Trout, 85–97. MIT Press.

Chang, H. 2022. *Realism for Realistic People*. Cambridge University Press.

Craver, C. F. and L. Darden. 2013. *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press.

Cronbach, L. J. and P. E. Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52 (4): 281–302.

Einstein, A. (1916) 1996. *The Collected Papers of Albert Einstein*. Volume 6 (English Translation Supplement), translated by A. Engel. Princeton University Press.

Ekman, P. 1993. “Facial Expression and Emotion.” *American Psychologist* 48 (4): 384–92.

Ekman, P. and D. Cordaro. 2011. “What Is Meant by Calling Emotions Basic.” *Emotion Review* 3 (4): 364–70.

Giere, R. N. 1999. *Science without Laws*. University of Chicago Press.

Goodman, N. 1978. *Ways of Worldmaking*. Hackett.

Kitcher, P. 2001. *Science, Truth, and Democracy*. Oxford University Press.

Kurth, C. 2019. “Are Emotions Psychological Constructions?” *Philosophy of Science* 86 (5): 1227–38.

LeDoux, J. E. 1996. *The Emotional Brain*. Simon & Schuster.

———. 2014. “Coming to Terms with Fear.” *Proceedings of the National Academy of Sciences* 111 (8): 2871.

LeDoux, J. E. and D. S. Pine. 2016. “Using Neuroscience to Help Understand Fear and Anxiety: A Two-System Framework.” *American Journal of Psychiatry* 173 (11): 1083–1093.

Locke, J. (1689) 1997. *An Essay Concerning Human Understanding*. Penguin Books.

MacCorquodale, K. and P. E. Meehl. 1948. “On a Distinction between Hypothetical Constructs and Intervening Variables.” *Psychological Review* 55 (2): 95–107.

Mach, E. 1914. *The Analysis of Sensations and the Relation of the Physical to the Psychical*. Translated from the 1st German edition by C. M. Williams, and revised and supplemented from the 5th German edition by S. Waterlow. Open Court.

Mackinnon, D. M., F. Waismann, and W. C. Kneale. 1945. “Symposium: Verifiability.” *Proceedings of the Aristotelian Society, Supplementary Volumes* 19: 101–164.

Makovec, D. 2019. “Introduction: Waismann’s Rocky Strata.” In *Friedrich Waismann: The Open Texture of Analytic Philosophy*, edited by D. Makovec and S. Shapiro, 1–25. Springer International.

———. 2025. “Open Texture in Science and Philosophy.” In *100 Years of “Tractatus Logico-Philosophicus”—70 Years after Wittgenstein’s Death: Proceedings of the 44th International Ludwig Wittgenstein Symposium*, edited by E. Heinrich-Ramharter, A. Pichler, and F. Stadler, 335–46. De Gruyter.

Makovec, D. and S. Shapiro, eds. 2019. *Friedrich Waismann: The Open Texture of Analytic Philosophy*. Palgrave Macmillan.

Pap, A. 1953. “Reduction Sentences and Open Concepts.” *Methodos* 5: 3–30.

Poincaré, H. (1905) 2001. “Science and Hypothesis.” In *The Value of Science: Essential Writings of Henri Poincaré*, edited by S. J. Gould, 1–178. The Modern Library.

Putnam, H. 1990. *Realism with a Human Face*. Harvard University Press.

———. 1996. “Irrealism and Deconstruction.” In *Starmaking: Realism, Anti-realism, and Irrealism*, edited by P. J. McCormick, 179–200. Bradford Books.

Ross, L. N. and J. Woodward. 2023. “Causal Approaches to Scientific Explanation.” In *The Stanford Encyclopedia of Philosophy*. Edited by E. N. Zalta. Spring 2023 edition. https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science/.

Russell, J. A. 1991. “In Defense of a Prototype Approach to Emotion Concepts.” *Journal of Personality and Social Psychology* 60 (1): 37–47.

———. 2003. “Core Affect and the Psychological Construction of Emotion.” *Psychological Review* 110 (1): 145–72.

———. 2012. “From a Psychological Constructionist Perspective.” In *Categorical versus Dimensional Models of Affect: A Seminar of the Theories of Panksepp and Russell*, edited by P. Zachar and R. D. Ellis, 79–118. John Benjamins.

Scarantino, A. and P. Griffiths. 2011. “Don’t Give up on Basic Emotions.” *Emotion Review* 3 (4): 444–54.

Scherer, K. R. 2009. “The Dynamic Architecture of Emotion: Evidence for the Component Process Model.” *Cognition and Emotion* 23 (7): 1307–351.

Selby, E. A. and T. E. Joiner, Jr. 2009. “Cascades of Emotion: The Emergence of Borderline Personality Disorder from Emotional and Behavioral Dysregulation.” *Review of General Psychology* 13 (3): 219.

Tal, E. 2020. “Measurement in Science.” In *The Stanford Encyclopedia of Philosophy*. Edited by E. N. Zalta. Fall 2020 edition. https://plato.stanford.edu/archives/fall2020/entries/measurement-science/.

van Fraassen, B. C. 2002. *The Empirical Stance*. Yale University Press.

van Loo, H. M. and J.-W. Romejin. 2015. “Psychiatric Comorbidity: Fact or Artifact.” *Theoretical Medicine and Bioethics* 36: 41–60.

Waismann, F. 1945. “Verifiability.” *Proceedings of the Aristotelian Society, Supplementary Volume* 19: 119–50.

Zachar, P. 2022. “The Psychological Construction of Emotion—A Non-Essentialist Philosophy of Science.” *Emotion Review* 14 (1): 3–14.

Received: 13 June 2024
Accepted: 19 June 2025
Published: 28 August 2025

Navigating the Waters of Emotion with a View Toward Cooperation

Cecilea Mun – North Park University, USA, cecileamun@icloud.com

Abstract

What emotions are is a central question in the science of emotion, and is often interpreted as a question about how one should define the theoretical term "emotion." It is also often interpreted as an invitation to understand the nature of emotions. In this paper, I demonstrate how these two interpretations can be related through an interdisciplinary, pluralist approach to the science of emotion. In doing so, I illustrate the limits of Klaus Scherer's proposed consensual, polythetic working definition with respect to some fundamental concerns in the science of emotion. I use the meta-semantic taxonomy of theories of emotion to argue that once psychological theories of emotion are more clearly delineated according to two fundamental concerns—the metaphysical and the meta-semantic—one can not only trace the logical implications of these fundamental concerns to their implications for the design of empirical research, but one can also more clearly understand that the pursuit of knowledge in the science of emotion requires both those who take emotions to be stars, as well as those who take them to be constellations.

Keywords: emotions, language, ontology, psychology, philosophy, science, unification

1. Introduction

Approximately twenty years ago, Klaus Scherer (2000) and James Russell (2003) argued for the convergence of emotion theories on a consensual definition of emotion. Approximately ten years later, a special section on defining emotions was published in *Emotion Review* (2010, volume 2, no. 4), followed by a second special section, edited by Russell (2012, volume 4, no. 4). More recently, Scherer observed that such efforts were unsuccessful, yet he also expressed some optimism about current and future efforts toward convergence: "While [previous] suggestions had little success in bringing about convergence in the last 20 years, now might be the time to start a concerted effort towards theory convergence in emotion science" (2022, 165). To this end, Scherer draws on Fiona Hibberd's (2019) suggestion regarding the productivity of a polythetic definition of emotion for the aim of eventual unification, and Gerd Gigerenzer's (2017) broad outline for a two-stage process of theory integration, to propose 1) a consensual, polythetic working definition of emotion and model, and 2) possible subsequent steps toward theory integration. One major aim in arguing for such a proposal is to "encourage the design of critical empirical studies to examine the proposed mechanisms" (Scherer 2022, 155). A second aim is "to encourage an unbiased discussion of the similarities and differences between concepts and mechanisms proposed by different theories, rather than theory integration or finding consensus at a low level" (165). These two aims, to some extent, echo Rainer Reisenzein's (2022) call for more theoretical psychology.

Specifically, Scherer argues that basic emotion theories, appraisal theories, social constructivist theories, and psychological constructionist theories share the following assumptions in common:

> That emotions 1) consist of an episodic process in response to a perceived event or situation of major significance, 2) which is characterized by recursive causal effects (forward and backwards) between several components that include the evaluation of the event in terms of its significance for the goals and values of the individual, 3) creating physiological reactions, motor expressions, and action tendencies and 4) that this process is partially accessible to consciousness, resulting in feelings that 5) can be categorized and subsequently labelled by the individual in terms of its subjective conceptual structure. (2022, 164)

Given these observations, Scherer proposes the following consensual, polythetic, working definition:

> The nature of an emotion can be summarised as follows: Individuals are exposed to stimuli, events, or situations generating an "overwhelming idea" of personal significance and potentially requiring some kind of action. This sets off a parallel, multi-level, and recursive process to determine, form, or construct the nature of this reaction. The first stage after elicitation is a subjective analysis and evaluation of eliciting stimulus/event/situation in terms of its consequences, implications, and requirements, involving a variety of attribution and appraisal mechanisms. The result of this evaluation process, which generally runs through several recursive cycles involving interactions between criteria, produces a synchronised effect on action tendencies and autonomic and somatic responses (including expressions). These in turn, are also evaluated by the appraisal system in the context of the results of situation evaluation until some degree of closure is achieved. The constantly changing evaluation results in the form of continuously updated schemata, which constitute the unconscious representation of an integrated feeling (*qualia*) of the emotional experience (core affect). Parts of this feeling representation can enter [consciousness] (Scherer, 2005b) and, especially when social communication of the felt experience seems desirable, give rise to categorization and eventually verbal labeling. (164–65)

Scherer also highlights the import of social, cultural, and historical factors that can shape an emotional episode, with the examples of valence appraisals, normative assessments, and the categorisation and labelling of emotional episodes (165). Furthermore, Scherer utilises his component process model of emotion as a means to integrate basic emotion theories, appraisal theories, social functional theories, social constructivist theories, and psychological constructionist theories (157–59).

I applaud Scherer's and Reisenzein's efforts to further advance theoretical research in the science of emotion, including interdisciplinary research, and I offer this article as a contribution toward such efforts. I also hope to encourage further discourse on the interdisciplinary concerns raised by Scherer, not only across disciplines, but also across publication venues, by responding to Scherer's call in this special issue of *Passion*. Scherer's proposal for a consensual, polythetic definition might seem to stop short of addressing the thematic question for this special issue: "Emotions—More Like Stars or Constellations?"[1] Yet, Scherer's

[1] The thematic question of this special issue may unwittingly lead to a lack of clarity or error in some responses. As I understand the question, the intent is to get at the debate between essentialist and non-essentialist theories of emotion by comparing essentialist positions to those that take emotions to be like stars, whereas non-essentialist positions are likened to those that take emotions to be more like constellations. One significant problem with this metaphor is that stellar taxonomy might more accurately be characterised as being drawn on epistemic considerations rather than any essentialist principles (see Ruphy 2010). To suggest that emotions are more like stars might then be, more accurately, to agree with a non-essentialist's position like de Sousa's (1984), which makes the question a non-starter. So, for the sake of the intent of this special issue, let us grant poetic license to the metaphor and assume that stars represent an essentialist position about what emotions are, whereas constellations represent a non-essentialist position.

proposal can be understood as a stepping-stone to addressing the concerns raised by this question. Granting that Scherer's proposed consensual definition, along with his proposed model, provides at least a starting point for the convergence and integration of basic emotion theories, appraisal theories, social constructivist (constructionist) theories, and psychological constructionist theories, it is also important to understand the potential limitations of this proposal, including for the unification of the science of emotion, as well as for the purpose of informing the design of empirical studies.

One approach to drawing out the implications of Scherer's proposal for a consensual, polythetic definition of emotion is to broaden our perspective on the categories of theories of emotion which we are willing to consider for our endeavours. For example, consider the categories of theories of emotion I proposed in my earlier works (2014; 2016; 2021). In these works, I recounted the contemporary landscape of theories of emotion from an interdisciplinary perspective, with a special emphasis on the disciplines of philosophy and psychology. I noted the similarities and differences between major theories of emotion within philosophical perspectives (e.g., cognitive theories, non-cognitive theories, social constructionist theories, prototype theories, and embodied perceptual theories) and psychological perspectives (e.g., basic emotion theories, appraisal theories, social constructivist theories, and psychological constructionist theories). By mapping out these various theories along the lines of two fundamental dimensions—a metaphysical dimension concerning the ontology of what emotions are, and a meta-semantic dimension concerning the significance of ordinary language to the science of emotion—I argued for fundamental distinctions between what I referred to as "realist theories," "instrumentalist theories," "eliminative-realist theories," and "eliminativist theories." These categories of theories of emotion constitute the current meta-semantic taxonomy of theories of emotion.[2] Given that Scherer's taxonomy of psychological categories of theories of emotion are subsumed by the meta-semantic taxonomy of theories of emotion, this taxonomy allows us to identify some limitations to Scherer's proposal, especially when the primary focus of concern is the unification of the science of emotion.

2. Fundamental Differences Between Theories of Emotion

The question about what the word "emotion" refers to can be understood as a question about whether the word "emotion" refers to a category of members that can possibly be independent of human conceptualisations (i.e., concrete things), members that are necessarily dependent on human conceptualisations (e.g., imaginations), or perhaps a mixture of both (e.g., artifacts), which would make that emotion category also necessarily dependent on human conceptualisations. The most important factor to attend to here, however, is the unifying principle, which is the principle that unifies the members of the emotion category into a single category and is thought to define that category.

Before we can determine exactly what kind of unifying principle is applicable, one must first understand what it means for the members that make up the *emotion* category to be concrete, imagined, or a combination of concrete and imagined particulars. In the first sense, the members of the category could possibly be mind-independent, and in the last two senses, they would necessarily be mind-dependent. In the first case, we can

[2] Although there are various alternative taxonomies of theories of emotion that one can appeal to for research purposes in the science of emotion (e.g., Griffiths 1997; Prinz 2004, Scarantino and de Sousa 2018; and Moors 2022), and each of these taxonomies serve their intended purposes, the uniqueness and the value of the meta-semantic taxonomy of theories of emotion is that, unlike these taxonomies of theories of emotion, the meta-semantic taxonomy of theories of emotion categorises theories in terms of fundamental differences such that each category is mutually exclusive with the others, except for very narrow, vague, borderline cases, and so are able to sufficiently track fundamental shifts in a theory.

refer to such particulars as objective kinds. In the two later cases, we can refer to such particulars as subjective kinds. The commitment to objective kinds versus subjective kinds can also be understood as a distinction that demarcates a realist theory of kinds from an anti-realist theory of kinds.[3] Such anti-realist theories of emotion can also be generally characterised as a kind of nominalism about emotion, which is a position that takes emotion words to be mere labels for categories.[4] The kind of realism and nominalism noted here can also be respectively related to the categories of essentialism and non-essentialism (see Hibberd 2019; Zachar 2022). Essentialist theories of emotion are theories of emotion that regard the referent of "emotion" to be an *objective kind* category. Non-essentialist theories, in contrast, regard the referent to be a *subjective kind* category. Furthermore, granting that *stars* metaphorically represent an essentialist position, we can also understand objective kind theorists to take emotions to be more like *stars*, whereas subjective kind theorists take emotions to be more like *constellations*.

Another related concern is the concern with the relationship between ordinary language and the technical language that experts use. For example, consider the word "water" and the word "H_2O." The first is considered an *ordinary language* word. We use the word in our everyday lives when we talk about the thing to which "water" refers. But experts, such as scientists and philosophers, might talk instead of "H_2O" rather than "water," and there are often questions as to whether experts and ordinary people are referring to the same category of things when they speak of "water."

Some emotion experts believe that their technical use of the word "emotion" refers to the same category of things to which the ordinary language use of the word "emotion" refers.[5] Consequently, they believe that they are speaking of the same kind of thing that ordinary people are speaking about when they speak of "emotion" or "emotions," and they believe that the work of the sciences is to help correct and refine ordinary language meanings. They believe that their scientific, technical term, which is shared with ordinary language, is a *trans-theoretical term* (i.e., it can be shared across theories, including ordinary language as a kind of theory).[6] Such an expert can be referred to as an *optimist about ordinary language emotion words*. They are optimistic about the intended referential meanings of ordinary language emotion words. They believe that appropriate ordinary and scientific uses of the word "emotion" share the same referent. Those who disagree with this presupposition can be referred to as *pessimists about ordinary language emotion words*. Pessimists believe that the referents of ordinary language emotion words are so different from what they are speaking about that they believe that ordinary people are referring to a different category of things compared to when they use the emotion words of their theoretical language.

Given the example I used about "water" and "H_2O," you might be more willing to side with the pessimists about the word "emotion" because it seems clear to you that *water* is not *H_2O*, and especially if by "H_2O" we mean *pure H_2O* and *water* is not *pure H_2O*. But now consider the ordinary English word "electricity." Does the ordinary English word "electricity" refer to the same thing as the technical word "electricity," when it is used by a physicist, engineer, or philosopher? This might be a trickier question to answer, and perhaps you would be more willing to side with the optimist when thinking about the word "electricity." This point about the relationship between ordinary language words and technical scientific words can also be related to questions about the meanings

[3] See Tuomas Tahko 2021 for an interesting discussion of realism and anti-realism, as well as the criteria of mind-independence for characterising natural kinds.

[4] For an interesting discussion of a trope nominalist theory of natural kinds, see Markku Keinänen 2015.

[5] For a discussion of the relationship between expert and ordinary knowledge, see Fodor 1994. Hibberd (2019) also mentions this fact.

[6] This term was coined by Hilary Putnam (1973).

of words across time, within the context of ordinary or scientific languages, and even about a word and its translation into a different language. For example, consider the word "whale." There was a time, long ago, when people (including experts) thought whales were fish. We now know that whales are mammals. In either the context of comparing only the meanings of the ordinary English word "whale" across time, or comparing only the meanings of the technical scientific word "*Cetacea*" across time, would you say that ordinary people or scientists were referring to the same category of things from one time to another?

If we consider whether an emotion expert would regard emotions to be an objective kind or a subjective kind, and we add to this what an emotion expert believes about the relationship between their technical, scientific emotion words and ordinary language emotion words, then we can construct a matrix with a total of four different fundamental categories of expert perspectives on (or frameworks about) what emotions are: *realism*, *instrumentalism*, *eliminative-realism*, and *eliminativism* (Mun 2021).

Realists are optimists about ordinary language emotion words, and believe that the word "emotion" refers to an objective kind category. Specifically, they believe that our experiences of different emotions (typically) share certain fundamental features, such as being products of an evolutionarily evolved system in which our mental processes (e.g., perceptions, thoughts, beliefs, and desires) are related to our physiological processes (e.g., the activation of our autonomic nervous system and motor cortex) in such a way that allows us to address challenges in the world for the purpose of survival. To this extent, one might say that realists believe that beings that have emotions have an *emotion system*. When thinking about an emotion system, you might analogously consider how beings that have perceptions have a perceptual system or how beings that digest food have a digestive system. To a lesser extent, one might say that the *emotion* category is like the *cat* category, or is like the category of *stars*: the members of the category are necessarily concrete particulars. I will further elaborate on this point in §5. For examples in psychology, see Paul Ekman's (2003) basic emotion theory, Richard Lazarus's (1991) appraisal theory, and Klaus Scherer's (1982; 1987; 2001; 2005; 2009; 2012; 2019) component process theory. For examples in philosophy, see Jesse J. Prinz's (2004) embodied appraisal theory of emotion.

Instrumentalists are also optimists about ordinary language emotion words, but they believe that the word "emotion" refers to a subjective kind category. To this extent, instrumentalist theories are logically contrary to realist theories: they can both be false at the same time, but only one can be true in relation to the other. Instrumentalists agree that ordinary people and experts are all speaking of the same kind of things when we use shared emotion words, but they deny that emotional beings have an evolutionarily evolved emotion system. They believe that emotions are instead products of different systems (e.g., the perceptual system, conative system, doxastic system, motor system, autonomic nervous system, etc.) working independently to respond to challenges in the world, without the need for an overarching emotion system. Consequently, emotion words, such as "emotion," refer to a category in which the members of that category are ultimately dependent on human conceptualisations (i.e., arbitrary ways in which human beings think about emotions) since it is these ways of thinking about emotions that are ultimately the basis on which emotional experiences are categorised. In short, instrumentalists believe that the members of the *emotion* category are something akin to imaginative projections (i.e., they are social constructions), or are more like *constellations*. For examples in psychology, see James Averill's (1980; 1986; 1997; 2004) social constructivist theory of emotion and Lisa Barrett's (2017a; 2017b) conceptual act theory. For philosophical examples, see Ronald de Sousa's (1987) theory of patterns of salience, Martha Nussbaum's (2001; 2016) theory of eudaimonistic assent, and Claire Armon-Jones's (1986) social constructionist theory of emotion.

The opposite, or logical contradictory, of instrumentalism is *eliminative-realism*. Whereas an instrumentalist is a subjective-kind theorist who is also an optimist about ordinary language emotion words, an eliminative-realist is an objective-kind theorist who is also a pessimist about ordinary language emotion words. Eliminative-realists believe that emotions can at least be categorised into distinct kinds of emotion systems (e.g., the panic system, fear system, rage system, seeking system, lust system, care system, and play system), if not an overarching emotion system. So, they believe that at least certain emotion types are objective kinds. They, like realists, believe that emotions are like *cats* or *stars*. Yet, they deny that the emotion words that experts use refer to the same things as the emotion words used by ordinary people. So, they are pessimists about ordinary language emotion words. One interesting consequence is that, although eliminative-realists believe that the *emotion* category is ultimately grounded in members that are more like *cats* or *stars* than *unicorns* or *constellations*, eliminative-realists also believe that the ordinary language word "emotion" does not refer to the same kind of thing that they are referring to when they speak of emotions. Consequently, instrumentalism and eliminative-realism have opposing truth-values: if instrumentalism is true, then eliminative-realism must be false, and *vice versa*; they cannot both be true or both be false at the same time. They are also logically contrary to realist theories, as well as eliminativist theories of emotion, but for different reasons. As with basic emotion theories, they are objective kind theories; as with eliminativist theories, they are pessimistic about ordinary language. Jaak Panksepp's (1998; 2008) basic emotion theory is a psychological eliminative-realist theory, and Paul Griffiths (1997) and Scarantino (2012; Scarantino and Griffiths 2011) provide philosophical examples.

Eliminativist theories are contradictory to realist theories. Eliminativists are pessimists about ordinary language emotion words, and they believe that how emotional experiences get categorised ultimately depends on human conceptualisations. Their reason for being pessimists about ordinary language emotion words are the same as the eliminative-realist's: they believe that ordinary language uses of emotion words are too imprecise to be useful for scientific endeavours. Their reason for believing that emotions are subjective kinds is the same as the instrumentalist's. They don't believe that there is any such thing as an emotion system or different types of emotion systems. In other words, eliminativists believe that the *emotion* category is constituted by members that are more like imaginative projections (or *constellations*) than *cats* (or *stars*). As such, eliminativist theories are also contrary to eliminative-realist theories, as well as instrumentalist theories, but for different reasons. James Russell's (2003; 2009; 2010; 2012) psychological constructionist theory is a psychological eliminativist theory, and I will later argue that George Mandler's ([1975] 1984) psychological social constructivist theory is also an exemplar of eliminative theories of emotion.[7, 8]

[7] Mandler's social constructivist theory, however, might be most accurately characterised as being positioned in the vague area between instrumentalism and eliminativism, and this is mostly a consequence of his pessimism about ordinary language.

[8] As far as I am aware, there are no eliminativist philosophical theories of emotion. I suspect that this is mostly because such a position is especially fruitful for empirical approaches that aim to falsify alternative theories, and especially difficult for philosophical approaches that take a more logical, argumentative approach. I can imagine the development of such a philosophical position, however, especially if one aligns the primary aim of such an endeavour with the critical aim of eliminativist psychological approaches, although the success of such an endeavour might require an especially keen philosopher. If they succeed, however, their impact would be immensely significant. One especially worth-while initial pursuit for such an endeavour might be to draw out the details of how such a critical aim might be pursued, especially given the fact that eliminativist theories do not share the same object of inquiry as the alternative theories which they would be aiming to falsify. I address how falsification might occur between realists and instrumentalists in §4; yet I do not similarly address how eliminativists might do so. Psychological eliminativist theorists have also sought to falsify realist approaches (e.g., DiGirolamo and Russell 2017). In any case, I will leave further considerations on these concerns for some future time.

3. Implications of the Meta-Semantic Taxonomy of Theories of Emotion

Given these four fundamental categories of theories of emotion, one can observe that the psychological taxa of theories of emotion discussed by Scherer (2022)—basic emotion theories, appraisal theories, social constructivist (constructionist) theories, and psychological constructionist theories—are subsumed by the meta-semantic taxonomy of theories of emotion. Both basic emotion theories (such as Ekman's) and appraisal theories (such as Lazarus's and Scherer's) are realist theories. Social constructivist theories (such as Averill's) and some psychological constructionist theories (such as Barrett's) are instrumentalist theories, whereas other psychological constructionist theories (such as Russell's) are eliminativist theories. Given these categorisations, according to the meta-semantic taxonomy, there are certain logical relations between these theories that would demarcate at least some limits to moving beyond the convergence on Scherer's proposed consensual, polythetic, working definition and component process model. To more clearly understand what these limits are, I presuppose the acceptance of Scherer's proposed definition and model, as a means for integrating basic emotion theories, appraisal theories, social constructivist theories, and psychological constructionist theories. I then explore at least some of the possible limits for full theory integration, and therefore the unification of theories of emotion.

As realist theories, there should be no problem with the full integration of basic emotion theories and appraisal theories, since these two psychological categories of theories of emotion are fundamentally the same. To some extent, one can trace the contemporary history of the integration of these two kinds of psychological theories through the history of the convergence between Ekman's basic emotion theory and Lazarus's cognitive-relational theory (see Lazarus 1990; Ekman 1999), as well as to Stanley Schacter and Jerome Singer's (1962) contributions, and the debates between Lazarus, Mandler, and Robert Zajonc (see Bozinovski 2018; Mandler 1990). The debates which led to the convergence between these two kinds of theories can be summarised as concerning the necessity of cognitive, non-automatic appraisals as aspects of the elicitation of emotions. The result of the convergence can be understood in terms of the acceptance of at least two kinds of elicitation processes by both kinds of frameworks. Given the foregoing discussion, to the extent that Scherer's component process model is a model for his appraisal theory of emotion, it should not be a surprise that basic emotion theories and appraisal theories are generally consistent with Scherer's proposed consensual definition and component process model (Scherer 2022, 161–62). One noteworthy observation is that such a convergence did not require an agreed upon, consensual, working definition, including an operational one, that was shared between the two kinds of theories, but instead a shared object of inquiry. As realist theories, both basic emotion theories and appraisal theories can be understood as ultimately sharing the same referent for the theoretical term "emotion." Such a condition is secured by the shared assumptions that emotions constitute an objective kind and the commitment that the language of their science shares the same referents as the referents of ordinary language emotion words (i.e., optimism about ordinary language).

The convergence of social constructivist theories on Scherer's consensual definition and model may also seem unproblematic. Scherer notes Averill's suggested definition of emotions as socially constructed syndromes or transitory roles, and he believes that such accounts might unproblematically converge with his proposed definition and model. He observes that such an account has been further developed in Dacher Keltner and Jonathan Haidt's (1999) social functionalist account, and that the focus of Mandler's (1990) social constructivist account is the role of cognitive schemata in determining an emotional experience (2022, 162). These theories can be related to the multi-level appraisal aspect of Scherer's proposed consensual definition and model, or they can be regarded as focusing primarily on the dynamic social aspects of emotion, which are essentially external to a subject's emotional experience, although these external factors and relations would be closely related (as with Mandler's, and Keltner and Haidt's accounts). Let us consider, however, the extent to which social constructivist theories can be integrated with basic emotion theories and appraisal theories, beyond their convergence on Scherer's proposed consensual definition and model.

The ease with which Keltner and Haidt's social functional account can converge with Scherer's proposed consensual definition and model is predicted by the meta-semantic taxonomy of theories of emotion. This is because, although Keltner and Haidt's social functional account is regarded as a social constructivist account, it is also a realist account. This is primarily because Keltner and Haidt's account also accept that emotions are objective kinds, and that ordinary language emotion words are transtheoretical with the emotion words of their theoretical language. Thus, for social functional accounts like Keltner and Haidt's, theory integration can move beyond Scherer's proposed consensual definition and model. In short, as with basic emotion theories and appraisal theories, full integration of these theories should not be problematic. A problem arises, however, when we consider the degree to which Averill and Mandler's accounts can be fully integrated with Scherer's component process theory, beyond convergence on Scherer's proposed consensual definition and model.

Although one might regard Averill's, Mandler's, and Keltner and Haidt's theories as constructivist (or constructionist) theories (Scherer 2022, 162), in accordance with their psychological taxa, one ought to note that while Keltner and Haidt's theory accepts the assumption that "emotions are thought of as relatively automatic, and rapid responses" (1999, 508), both Averill and Mandler's accounts reject this Jamesian premise. As Averill explains,

> In cognitive terms, emotions may be conceived of as belief systems or schemas that guide the appraisal of situations, the organization of responses and self-monitoring (interpretation) of behaviour. When conceived in this way, the question arises, What is the source of emotional schemas? The more traditional answer to this question is that emotional schemas became hardwired into the nervous system during the course of evolution—that they represent innate affect programmes (Izard, 1977; Tomkins, 1981). By contrast, a constructivist view assumes that emotional schemas are the internal representation of social norms or rules. (1986, 100)

Mandler notes that James's "particular constructions depended entirely on patterns of visceral and muscular feedback and are no longer found acceptable" (1990, 22; see also Mandler [1975] 1984).

The rejection of this Jamesian principle sets Averill's social constructivist theory apart from basic emotion and appraisal theories, although, like basic emotion and appraisal theories, his theory is optimistic about ordinary language emotion words (Mun 2021, 45–48). This makes Averill's theory an instrumentalist theory, rather than a realist theory, according to the meta-semantic taxonomy of theories of emotion. Social constructivist theories such as Averill's are, therefore, logically contrary to appraisal theories (such as Scherer's component process theory), as indicated by the meta-semantic taxonomy. Thus, full integration between these kinds of social constructivist theories (e.g., Averill's) and appraisal theories (e.g., Scherer's and Lazarus's), and other social constructionist theories (e.g., Keltner and Haidt's), would not be possible without a fundamental change in at least one of the theories.

This conclusion is an implication of the logical relations that hold between realist and instrumentalist theories, which place rational constraints on the design of empirical research. That both realist theories and instrumentalist theories may converge on Scherer's consensual definition and model speaks to the possibility that both realist theories and instrumentalist theories may be false. Thus, although empirical research based on Scherer's proposed model would be fruitful, especially in possibly falsifying both realist and instrumentalist theories, such research would not get at concerns regarding the fundamental nature of emotion on which realists and instrumentalists disagree. When theory integration is pursued beyond Scherer's consensual definition and model, one should observe that appraisal theories (along with basic emotion theories and other realist theories, such as Keltner and Haidt's) would ultimately be at odds with social constructivist theories (such as Averill's).

One way to draw out the logical constraints on empirical research is to consider Kristen Lindquist et al.'s (2022) social constructivist (constructionist) approach. Lindquist et al. present a cultural evolutionary framework that takes emotions to be cultural artifacts that evolved through social transmission within and between groups, which are underpinned by neurological mechanisms linked to physiological and action regulation (Lindquist et al. 2022, 670).[9] One consequence of Lindquist et al.'s framework is that it predominantly focuses on empirical research on adult subjects, including in the area of emotion concepts (e.g., Satpute et al. 2016; Brooks et al. 2019), while also admitting that linguistic differences may not map on to experiential differences (672). Lindquist et al.'s framework, which is an instrumentalist framework, is fairly consistent with realist frameworks with respect to adult human emotions. That this is the case can be explained by the fact that realist theories do not necessarily deny that emotions are socially constructed. This is also consistent with the logical implication that the theories are contraries: both realist and instrumentalist theories can be wrong about the social construction of emotions.

Where Lindquist et al.'s theory diverges from realist theories is that it claims that emotions are *only* artifacts. As with other instrumentalist views, they reject anything like an emotion system, including for specific emotion types. Accordingly, one might be able to adjudicate between realist theories (such as Ekman's, Scherer's, and Keltner and Haidt's) and instrumentalist theories (such as Lindquist et al.'s and Averill's) by focusing on identifying possible innate aspects of an emotion system in infancy, as well as their developmental trajectory into adult emotions. One place to start is with research on infant emotion perceptions (see Nelson and Leppänen 2009). I will not go into detail here, especially since there is a paucity of empirical research on infant emotion perception, including in cross-cultural research, but one might consider Carlijn van den Boomen et al.'s (2019) results, which found that infants between 9–10 months are better able to discriminate happy faces from angry and neutral faces.

If one assumes that an emotion perception system is a significant aspect of an emotion system, and that system comes online in early infancy (after approximately 7 months), then an explanation as to why infants might first develop the ability to perceive positive emotions from angry or neutral emotions seems forthcoming: the perception of positive emotions is necessary for engendering social bonds (such as attachment, trust, etc.), which are necessary preconditions for socialisation (including norm and concept acquisition), and are therefore secured by evolution. If emotions are entirely socially constructed as artifacts and require the input of emotion concepts, the question remains how infants between 9–10 months can discriminate positive from negative and neutral emotions *at all.*

Although, an explanation for why they might initially discriminate positive from negative and neutral emotions might seem to be forthcoming: instrumentalists might conjecture that the rate of exposure to positive versus negative emotion faces would explain this difference. Yet, a problem arises when one considers

[9] A cultural evolutionary framework for emotion need not, however, be an instrumentalist approach. For a sketch of a realist approach, see Mun 2022.

that such discrimination can occur as early as 9–10 months.[10] What explains such discerning abilities at this age if not an innate, evolutionarily evolved system? Thus, although instrumental social constructivist theories share some overlapping considerations with realist theories, such as basic emotion theories and appraisal theories, they are fundamentally distinct kinds of theories. Accordingly, full theory integration between these theories would not be possible without a fundamental theory change in at least one of the theories. A similar conclusion can be drawn for Barrett's conceptual act theory (2017a; 2017b). Although Barrett's theory is regarded as a psychological constructionist theory, according to the psychological taxa of theories of emotion it is an instrumentalist theory (see Mun 2021 for a detailed explanation).

Russell's psychological constructionist theory, unlike Barrett's theory, is an eliminative theory. As such, it is logically contrary to instrumentalist theories, such as Barrett's, Lindquist et al.'s, and Averill's (see Mun 2021 for a detailed explanation; see also Zachar 2022). Although Russell's theory will overlap with instrumentalist theories, especially with respect to claims about the subjective kind nature of emotional experiences, Russell's theory also denies the existence of emotions. According to Russell, emotions are not simply something like imagined projections, but they nonetheless do not constitute a legitimate psychological category for scientific research. What *do* constitute psychologically legitimate categories, according to Russell, are the physiological (core affect), behavioural (actions/reflexes), and narrative (scripts/concepts) aspects that constitute what Russell refers to as an emotional episode or a meta-emotional experience. These are the elements of Russell's psychological constructionist theory, as an eliminativist theory, that overlap with instrumentalist theories such as Barrett's, Lindquest et al.'s, and Averill's. Thus, along the lines of these ontological concerns, both kinds of theories may be proven false.

Russell's theory, however, is also pessimistic about ordinary language emotion words. It takes ordinary language to misidentify the referents of ordinary language emotion words, and as such severs the connection between his theoretical language of emotion and ordinary language emotion words, unlike instrumentalist theories. The result is a rejection of the emotion category as a legitimate category for scientific study, which stands in opposition to the instrumentalist's claim that the emotion category constitutes a legitimate category for scientific research. Thus, as with realist and instrumentalist theories, instrumentalist and eliminativist theories can both be simultaneously proven false, yet only one of the two kinds can be true. A similar conclusion can also be drawn about Mandler's social constructivist theory ([1975] 1984). As with Russell's psychological constructionist theory, Mandler's social constructivist theory is pessimistic about ordinary language emotion words—in contrast to other social constructivist theories (e.g., Averill's)[11]—and, also like Russell's psychological constructionist theory, it holds that emotions are subjective kinds (similar to instrumentalist theories). As such, Mandler's social constructivist theory is more closely aligned with psychological constructionist theories like Russell's, rather than those like Barrett's and other social constructivist theories (such as Averill's).

[10] Such questions might be resolved through additional empirical research, including cross-cultural research comparing infants' abilities to discriminate between positive and negative emotion between cultures that encourage smiling and those that do not (see Krys et al. 2016, although in such research the rate at which people smile at infants should be specifically considered).

[11] In chapter one, in the section titled "Psychology and Common Language," Mandler discusses the import of ordinary language to the science of emotion. One might conclude from parts of this discussion that Mandler's theory ought to be regarded as an instrumentalist theory. Yet the parts of his discussion which suggest this correspond more closely with what Mun referred to as the *fundamental base for interdisciplinary inquiry in the science of emotion* (Base^e), which is a fundamental principle to which any adequate theory of emotion should adhere (2021, ch. 4). Other parts of Mandler's discussion, however, make clear that he is a pessimist about ordinary language emotion words: "Those who have looked to ordinary language as the royal road to developing a satisfactory scientific language (both syntactically and semantically) often fail to apply a fundamental distinction in the primary *function* of these two languages" (Mandler [1975] 1984, 7).

One way to draw out the implications of these logical constraints on empirical design is to consider the phenomenon of pretence emotions (see Mun 2021). For example, one might consider the implication of emerging empirical research on real and fake emotions (e.g., Saxen et al. 2017; Jia et al. 2021).[12] Granting that pretence emotions may constitute a complex category of experiences—requiring the acceptance of degrees between real and fictitious emotions (Pugmire 1994), as well as the teasing apart of emotional responses to fictions (Radford and Weston 1975) and genuine emotional responses—empirical research on the neural–physiological differences between genuine and pretence emotions might lead to further progress for realist theories. Russell's eliminativist theory also does not provide any metaphysically grounded theoretical constraints against accepting pretence emotions as instances of emotion. One can also say something similar about Mandler's social constructivist account.[13] So, it might be difficult for Russell's eliminativist theory to make sense of the results of empirical research which aims to distinguish real emotions from pretence emotions.

One might also conclude that instrumentalists would have an equally difficult time as eliminativists in making sense of pretence emotions. As with eliminativist theories, which take emotions to be subjective kinds, instrumentalists also regard emotions as something like imagined projections. Thus, they would have an equally difficult time making sense of the distinction between genuine and pretence emotions, which is currently being borne out by emerging research. Unlike eliminativists, however, instrumentalists can fall back on ordinary language, given their optimism about ordinary language, to at least restrict the instances of the category under study. According to ordinary language, pretence emotions are not genuine cases of emotions. So, such experiences can be readily rejected by instrumentalists as constituting members of the emotion category. In contrast, realist theories would have the most resources to make sense of this distinction, once concerns about vagueness are bracketed for later consideration.[14] What distinguishes a real or genuine emotion from a non-genuine pretence emotion would be the activation of the emotion system, whatever it may be.

Another way to draw out the logical constraints on empirical design between realist and eliminativist theories is to consider the difference between Mandler's appeal to likings in support of his rejection of the Jamesian premise that emotions are discrete patterns of behaviour, experience, and neural activity. Mandler relies primarily on observations that automatic "affective" reactions are typically slower than cognitive reactions to argue against the claim that emotions can bypass cognitions (1990, 34–38). Yet, if we consider Scherer's Table 1,

[12] On a tangential note, it might also be important for researchers to consider the ethical implications of future technology that may be developed on the basis of such empirical research. Consider, for example, the warning Barrett conveyed regarding the reliance on emotional responses in the U.S. judicial system (Barrett 2017a), and the potential detrimental effect that such technology might have on marginalised populations due to either coding bias or an undiagnosed affective disability.

[13] Mandler does on occasion suggest that behaviours—which may be the kind of category under which pretence behaviours might be placed—ought not to be taken as emotions: "The behavior itself, is not to be considered emotional within the context of this model" ([1975] 1984, 121). As Mandler continues, he elaborates by noting that it is the "cognitive evaluations, which in turn determine the phenomenal experience of emotion" (121), and that "some cognitive interpretation of the environment produces arousal, and the perception of that arousal together with some cognition of the situation generates emotional experience" (123). Arousal is therefore a necessary condition, yet not a sufficient one, for an emotional experience, and arousal and cognitive evaluation are jointly necessary and sufficient. An emotion system, however, is not posited to causally connect relevant arousals to relevant cognitive evaluations, and it is this causal untethering which leaves open the possibility of Mandler's theory admitting pretence emotions as instances of emotion. (One should also note the striking similarities between Mandler's theory and Russell's at this point.)

[14] Vagueness, although interesting and perhaps also unavoidable for the science of emotion—not only given that any science is ultimately a human endeavour and the emotions that are of central concern are ultimately human experiences—need not be of too much concern for the time being since, even granting vagueness, there are in fact clearly identifiable cases of genuine emotions and non-genuine, pretense emotions (such as those induced by method acting).

of the "design feature delimitation of different affective states/dispositions," likings are not regarded as emotions. They are instead categorically distinguished from emotions, as attitudes. These facts demonstrate how realist theories are logically contradictory with eliminativist theories. Realist and eliminativist theories have opposing truth-values: when one is true the other must be false. Accordingly, the full integration of theories, beyond a convergence on Scherer's consensual definition and model, would not be possible between realist theories and eliminativist theories, without at least some fundamental change in one of these theories.

4. A Pluralist Approach to a Unified Science of Emotion

I introduced the meta-semantic taxonomy of theories of emotion, explained how it can be related to aspects of the psychological taxonomy of theories of emotion, and illustrated how the meta-semantic taxonomy can clarify the fundamental differences between theories of emotion beyond the discussed aspects of the psychological taxonomy of theories of emotion. I also demonstrated how the logical relations between theories of emotion implied by the meta-semantic taxonomy lead to substantive constraints on empirical design, especially when one considers the possibility of theory integration beyond Scherer's proposed consensual definition. Metaphysical claims about whether emotions are objective or subjective kinds, and claims about the import of ordinary language emotion words in the science of emotion, not only have significant theoretical implications, but these implications can be traced to questions about the constitution of the object of study, which ultimately informs the design of empirical research.

Although vagueness remains about exactly which instances of experience constitute a realist's (or any other theoretical kind's) category of emotion,[15] one can delineate a realist's object of study from an instrumentalist's object of study (e.g., a category admitting instances of human infant emotions versus one that does not). Consequently, instrumentalist theories may lack the ontological resources to adequately explain phenomena associated with infant emotions (e.g., infant emotion recognition). An eliminativist's object of study can also be delineated from a realist's and instrumentalist's object of study to the extent that the eliminativist's category of *emotion* (or more accurately, emotional episodes) would admit instances that both a realist and instrumentalist would have the resources to reject as appropriate instances of their object of study (e.g., pretence emotions and likings[16]).

We were also able to observe the differential effects of the fundamental metaphysical claims regarding whether emotions are an objective or subjective kind, and the fundamental meta-semantic claims about whether ordinary language emotion words are or are not transtheoretical words, especially by observing the implications that these differences between instrumentalist and eliminativist theories of emotion have on empirical design. Although both instrumentalists and eliminativists lack the theoretical resources to reject pretence emotions and likings as instances of emotion from the metaphysical perspective, instrumentalists can reject such phenomena as legitimate instances of emotion, given their optimism about ordinary language

[15] That there may be vagueness with objective kind categories may not be very problematic. Given the complexity of emotion, vagueness is to be expected, especially at this point in our scientific endeavours. Furthermore, it just might be an objective fact that nature has fuzzy joints. It may also not be very problematic that objective kinds are interest relative. What matters is not whether objective kinds are interests relative, but whether they need be.

[16] I take it that there is considerable vagueness in ordinary language as to whether likings are emotions.

emotion words (i.e., from a meta-semantic perspective).[17] This is especially significant because it allows us to further understand the significance of optimism about ordinary language emotion words for the science of emotion.

De Sousa (1987) associates what I refer to as realism with the *modern view*, which takes natural kind names (e.g., emotion) to be rigid designators. Saul Kripke ([1972] 1980) is the philosopher most often identified with the notion of rigid designation. For Kripke, proper names and natural kind terms, as eloquently recounted by de Sousa, get their reference by ostension (1987, 563). I argue here that one might also associate rigid designation with optimism about ordinary language emotion words. Accordingly, realists and instrumentalists might be said to share similar intuitions about which instances in the world are and are not emotions,[18] but they disagree on their metaphysical assumption about those instances (whether they are objective or subjective kinds).

For Kripke, the rigid designation of natural kind terms, such as *water*, *gold*, *heat*, *tiger*, and *cow*, were especially significant because this notion offered conditions for falsification. Rigid designation was a method for fixing the referent of a term, even when the essential characteristics of the referent were unknown at the time of its fixing, and identified only in virtue of a stipulative definition. Once a referent is fixed, identifying the nature of the referent is a matter of scientific discovery, which might result in the falsification of the stipulative definition. As Kripke explains with his example of *cat*: "Cats might turn out to be automata, or strange demons … planted by a magician. Suppose they turned out to be a species of demons. Then … the inclination is to say, not that they there turned out to be no cats, but that cats have turned out not to be animals as we originally supposed" ([1972] 1980, 122; see also Kripke 2011). Thus, if one associates optimism about ordinary language emotion words with Kripke's notion of rigid designation, one should notice that realists and instrumentalists have placed opposing metaphysical bets on future scientific discoveries about what exactly emotions are, and at least some arguments on one side can be taken as putatively falsifying the other. Yet, as contrary theories, these falsifying results are limited, since both kinds of theories can be proven false, while only one kind between the two can be proven true. And this is where contradictory theories make their contribution to the overall scientific enterprise and aim toward unification. Although at least seemingly talking past each other,[19] contradictory theories are theoretical alternatives that would be welcomed in light of the possibility that both realist and instrumentalist theories are proven false, while also serving as a means for falsifying realist or instrumentalist theories.

Such a conclusion runs counter to Mandler's conclusion about the productivity of ordinary language for the science of emotion: "To assert that such a truth is available by a proper examination of our phenomenal selves or by the proper analysis of language is, at least, a hinderance and, at worst, a wall that keeps us from playing that most productive game—science" ([1975] 1984, 7). And one might attempt to argue that this does not bode well for the pessimist about ordinary language. One might even go so far as to argue that the crisis in the science of emotion is a consequence of such untethering. Such a conclusion, however, would fail to see the forest for the trees. One might instead conclude that on the waters of the science of emotion, guided by their relevant lights (stars or constellations), realist and instrumentalist theories cast the narrowest methodological net with their object of study, and eliminative-realist and eliminative theories can cast wider nets. These similarities

[17] One might provide a similar analysis of eliminative-realist positions, but I do not do so here since the primary theories of interest for this paper are the psychological basic emotion theories, appraisal theories, social functional theories, social constructivist theories, and psychological constructionist theories.

[18] Although there might be some vagueness with respect to cases like likings.

[19] Contradictory theories do not necessarily disagree on which particular instances of an experience make up the category of study.

and differences, however, may be necessary for the productivity of the science of emotion. Once theories are categorised in accordance with their fundamental differences, one can more clearly understand how each kind of theory takes a logical position of falsification in relation to the other three kinds, so as to contribute to the joint, cooperative effort in understanding the complexities of emotion, and to ultimately arrive at a true, unified theory of emotion. Thus, each kind of theory makes a unique and necessary contribution to the science of emotion, and as such, although theory integration to the extent proposed by Scherer may be fruitful, moving beyond that extent, without fundamental shifts, may ultimately be impossible in the science of emotion.

References

Armon-Jones, C. 1986. “The Thesis of Social Constructionism.” In *The Social Construction of Emotions*, edited by R. Harré, 32–56. Blackwell.

Averill, J. R. 1980. “A Constructivist View of Emotion.” In *Emotion: Theory, Research and Experience, Vol. 1: Theories of Emotion*, edited by R. Plutchik and H. Kellerman, 305–339. Academic Press.

———. 1986. “The Acquisition of Emotions During Childhood.” In *The Social Construction of Emotions*, edited by Rom Harré, 98–118. Blackwell.

———. 1997. “The Emotions: An Integrative Approach.” In *Handbook of Personality Psychology*, edited by R. Hogan and J. A. Johnson, 513–41. Academic Press.

———. 2004. “Everyday Emotions: Let Me Count the Ways.” *Social Science Information* 43 (4): 571–80.

Barrett, L. F. 2017a. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

———. 2017b. “The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization.” *Social Cognitive and Affective Neuroscience* 12 (1): 1–23.

Bozinovski, S. 2018. “Cognition-Emotion Primacy Debate and Crossbar Adaptive Array in 1980-1982.” *Procedia Computer Science* 145: 105–111.

Brooks, J. A., J. Chikazoe, N. Sadato, and J. B. Freeman. 2019. “The Neural Representation of Facial-Emotion Categories Reflects Conceptual Structure.” *Proceedings of the National Academy of Sciences* 116 (32): 15861–70.

De Sousa, R. 1984. “The Natural Shiftiness of Natural Kinds.” *Canadian Journal of Philosophy* 14 (4): 561–80.

———. 1987. *The Rationality of Emotion*. MIT Press.

DiGirolamo, M. A. and J. A. Russell. 2017. “The Emotion Seen in a Face Can Be a Methodological Artifact: The Process of Elimination Hypothesis.” *Emotion* 17 (3): 538–46.

Ekman, P. 1999. “Basic Emotion.” In *Handbook of Cognition and Emotion*, edited by T. Dalgleish and M. Power, 45–60. John Wiley and Sons.

———. 2003. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times Books.

Fodor, J. A. 1994. *The Elm and the Expert: Mentalese and Its Semantics*. MIT Press.

Gigerenzer, G. 2017. “A Theory Integration Program.” *Decision* 4 (3): 133–45.

Griffiths, P. E. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press.

Hibberd, F. J. 2019. “What Is Scientific Definition?” *The Journal of Mind and Behavior* 40 (1): 29–52.

Jia, S., S. Wang, C. Hu, P. J. Webster, and X. Li. 2021. “Detection of Genuine and Posed Facial Expressions of Emotion: Databases and Methods.” *Frontiers in Psychology* 11: 580287.

Keinänen, M. 2015. “A Trope Nominalist Theory of Natural Kinds.” In *Nominalism About Properties: New Essays*, edited by G. Guigon and G. Rodriguez-Pereyra, 156–74.

Keltner, D. and J. Haidt. 1999. “Social Functions of Emotions at Four Levels of Analysis.” *Cognition and Emotion* 13 (5): 505–21.

Kollareth, D., Mariko K., and J. A. Russell. 2019. “Shame is a Folk Term Unsuitable as a Technical Term in Science.” In *Interdisciplinary Perspectives on Shame: Methods, Theories, Norms, Cultures, and Politics*, edited by C. Mun, 3–26. Lexington Books.

Kripke, S. A. (1972) 1980. *Naming and Necessity*. Harvard University Press.

———. 2011. “Identity and Necessity.” In *Philosophical Troubles: Collected Papers*, *Volume 1*, edited by S. A. Kripke, 1–26. Oxford Academic.

Krys, K., C.-M. Vauclair, C. A. Capaldi, V. M.-C. Lun, M. H. Bond, A. Domínguez-Espinosa, C. Torres, O. V. Lipp, L. S. S. Manickam, C. Xing, R. Antalíkova, V. Pavlopoulos, J. Teyssier, T. Hur, K. Hansen, P. Szarota, R. A. Ahmed, E. Burtceva, A. Chkhaidze, E. Cenko, P. Denoux, M. Fülöp, A. Hassan, D. O. Igbokwe, I. Işık, G. Javangwe, M. Malbran, F. Maricchilo, H.

Mikarsa, L. K. Miles, M. Nader, J. Park, M. Rizwan, R. Salem, B. Schwarz, I. Shah, C.-R. Sun, W. van Tilburg, W. Wagner, R. Wise, and A. A. Yu. 2016. "Be Careful Where You Smile: Culture Shapes Judgments of Intelligence and Honesty of Smiling Individuals." *Journal of Nonverbal Behavio*r 40 (2): 101–116.

Lazarus, R. S. 1990. "Constructs of the Mind in Adaptation." In *Psychological and Biological Approaches to Emotion*, edited by N. L. Stein, B. Levanthal, and T. Trabasso, 3–19. Lawrence Erlbaum Associates.

———. 1991. *Emotion and Adaptation*. Oxford University Press.

Lindquist, K. A., J. C. Jackson, J. Leshin, A. B. Satpute, and M. Gendron. 2022. "The Cultural Evolution of Emotion." *Nature Reviews: Psychology* 1: 669–81.

Mandler, G. (1975) 1984. *Mind and Body: Psychology of Emotion and Stress*. W.W. Norton and Company.

———. 1990. "A Constructivist Theory of Emotion." In *Psychological and Biological Approaches to Emotion*, edited by N. L. Stein, B. B. Leventhal, and T. Trabasso, 21–44. Erlbaum.

Moors, A. 2022. *Demystifying Emotions: A Typology of Theories in Psychology and Philosophy*. Cambridge University Press.

Mun, C. 2014. "A New Foundation for the Disciplines of Philosophy and Psychology: Unification without Consilience." PhD diss., Arizona State University.

———. 2016. "Natural Kinds, Social Constructions, and Ordinary Language: Clarifying the Crisis in the Science of Emotion." *Journal of Social Ontology* 2 (2): 247–69.

———. 2021. *Interdisciplinary Foundations for the Science of Emotion: Unification without Consilience*. Palgrave Macmillan.

———. 2022. "The Science of Emotion: Mind, Body, and Culture." *Philosophies* 7 (6): 144.

Nelson, C. A. and J. M. Leppänen. 2009. "Tuning the Developing Brain to Social Signals of Emotions." *Nature Reviews: Neuroscience* 10 (1): 37–47.

Nussbaum, M. C. 2001. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press.

———. 2016. *Anger and Forgiveness: Resentment, Generosity, and Justice*. Oxford University Press.

Panksepp, J. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.

———. 2008. "Carving 'Natural' Emotions: 'Kindly' from Bottom-Up but Not Top-Down." *Journal of Theoretical and Philosophical Psychology* 28 (2): 395–422.

Prinz, J. J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press.

Pugmire, D. 1994. "Real Emotion." *Philosophy and Phenomenological Research* 54 (1): 105–122.

Putnam, H. 1973. "Meaning and Reference." *The Journal of Philosophy* 70 (19): 699–711.

Radford, C. and M. Weston. 1975. "How Can We Be Moved by the Fate of Anna Karenina?" *Proceedings of the Aristotelian Society, Supplementary Volumes* 49: 67–93.

Reisenzein, R. 2022. "Tasks for a Theoretical Psychology of Emotion." *Cognition and Emotion* 36 (2), 171–87.

Ruphy, S. 2010. "Are Stellar Kinds Natural Kinds? A Challenging Newcomer in the Monism/Pluralism and Realism/Antirealism Debates." *Philosophy of Science* 77 (5): 1109–20.

Russell, J. A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110 (1): 145–72.

———. 2009. "Emotion, Core Affect, and Psychological Construction." *Cognition and Emotion* 23: 1259–83.

———. 2010. "Descriptive and Prescriptive Definitions of Emotion." *Emotion Review* 2: 377–78.

———. 2012. "Introduction to the Special Selection: On Defining Emotion." *Emotion Review* 4 (4): 337.

Satpute, A. B., E. C. Nook, S. Narayanan, J. Shu, J. Weber, and K. N. Ochsner. 2016. "Emotions in 'Black and White' or Shades of Gray? How We Think About Emotion Shapes Our Perception and Neural Representation of Emotion." *Psychological Science* 27 (11): 1428–42.

Saxen, F., P. Werner, and A. Al-Hamadi. 2017. "Real vs. Fake Emotion Challenge: Learning to Rank Authenticity from Facial Activity Descriptors." In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 3073–78. IEEE.

Scarantino, A. 2012. "How to Define Emotions Scientifically." *Emotion Review* 4 (4): 358–68.

Scarantino, A and P. Griffiths. 2011. "Don't Give Up on Basic Emotions." *Emotion Review* 3 (4): 444–54.

Scarantino, A. and R. de Sousa. 2018. "Emotion." In *The Stanford Encyclopedia of Philosophy*. Edited by E. N. Zalta. Winter 2018 edition. https://plato.stanford.edu/archives/win2018/entries/emotion.

Schachter, S. and J. E. Singer. 1962. “Cognitive, Social, and Physiological Determinants of Emotional State.” *Psychological Review* 69: 379–99.

Scherer, K. R. 1982. “Emotion as a Process: Function, Origin and Regulation.” *Social Science Information/Information sur les Sciences Sociales* 21: 555–70.

———. 1987. “Toward a Dynamic Theory of Emotion: The Component Process Model of Affective States.” *Geneva Studies in Emotion and Communication* 1 (1): 1–98.

———. 2000. “Psychological Models of Emotion.” In *The Neuropsychology of Emotion*, edited by J. Borod, 137–62. Oxford University Press.

———. 2001. “Appraisal Considered as a Process of Multilevel Sequential Checking.” In *Appraisal Processes in Emotion: Theory, Methods, Research*, edited by K. R. Scherer, A. Schor, and T. Johnstone, 92–120. Oxford University Press.

———. 2005. “What Are Emotions? And How Can They Be Measured?” *Social Science Information* 44: 695–729.

———. 2009. “Emotion as Emergent Processes: They Require a Dynamic Computational Architecture.” *Philosophical Transactions of the Royal Society B* 364 (1535): 3459–74.

———. 2012. “Neuroscience Findings Are Consistent with Appraisal Theories of Emotion; but Does the Brain ‘Respect’ Constructionism?” *Behavioral and Brain Sciences* 35: 163–64.

———. 2019. “Towards a Prediction and Data Driven Computational Process Model of Emotion.” *IEEE Transactions on Affective Computing* 12 (2): 279–92.

———. 2022. “Theory Convergence in Emotion Science is Timely and Realistic.” *Cognition and Emotion* 36 (2): 154–70.

Tahko, T. E. 2021. *Unity of Science*. Cambridge University Press.

van den Boomen, C., N. M. Munsters, and C. Kemner. 2019. “Emotion Processing in the Infant Brain: The Importance of Local Information.” *Neuropsychologia* 126: 62–68.

Wundt, W. 1987. *Outlines of Psychology*. Translated by C. H. Judd. Whilhelm Englemann.

Zachar, P. 2022. “The Psychological Construction of Emotion—A Non-Essentialist Philosophy of Science.” *Emotion Review* 14 (1): 3–14.

Received: 06 June 2024
Accepted: 20 June 2025
Published: 28 August 2025

Investigating the Fundamental Base of Emotion Science

Juan R. Loaiza – Universidad Alberto Hurtado, Chile, jloaiza@uahurtado.cl

Abstract

How should we investigate folk emotion concepts for the purposes of anchoring scientific emotion concepts? In this article, I expand on Mun's ideas on what she calls the fundamental base for interdisciplinary inquiry in the science of emotion. I argue that first-personal experiences should not be identified with the *explananda* of emotion science. This is because these experiences do not provide a theoretically neutral and intersubjectively accessible ground on which to anchor scientific concepts of emotions. Instead, I propose a pragmatic account of how to investigate the fundamental base, drawing on Haslanger's distinction between manifest and operative concepts, as well as some of her views on social constructionism and semantic externalism. This pragmatic account covers the relevant sources of evidence considered by Mun, but also calls for an investigation of cultural variations in emotion concepts, scripts, and norms. The upshot of the pragmatic account is a more expansive research programme that maintains the interdisciplinary spirit of emotion science.

Keywords: folk emotion concepts, first-person experience, externalism, social construction

What should emotion science explain? Emotions, of course. But how are emotions to be individuated and identified for the purposes of emotion science? Answering this question is not an easy task. On one view, emotions should be individuated and identified through ordinary intuitions. This is what Mun (2021) calls the "*fundamental base for interdisciplinary inquiry in the science of emotion*." But how should we identify those ordinary intuitions? In other words, how should we characterise that fundamental base for emotion science?

In this paper, I argue for an account of the fundamental base for emotion science, and I propose a project for the analysis of emotion concepts. With Mun, I accept that the fundamental base can be initially characterised through ordinary intuitions about emotions. However, in contrast with Mun, I claim that ordinary intuitions should not include first-person experiences. Instead, I propose a pragmatic account of the fundamental base that, following insights from semantic externalism, identifies ordinary intuitions with the public use of emotion concepts. The upshot of this account is an empirically tractable project in emotion research that is compatible with an interdisciplinary science, and that raises important epistemic and ethical questions towards constructing a scientific theory of emotion.

My argument will proceed as follows. First, I will rehearse some arguments in the literature in support of the idea that emotion science should be based on ordinary intuitions or some form of folk psychology. Second, I will focus on Mun's account of the fundamental base, and show the merits and limitations of characterising the fundamental base in terms of first-person intuitions. To achieve this, I present an argument based on considerations concerning theoretical and epistemological difficulties with relying on first-person experience for the purposes of identifying the fundamental base. Third, I present my pragmatic account of the fundamental base, drawing from Haslanger's (2005, [2000] 2012, [2006] 2012) ideas on semantic externalism and social construction. Fourth, I conclude by raising some epistemic and ethical questions for the project of investigating the fundamental base and formulating scientific concepts of emotions.

1. Emotion Science and Folk Intuitions

The question of how to define emotion concepts in science can be traced back to the question of whether emotions constitute natural kinds. The motivation behind framing the question in terms of natural kinds stems from the assumption that one of the aims of scientific inquiry is to discover how objects and phenomena in the world belong together and how they differ from others. Importantly, these distinctions are presumably independent of human interests, such that they exist independently of human practices and thus are one of the objects of discovery for science (Bird and Tobin 2022). Framed in this way, the problem is, first, to determine what explains whether objects in the world belong in a class or not—that is, to establish some criterion for natural kindhood—and second, to evaluate whether this criterion applies in the case of emotions.

There have been two important challenges to the project of identifying emotions with some natural kind (or set of natural kinds). First, there is Paul Griffiths' (1997) claim that the vernacular concept of "emotions" does not refer to any single natural kind because the phenomena captured by this concept belong to three different classes: basic emotions, arguably identified with affect programs; higher cognitive emotions, identified with evolved solutions to cooperation problems; and socially constructed emotions. Second, there is Lisa Barrett's (2006; 2017) claim that emotions do not have a neural or physiological *fingerprint*, a pattern to which each emotion category corresponds, and that they therefore cannot be said to form natural kinds or to support typological thinking (Barrett and Lida 2025). Other scientists, along with Barrett, have also published evidence that arguably casts doubt on the natural kind assumption (e.g. Lindquist et al. 2012; Touroutoglou et al. 2014).

Faced with these challenges, researchers have opted for two roads. One, defended by Griffiths, is eliminativism, i.e., the claim that there should be no "emotion science" proper, and that the vernacular concept of emotions plays no role in scientific inquiry. James Russell (2009) has defended a similar view, arguing that folk emotion concepts are not useful tools for scientific inquiry. The second kind of challenge is some form of revisionism, according to which there can be a science of emotions, but we need a separate conceptual framework to capture what it is that emotion science is studying. This new conceptual framework would presumably satisfy the demands of scientific inquiry, such as supporting explanations and predictions, and, crucially, should be allowed to differ from folk understandings of emotions. This is the idea behind Andrea Scarantino's (2012) distinction between the Folk Emotion Project and the Scientific Emotion Project.

Current emotion science leans towards this latter alternative, offering revised concepts of emotions that are presumably more scientifically fruitful than their vernacular counterparts. However, as with any analytic endeavour, any concept that is a candidate to be the *analysans* of a vernacular concept faces the risk of differing so much from it that it ends up being an entirely different concept altogether. One such objection was raised by P. F. Strawson (1963) against Rudolf Carnap's ([1950] 1963) idea that we could explicate folk concepts to construct more scientifically fruitful ones. This is also a common theme in discussions on conceptual engineering, a project that inherits from Carnap in calling for the construction of concepts that are more suitable for specific epistemic, ethical, or political projects than the currently operating (often folk) concepts (Díaz-León 2020; Thomasson 2020).

The solution to this problem is to somehow *anchor* scientific concepts in folk concepts, such that we can reasonably claim both that the new scientific concepts are better suited for empirical research and that they still refer to some extent to the phenomena we refer to when we use folk concepts. This is important to guaranteeing that emotion science remains a science of *emotion*, which is what the theory was supposed to explain in the first place. How this anchoring relation works remains to be worked out.

The problem of how to explicate the anchoring relation can be divided into two parts. First is the question of what the properties of this relation are, and which formal and epistemic criteria this relation should respect. The second is which domain such a relation is to be defined over, that is, how to characterise the domain of folk emotion concepts that will constitute the basis for scientific theory construction. I will discuss the second question throughout the rest of this article. Yet, something should also be said about the first.

In my view, the anchoring relation should first be such that it preserves important elements of the extension and the anti-extension of the original concept. In other words, scientific concepts should be similar enough in extension to folk concepts. I shall call this the *similarity criterion*, honouring Carnap's similarity criterion of explication. It is important to clarify, though, that the conditions under which an element may be included or excluded from the extension of a given category should be determined in actual scientific practice, based on theoretical values such as coherence, explanatory power, or other sets of epistemic (and even non-epistemic) criteria, rather than being decided *a priori*. This entails that deciding on limit cases and constructing scientific concepts anchored in folk concepts is an ongoing endeavour, rather than a one-off conceptual analysis project.

Second, the domain of the anchoring relation, i.e., folk concepts, must provide the grounds on which evidence will be obtained to evaluate different scientific theories of emotion. Otherwise, we would be identifying the *explananda* by presupposing theoretical constructs that already fit the explanation we want to obtain. Therefore, folk categories must be identified in ways that are as theoretically neutral as possible, at least regarding scientific theories of emotion, which will be evaluated based on evidence obtained by using these categories. I shall call this the *neutrality criterion*.

Lastly, researchers must be able to eventually agree to some extent on the reference of the folk concepts in question. This entails that it must be possible to settle disagreements that may emerge when identifying what it is that we are talking about when using folk concepts. To be able to settle such disagreements, we must have some way to ostensively pick out the phenomenon to which folk concepts refer, even if we temporarily disagree on how to describe them. This would enable us to compare and contrast our conceptions while maintaining a fixed reference, enabling us to eventually construct a unified framework in which to construct scientific concepts. I shall call this the *intersubjectivity criterion*.[1]

In sum, to anchor scientific concepts of emotion in folk concepts means to construct theoretical concepts that are similar enough in extension to folk concepts, which requires a method of identifying the latter. To identify folk concepts, we need some framework that is theoretically neutral (relative to the theories of emotion that we will evaluate based on how they respect the extensions of folk concepts) and that allows us to ostensively pick out the phenomena that emotion science is supposed to explain.

[1] This criterion can also be called an *objectivity criterion* if we understand objectivity as a form of intersubjective agreement. In this sense, the intersubjectivity criterion would warrant what Douglas (2004) calls *concordant objectivity*.

With these criteria in mind, how should we identify folk concepts to which emotion science should be anchored? In the next section, I will discuss Mun's account of what she calls the *fundamental base* of emotion science. I will argue that, by relying on first-person experiences, Mun's account does not satisfy the neutrality or intersubjectivity criteria. If this is correct, a new account of the fundamental base of emotion science is required.

2. The Problem of the Fundamental Base

2.1 Mun's Account of the Fundamental Base

The claim that emotion science should be based on ordinary intuitions is what Mun (2021) calls the *fundamental base* of emotion science. In her words, it is the claim that "an adequate theory of emotion must recognise that *ordinary folk intuitions* serve the fundamental purpose of identifying the explananda for the science of emotion" (93; my emphasis).

Although Mun offers some clues as to what she means by ordinary folk intuitions, they are still difficult to define precisely. First, Mun clarifies that when she speaks of ordinary folk intuitions, she means "the *ordinary experiences* that many people have which allow them to determine not only whether or not someone is experiencing an emotion, but also (for many) which allow them to determine what kind of emotion they are experiencing to a certain extent" (94; my emphasis).

This suggests that ordinary folk intuitions are the *experiences* which allow people to identify emotional experiences in others as well as themselves. Importantly, this use of the term "intuition" differs from other standard uses in the literature which understand intuitions as beliefs, dispositions, and seemings, among others (Pust 2024).

Immediately afterwards, Mun writes: "These ordinary intuitions are an object of study for the science of emotion, and such research is typically carried out as studies on emotion recognition and attribution" (94). By emotion recognition and attribution, Mun means recognising an emotion in another and identifying an emotion type, respectively, based on facial expressions, vocal cues, and bodily gestures (I shall refer to the collection of these three cues as the "emotional expression"). To illustrate this research, Mun reconstructs the findings from DiGirolamo and Russell (2017) that show (in her interpretation) that, although forced choice paradigms in emotion recognition and attribution studies artificially inflate agreement between subjects when they are asked to categorise an emotional expression, there is still some degree of uniformity in how subjects recognise and categorise emotions in others when using open choice paradigms. For Mun, this suggests that there is some uniformity to the "intuitions" that subjects use to categorise emotional expressions, which supports the idea that emotion science should begin from such intuitions.

Finally, Mun discusses whether emotion recognition and attribution require knowing the word associated with a given emotional expression, which she denies. In her view, emotion recognition and attribution are processes that need not involve labelling an expression, but only require reactions to others' expressions and behaviour. These processes, when they occur without labelling, exemplify, in Mun's words, "[our] recognitions of our ordinary intuitions about emotions, and it is with these intuitions, along with our first-person emotional experiences, that the science of emotion must begin" (2021, 98).

At this point, we can identify some of the problems with Mun's presentation of ordinary intuitions as the basis of emotion science. As I understand Mun's view, her claim is that emotion science should begin by investigating how people recognise that others are experiencing an emotion and how they identify which emotion they are experiencing. Yet, it is unclear which role experiences—especially *first-person emotional experiences*, which she emphasises—plays in identifying the *explananda* of emotion science. The studies on emotion recognition and attribution which Mun considers, although interesting in their own right, do not offer much information about the first-person experiences which enable subjects to attribute and recognise emotions. On the face of it, these studies only offer information about the general pattern of behaviour that can be observed by a third-person observation method across many subjects, not about what their experience is or how these subjects are using their experiences to identify emotions in others. Furthermore, while ordinary intuitions were initially identified with "the ordinary experiences [which allow emotion recognition and attribution]," in the latter formulation, "ordinary intuitions" and "first-person emotional experiences" appear as separate (and the invitation is to study one "along with" the other). Hence, it is unclear whether ordinary intuitions are the same as first-personal experiences, what the role of emotion recognition and attribution studies ought to be, or how ordinary intuitions and first-personal experiences should be understood in general.

In my view, there are at least two interpretations of Mun's idea. If we take the first formulation at face value, ordinary intuitions are first-person experiences which people use to recognise and attribute emotions in others. If we take the latter presentation more seriously, ordinary intuitions are separate from first-person experiences, raising the question of what these ordinary intuitions are and how they differ from first-person experiences. Put succinctly, either ordinary folk intuitions are identified with first-person experiences or they are separate and left undefined.

Despite these obscurities, there are some ideas that can be extracted more clearly from Mun's presentation. Under either interpretation of what ordinary intuitions are, first-person experiences are a necessary part of the fundamental base of emotion science, either because they are identical to ordinary intuitions, or because they are part of the fundamental base of emotion science, "along with" ordinary intuitions (whatever they may be). Additionally, for Mun, ordinary intuitions and first-person experiences (understood either as identical or as separate but equally important) can be studied by examining emotion attribution and recognition. This implies that emotion attribution and recognition studies offer information about ordinary intuition and first-person experiences as they influence attribution and recognition.

In what follows, I will argue that the first interpretation leads to an untenable account of ordinary intuitions that does not help us to clarify the fundamental base of emotion research. This would leave us with the second interpretation, that "ordinary intuitions" are left unanalysed. If we accept this second interpretation, then the remaining task is to offer an analysis of ordinary intuitions that makes them scientifically tractable and suitable for investigation, from which to begin theorising about emotions, a task that I undertake later.[2]

2.2 Against First-Person Approaches to the Fundamental Base

Let us assume that Mun does rely on first-person experiences as necessary elements of ordinary intuitions. Does this account of the fundamental base offer an account of folk categories that satisfies the similarity, neutrality, and intersubjectivity criteria mentioned above? In what follows, I will argue that even if this account satisfies the similarity and neutrality criteria, it does not satisfy the intersubjectivity criterion. This is because what

[2] This means that if my interpretation of Mun is incorrect, and what she has in mind is not that ordinary intuitions involve first-person experiences, the remainder of the argument can still be understood as an expansion on her overall project.

Mun identifies with the fundamental base, i.e., "ordinary intuitions and first-person experiences," does not provide grounds for picking out *explananda* intersubjectively. Consequently, we need some other account of the fundamental base that can satisfy all of the criteria and that offers a scientifically tractable base on which to anchor emotion science.

Let us consider how we can identify the phenomena that lie at the fundamental base according to Mun's account. Recall that Mun's fundamental base consists of the ordinary intuitions and first-person experiences that enable emotion recognition and attribution. To identify them, according to Mun's story, we can rely on the methods used in emotion recognition and attribution studies, such as forced or open choice paradigms and the like. Yet, as I argued in the previous section, what are important to (or at least a necessary part of) the fundamental base are the first-person experiences which subjects use to recognise and attribute emotions in others. The methods used in emotion recognition and attribution studies, which lead to agreement scores rather than self-reports, must be then interpreted as offering indirect information on first-person experience and how it is (consciously or unconsciously) deployed in emotion recognition and attribution.

Does first-person experience help to identify the *explananda* of emotion science? In my view, it is not possible to satisfy both the neutrality criterion and the intersubjectivity criterion mentioned before if we identify the *explananda* of emotion science with, or by means of, first-person experience. Either we need a purely phenomenal description of what occurs in emotion recognition and attribution, and first-person experience (which cannot be accessed by others, violating the intersubjectivity criterion), or we need a theory of emotion that describes our phenomenology in some way as to explain how first-person experience plays a role in emotion recognition and attribution (which violates the neutrality criterion). Hence, the appeal to first-person experience is a non-starter when it comes to identifying the fundamental base for emotion science. Let us unpack this claim.

First, let us assume that what help to identify the *explananda* of emotion science are the purely phenomenal aspects of first-person experience. To get at them, we must obtain a description of this phenomenology such that it is clear what emotion science is purported to explain. However, how can such descriptions be intersubjectively compared and contrasted so as to identify the *explananda* of a scientific research programme? Given that we do not directly share epistemic access to the first-person experience of others, this means that we would not have epistemic access to what emotion science is supposed to explain. This would violate the intersubjectivity criterion, since it would make public identification of the *explananda* of emotion science impossible (insofar as science involves not only first-person descriptions of mental states but also communities correcting and jointly investigating a common phenomenon of interest).

There seems, however, to be a way out. Even if we do not have first-person epistemic access to others' experiences, it might not be the case that first-person experiences are completely epistemically inaccessible. After all, we have third-person methods of observation that may inform us about others' experiences. Examples of these methods may include precisely the methods Mun has in mind, that is, choice methods involved in emotion recognition and attribution, among others. Interpreted as offering indirect information on first-person experiences, these methods would provide evidence of the first-person experiences of subjects, even beyond the information that can be obtained by self-report.

Yet, to make sense of how these methods give us information about first-person experiences, we already need to presuppose a theoretical background that makes explanations bridging third-person observation with first-person experience intelligible. Not only would this require, for instance, a theory of consciousness (which

Mun discusses), but also a theory of emotion that already explains how our methods of observation map onto first-person experience, and how first-person experience affects the way in which we recognise and categorise emotions in others. Hence, even if we had a workable theory of first-person experience in a form that allowed intersubjective validation (that is, which connected third-person methods of observation with first-person emotional experiences), we would be violating the neutrality criterion, i.e., we would presuppose a theoretical framework, which is what we are supposed to evaluate after (and not before) identifying the fundamental base.[3]

In sum, identifying the fundamental base of emotion research, what identifies the *explananda* of emotion science, by appealing to first-person experience, is untenable. At best, it introduces important theoretical problems, and at worst it is epistemically unsound, given the risk of circularity (i.e., of violating the neutrality criterion) or lack of intersubjective access to the phenomena under investigation (i.e., violating the intersubjectivity criterion). We need, therefore, another way to fix what would constitute the fundamental base of emotion science. In other words, we need a different way to analyse the so-called "ordinary intuitions" involved in starting, and more importantly, evaluating a scientifically tractable theory of emotion.

In what follows, I will propose one such analysis of "ordinary intuitions" that respects the neutrality and intersubjectivity criteria. What we need, I shall argue, is an ostensive device that can help make reference to the *explananda* of emotion science without presupposing a specific theory of emotions (and hence no specific explanation of the role of first-person experience in emotion recognition and attribution), while allowing for intersubjective access to the phenomenon of interest, and that can enable means of comparison that are key to any scientific enterprise. These devices, I argue, are the concepts that operate in sociolinguistic exchanges about emotions.

3. A Pragmatic Account of the Fundamental Base

In this section, I will propose an account of "ordinary intuitions" in terms of sociolinguistic practices. My claim is that so-called "ordinary intuitions" about emotions should be identified through the referents of emotion vocabularies (folk emotion concepts, descriptions, etc.) and how they operate in social practices. To unpack this claim, I will follow Sally Haslanger's (2005, [2000] 2012, [2006] 2012) semantic externalism and her analyses of folk concepts of gender and race. While I do not want to commit to the idea that emotions are social kinds,[4] I believe that Haslanger's analyses of these concepts offers a framework that works well for other types of folk concepts. This helps maintain the similarity criterion, along with the neutrality and intersubjectivity criterion, which would lead to a better account of the so-called "fundamental base" of emotion science.

[3] Later, I will argue for an approach based on semantic externalism. While this may also be a theoretical framework, it does not constitute a theory of emotion, and it is plausibly less committal than a full-blown theory of phenomenal consciousness.

[4] There is a fundamental difference between referring to a phenomenon in the context of social interaction and holding a phenomenon to be a social kind. Kindhood is a relation concerning what warrants similarity judgments between a group of objects, often identified with essences (Kripke 1980) or mechanisms underlying the properties these objects exhibit (Boyd 1991). Claiming that some phenomena (such as emotions) form a social kind implies the claim that these phenomena are united (epistemically or metaphysically) by social properties. Yet, referring to a phenomenon in the context of social interaction means identifying devices that refer to some objects under some description in social interaction, without specifying which properties warrant similarity between the objects. For instance, in social interactions we can refer to eyes and describe them in a myriad of ways (beautiful, aggressive, passionate, etc.), but this does not imply that they form social kinds. My claim that we can attend to how we refer to emotions in social interactions, without committing to the view that emotions form social kinds, is analogous to claims about eyes in this regard.

3.1 Haslanger on Folk Concepts

Haslanger is interested in the question of how folk concepts of social kinds relate to more technical or theoretical concepts that can be used in the social sciences and in political discourse. She studies cases such as the concept of "woman," "man," "Black," "White," and other social kind concepts where everyday use may not map onto theoretical uses of these terms (similar to the case of folk and scientific emotion concepts). For instance, when people use concepts of race in everyday discourse, they might assume that race refers to some biological grouping of human beings that, in reality, does not exist. While people who use race terms in this way might fail to refer to what they think they are referring to, it is still problematic to claim that race does not exist at all. After all, people are affected by racial categories, and therefore there appears to be a social practice that is configured around these concepts. To resist such practices, Haslanger claims that we should be sensitive to differences between what people have in their minds when they use race and other social kind concepts, and the social practices to which these concepts actually refer.

One way to understand the differences between what people think they mean and what they actually mean is by invoking a distinction between *manifest* and *operative* concepts. According to Haslanger, manifest concepts are those that are accessible to people through introspection, that is, they are what people have in their minds and can offer as a definition when asked about what they mean. Operative concepts, on the other hand, are concepts that describe actual social meanings around which people coordinate their behaviour. Importantly, manifest and operative concepts often do not coincide. This is because individuals may not be aware that what they think by the use of a given concept does not adequately capture how they are actually behaving.

Haslanger uses the illustrative example of the concept "parent." When used in the context of her daughter's school, many people have in mind a concept of "parent" as that of "progenitor." Taken at face value, this use of the concept excludes parents who did not give birth to their children, such as adoptive parents or other caregivers. However, Haslanger notes that the concept "parent," in actual social practice, does include these other forms of parenthood, since it is expected that all caregivers, regardless of their biological relation to their children, respond to calls addressed to "parents" (e.g., when invited to "Parents' Night" at school). Hence, we can distinguish between what people might have in mind when offering a definition of "parent" (i.e., the often biology-laden concept), and the concept that adequately captures the actual social practice around which the community is organising their behaviour. The first is what Haslanger calls the *manifest* concept; the latter is the *operative* concept.

How can we cash out the difference between *manifest* and *operative* concepts? Haslanger explains this difference in terms of the distinction between semantic internalism and externalism. According to semantic internalism, the meaning of a term is bound to what speakers have in mind when they use the term and what is accessible to them when explaining their meant use of the term. Hence, to study the meaning of a term, according to the internalist, one should examine through introspection what one means (e.g., speaker intentions or conceptions) when one uses a term. Semantic externalism, in contrast, is the idea that the meaning of a term is bound to external facts about the use of the term that may not be known by speakers. As Wikforss (2008) explains it, externalists claim that it is not psychological states that determine meaning, but external (physical or social) facts about the environment which may or may not be known to speakers.[5] In view of the distinction between semantic internalism and externalism, the difference between manifest and operative concepts can

[5] To be clear, this is what Wikforss (2008) calls *foundational externalism*. She distinguishes two other forms of externalism, namely, *externalist semantics* and *psychological externalism*. Here I restrict externalism to *foundational externalism*, a claim about meaning determination, rather than a claim about the semantic value of terms (*externalist semantics*) or the content of mental states (*psychological externalism*).

be cashed out in terms of what determines meaning: while the meanings of manifest concepts are determined internally (by the psychological states of speakers when using a term), the meanings of operative concepts are determined externally (by physical or social facts about their environments and the linguistic communities who use them).

What kinds of concepts are amenable to analysis under the distinction between manifest and operative concepts? Haslanger thinks that to analyse social kinds, we must adopt an externalist view of the semantics of social kind terms. This is because, as illustrated by the previous example, people might not be sensitive to how a concept is actually operating in practice, and they might err about the role a concept plays when deployed within a community. In other words, their psychological states might not reflect the actual ways in which they are organising their behaviour (e.g., what they actually expect from others) when using some term, such that they might think they mean something when they expect different reactions from others (e.g. a school principal who is still including non-progenitor caregivers under the extension of "parent" and expecting certain forms of behaviour despite their explicit beliefs about the meaning of the term).

The distinction between manifest and operative concepts need not be limited to terms referring to social kinds such as "parent," even though those are Haslanger's focus. Any time our mental states about referents (i.e., intensions) may differ from what these referents (i.e., extensions) are, a distinction can be drawn between what we believe, our intentions, and what we are talking about. These are the traditional cases for semantic externalism about natural kinds such as water or arthritis (Putnam 1975; Burge 1979). Consequently, put in terms of folk theories and concepts, whenever we have a folk theory which involves some form of reference to a phenomenon (e.g., some vocabulary), we may analyse it in terms of what users of such a theory think and how such a theory operates, and we must be open to the possibility that these two levels of analysis do not match.

With Haslanger's general framework for analysing social kind concepts in mind, let us now apply it to the case of the fundamental base for the science of emotion. First, I will argue that the ordinary intuitions and folk emotion concepts relevant to approaching the fundamental base must be coined in terms of operative rather than manifest concepts. Here I will mostly follow Haslanger's strategy concerning social kinds, adding some nuances more specific to the case of emotion concepts. Second, I will present some interesting sources of evidence to which researchers can turn to examine the fundamental base. These include studies in emotion attribution and recognition, following Mun, but also other sources such as cross-cultural research on variations in emotion terms, and research on emotion scripts.

3.2 Back to the Fundamental Base

Let us take stock of the argument so far. I have argued that scientific emotion concepts must be somehow anchored to folk emotion concepts, and that this anchoring relation must satisfy three criteria: scientific emotion concepts must be similar enough to their folk counterparts (*similarity*); folk emotion concepts must be understood in theory-neutral terms relative to a theory of emotion (*neutrality*); and some form of intersubjective contrast must be possible, to settle disagreements about how we understand the fundamental base (*intersubjectivity*). I have also argued that Mun's appeal to first-person experience does not satisfy the neutrality and intersubjectivity criteria, leaving the question of how to identify the *explananda* of emotion science (i.e., the fundamental base) undetermined. How then should we identify the *explananda* of emotion science?

Let us begin with what, I believe, is a reasonable assumption, and one in the spirit of Mun's idea of the fundamental base: emotion science should begin by identifying the phenomena we usually make reference to by using emotion vocabulary (the term "emotion," emotion categories, descriptions, etc.). One traditional approach to such questions was to investigate these terms and their respective concepts *a priori*, but this would, at best, reveal beliefs about emotions rather than clarify what we are talking about (i.e., some subset of manifest concepts of emotion). Following the insights that follow from semantic externalism, and Haslanger's ideas concerning manifest and operative concepts, I propose we identify the *explananda* of emotion science with the referents of *operative concepts* of emotion.[6] In other words, investigating the fundamental base of emotion science is to investigate the phenomena to which communities refer with their use of vocabulary such as "angry," "sad," "happy," and so on, which involves understanding how such vocabulary operates in practice (rather than what we believe about these terms).[7]

Investigating operative concepts, as with other projects of discovering what we are talking about, is an empirical matter. We must investigate when these concepts are applied, consciously or not, and to which sorts of phenomena they make reference. Crucially, rather than taking them at face value, these concepts must work as *ostensive devices*, as means by which we signal what we mean. To clarify this, consider the following analogy: if we want to investigate water, we should attend to the substance referred to by the term "water," using the term as an ostensive device to make reference to (rather than describe) the phenomenon that we are interested in. We would not take beliefs about water at face value, but as ways of making the reference explicit so that we can be sure what it is that we are supposed to investigate.

Appealing to semantic externalism in this way allows the investigation of the fundamental base to satisfy the criteria laid down above. First, we are guaranteed to maintain some similarity between folk and scientific concepts, insofar as reference must be maintained in both domains. In other words, by maintaining referents fixed for the clearer cases under the extension of folk concepts, scientific concepts will keep sufficient similarity with their folk counterparts. Second, by taking folk concepts as ostensive devices, and not as informing us about the nature of emotions themselves, we are respecting the neutrality criterion, since we are not presupposing a theoretical framework to understand emotions, nor one that explains the role some other phenomenon (such as first-person experience) plays in emotional processes. Lastly, given that we are fixing reference in terms of the public application of emotional vocabulary, the account I'm proposing satisfies the intersubjectivity criterion by providing means of reference that are publicly accessible and that allow disagreement and consensus on what is it that emotion science should explain.

To make this proposal clearer, I will dedicate the remainder of this article to illustrating how research on the fundamental base can advance by considering different aspects of how emotion concepts, vocabularies, terms, descriptions, and the like, operate. I will focus on three sources of information about how these vocabularies operate in practice, namely: emotion recognition and attribution studies (reconsidered), cross-cultural variations of emotion vocabularies, and research on emotion norms and scripts. This is by no means an exhaustive list of potential sources of evidence, but an illustration of how to expand the examination of the fundamental base to other areas of research and disciplines.

[6] Another alternative is to assume that we can identify reference by identifying natural kinds of emotions (see e.g. Scarantino 2012). Yet, this would already assume that emotions form natural kinds, an assumption that is contentious in the literature (Barrett 2006). While I believe some version of natural kind externalism about emotions is correct, I will not expand on it here, and I will assume a more modest approach to what unifies emotion categories in science.

[7] This vocabulary need not only involve names for emotion categories, but may also include descriptions and other linguistic forms of reference.

3.3 Emotion Attribution and Recognition

Let us begin by discussing the two sources already mentioned by Mun, namely, studies on emotion attribution and emotion recognition. I have argued that, on Mun's account, emotion attribution and recognition studies must be interpreted as providing information about first-person experience. Given the externalist framework I am proposing, which rejects first-person experience as fundamental to identifying the *explananda* of emotion science, should emotion recognition and attribution studies still be considered part of the investigation of the fundamental base?

One way of using emotion attribution and recognition studies is to interpret them, not in terms of how they inform us about first-person experience, but in terms of what they inform us concerning operative concepts of emotion. Following DiGirolamo and Russell (2017), for instance, we can claim that folk emotion categories are relatively stable despite subjects choosing among open-ended options. This does not imply that emotions are, by nature, stable kinds, but rather that the *explananda* of emotion science involve some degree of stability. This approach mirrors what Scarantino (2012) suggests as the Folk Emotion Project, which interprets findings in experimental psychology as findings concerning folk concepts of emotion. Making such interpretations clear is important, since appealing to emotion attribution and recognition studies would otherwise presuppose a set of theoretical concepts which would violate the neutrality criterion.

Other findings in emotion attribution and recognition studies may be interpreted along similar lines. For example, empirical evidence suggests that both emotion attribution and recognition are highly context-sensitive. Sabini and Silver (2005) show that distinctions between shame and embarrassment, as well as between envy and anger, are sensitive to contextual factors. According to the researchers, subjects' judgements of shame and embarrassment are differentiated depending on whether they perceive that an event has revealed a real flaw in themselves (in the case of shame) or has only appeared to reveal a flaw, but the flaw is not considered real (in the case of embarrassment). In the cases of envy and anger, envy is attributed in cases where a person unwarrantedly accuses another person's accomplishment, while anger is attributed in cases where such an accusation is taken as warranted. In both cases, there are social norms that dictate whether a flaw is considered real or not, or whether an accusation is considered warranted or not, which in turn determine the emotion to be attributed. Similar studies include findings that show that subjects use contextual information even under the instruction to disregard context (Ngo and Isaacowitz 2015), that subjects suffering from depression are much less sensitive to contextual cues, which affects their accuracy in emotion recognition (Rottenberg et al. 2005), and that subjects are prone to using contextual information to disambiguate between possible emotions to attribute in a given situation (Hareli et al. 2018).

Overall, the case for context-sensitivity effects seems well-supported by empirical evidence in experimental psychology. Interpreted as offering information about the operation of folk concepts, these findings shed light on how emotion concepts operate in actual practices and help us to identify the phenomena that emotion science is supposed to explain and theorise about. Yet, these are not the only studies that help us understand the operation of folk emotion concepts or that influence our understanding of the fundamental base of emotion science.

3.4 Emotions and Cross-Cultural Linguistics

Another important source of empirical evidence relevant to investigating the fundamental base are studies on how different emotion terms operate across cultures. If part of the fundamental base concerns the tools with which people describe their emotions, investigating the use of emotion terms seems quite straightforward.

Yet, one important aspect to emphasise is that such an investigation must be carried out relative to specific languages. This is because there are considerable difficulties in translating emotion terms across languages, and while these difficulties should be addressed, differences between linguistic contexts must be acknowledged. For example, a considerable amount of emotion research has presupposed that folk concepts coming from English are the basic starting point for a science of emotion. This yields a limited research programme that does not properly account for differences across languages, which at the very least leads to dire epistemological issues for theorising about emotions. This line of criticism has been pressed with great emphasis by Anna Wierzbicka (1999; 2009; 2014; see also Levisen 2019).

Wierzbicka has shown several cases where translation of emotion terms across cultures fails. For example, the English term "sadness" translates into two terms in Russian, namely, "grust" and "pečal." Further research on translation has shown similar results. Barger et al. (2010) show that translating "disgust" into Chinese yields four different terms: "taoyan," "yanwu," "exin," and "otu." However, when back-translating these terms into English, the researchers note that only "exin" translates to "disgust," although it contains themes related to "anger" as well, and "yanwu" and "taoyan" seem to be inadequate translations, missing some core themes of the English term "disgust." This suggests that English terms do not map neatly onto terms in other languages.

Research on the fundamental base of emotion science should take such translation problems seriously. This is because, if we prioritise English vocabularies, "sadness" is bound to be considered a single discrete category; but if we prioritise Russian vocabularies, we would be motivated to split this category into two types of emotion. Such decisions inform further taxonomical and empirical questions and are therefore a vital part of the investigation of the fundamental base.

Wierzbicka's solution is to construct a minimalistic metalanguage that enables translation using universal semantic primes, a metalanguage that she calls the *Natural Semantic Metalanguage* (NSM). As Goddard describes it: "Using the NSM metalanguage allows us to decompose complex language-specific concepts into configurations of simple concepts that are shared across languages. This allows a very high degree of semantic resolution and enables us to access language-specific meanings using rigorous, evidence-based procedures of semantic analysis" (2015, 294).

By investigating how emotion concepts vary across cultures, we can understand how wide the fundamental base of emotion science should be. Given Wierzbicka's findings, we should work towards a science of emotion that is not limited to how English emotion concepts operate. This exemplifies how research in cross-cultural linguistics can inform what we consider to be at the base of the science of emotion.

3.5 Emotion Scripts and Norms

Cross-cultural variations in emotions are not limited to variations in emotion terms. Emotional behaviours also vary across cultures. These include variations in expression, action tendencies, and social norms attached to emotional responses. These variations can be summed up under a common concept of variations in emotional *scripts* (Wierzbicka 1994; Eickers and Prinz 2021; Eickers 2024).

Eickers defines scripts as follows: "Scripts are normative, context-sensitive, nested knowledge structures that describe behavior in terms of corresponding events, situations, social roles, individuals, or mental state types in a way that guides action" (2024, 7–8). These structures are paramount to social cognition, as they are determinants of how people behave regarding social norms. In the case of emotions, scripts involve our knowledge of how emotional expressions, action tendencies, and other aspects of emotional behaviour, occur in concrete social settings. Hence, investigating emotional scripts can offer important information on operative emotion concepts that constitute the fundamental base, since they are part of the practices to which such concepts refer.

There are two aspects of scripts that are worth mentioning for the present argument. First, scripts do not necessarily refer to internal representational states or beliefs social agents have about the social world. This is important because it marks the difference between using them to investigate operative concepts of emotion and using them to investigate manifest concepts. While scripts can be made explicit and used consciously by individuals (e.g. when I learn how to behave in a specific context by learning the appropriate sequence of acceptable behaviours), they need not be used consciously. Most of our social lives actually occur without us reflecting on our behaviour, and emotional behaviours are no exception. This supports the externalist view I am arguing for, since, by investigating scripts, we can make sense of factors pertaining to emotional behaviour beyond beliefs and other internal states of individual minds.

Second, emotion scripts support and expand on both of the aforementioned areas of research, namely, investigating forms of emotion attribution and recognition from a cross-cultural perspective. Scripts are highly context-sensitive and culturally varied, and they help explain variations in the context of emotions. As Eickers and Prinz put it:

> Evidence for learning of scripts can be found in the aforementioned fact that emotions differ cross-culturally. For example, there are culturally specific behaviors such as bowing, flicking-off, and clapping. In addition, cultures differ in whether they encourage people to act aggressively when angry or to quietly sulk or brood (see Goddard (1996) on Malaysia), while other cultures discourage brooding (see Briggs (1970) on the Ifaluk). This suggests that anger is not an automatic response with a fixed action tendency, but rather a role that we act out in culturally prescribed ways. (2021, 356)

Besides the evidence Eickers and Prinz mention, there are other studies that exemplify how emotion scripts can serve to examine folk emotion concepts. For instance, Uchida and Kitayama (2009) show that US American and Japanese descriptions of happiness differ in how positive these emotions are and how much they are related to social relations. For US American populations, happiness is much more related to personal achievements and positive hedonic experiences than for Japanese populations, who judge happiness to involve ambivalent experiences that can even be socially disruptive.

Not only do cultural variations in emotion scripts help to account for variations in emotion terms, but they are also an important piece of emotion recognition and attribution. As I explained above, emotion recognition and attribution are context-sensitive. One way to understand such effects is by appealing to emotional scripts. If emotional scripts are part of the knowledge structure deployed to recognise and attribute emotions to others, and such structures involve the use of contextual information to specify and disambiguate information in the behaviour of others, it is natural to conclude that context-sensitivity in emotional attribution and expression is due to the way these scripts operate when people are engaging in these processes. In Eickers and Prinz's terms, emotion recognition is, after all, a social skill.

In sum, I take emotion recognition and attribution, emotion terms, and emotion scripts to be constitutive of the fundamental base of emotion research, the so-called "folk intuitions" at play that underlie emotion science and that are intended to anchor theoretical concepts of emotions. This invites an investigation of the fundamental base that is interdisciplinary, follows Mun's account of the relation between folk emotion concepts and scientific theories, and also expands to add other areas of research that can contribute to our understanding of emotions.

4. Conclusion

In this paper, I have argued for an account of the nature of the fundamental base of emotion science and a way to examine folk intuitions about emotions. While Mun (2021) advocates for grounding emotion science in ordinary intuitions, there are important philosophical challenges in determining the nature of these intuitions. Specifically, there are reasons to exclude first-person experiences from the scope of the fundamental base of emotion science. Instead, I posit an alternative approach that expands on Mun's, arguing that the fundamental base for emotion science should be characterised by the public use of emotion concepts. This pragmatic stance aligns with the goal of establishing an empirically tractable framework in which to theorise about emotions.

I started by first presenting existing viewpoints supporting the reliance on ordinary intuitions or folk psychology in emotion science. Subsequently, I examined Mun's account of the fundamental base, highlighting both its merits and limitations, particularly concerning first-person experiences. I focused on concerns related to how relying on first-person experience to identify the fundamental base does not satisfy the neutrality and intersubjectivity criteria that an account of the *explananda* of a scientific research programme should satisfy. These arguments led to a pragmatic account informed by Haslanger's ideas on semantic externalism and social construction, and which, in my view, offers a promising alternative that circumvents the challenges to the use of first-person experiences in identifying the *explananda* of emotion research.

On this pragmatic account, the explananda of emotion science should be identified by investigating operative concepts of emotion—that is, how folk emotion concepts operate in actual sociolinguistic practices. This approach shifts the focus from individual subjective experiences to shared, public understandings of emotions, which can provide a more stable foundation for emotion research. By grounding emotion science in these operative concepts, we can avoid the pitfalls of relying on first-person experiences, which are inherently subjective and difficult to validate across different individuals.

Furthermore, this pragmatic stance not only addresses methodological challenges, but also has broader implications for the interdisciplinary nature of emotion research. I exemplified how research on folk emotion concepts and the fundamental base can unfold from an externalist perspective. First, I reconsidered the role of studies on emotion attribution and recognition, suggesting that we interpret these findings as concerning the operation of folk emotion concepts rather than as indicators of first-person experiences. Second, I presented an argument from cross-cultural linguistics about how we can broaden the scope of emotion science beyond English folk emotion concepts. Lastly, I considered research on emotion scripts and norms, which provides an understanding of shared social conventions concerning emotional phenomena.

As a final note, someone might think that the account I have offered relies on an unfair misrepresentation of Mun's original idea concerning ordinary intuitions and emotional experience. While I have offered reasons to believe that, in Mun's account, ordinary intuitions and first-personal experiences are problematically intertwined, perhaps there is some other interpretation that can overcome the problems I have raised. For instance, first-personal experiences might not be as central as I understand them to be to determining the referents of emotion words and expressions, despite textual evidence to the contrary. In that case, I would urge the reader to consider my proposal as an expansion of Mun's ideas, to include insights from semantic externalism, cross-cultural linguistics, and social cognition into the study of the fundamental base. Put succinctly, either my argument overcomes an unclarity that is present in Mun's account, or expands on a correct version of her work.

In conclusion, while Mun's work provides valuable insights into the importance of grounding emotion science in ordinary intuitions, my proposed pragmatic account offers a more comprehensive framework. By moving beyond first-person experiences and instead focusing on public, shared understandings of emotions, we can create a foundation for emotion research that is both empirically tractable and interdisciplinary. This shift not only resolves key philosophical challenges, but also opens new avenues for addressing critical epistemic and ethical questions in the field of emotion research.

5. References

Barger, B., R. Nabi, and L. Y. Hong. 2010. "Standard Back-Translation Procedures May Not Capture Proper Emotion Concepts: A Case Study of Chinese Disgust Terms." *Emotion* 10 (5): 703–711.

Barrett, L. F. 2006. "Are Emotions Natural Kinds?" *Perspectives on Psychological Science* 1 (1): 28–58.

———. 2017. "The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization." *Social Cognitive and Affective Neuroscience* 12 (1): 1–23.

Barrett, L. F. and T. Lida. 2025. "Constructionist Theories of Emotions in Psychology and Neuroscience." In *Emotion Theory: The Routledge Comprehensive Guide. Volume I: History, Contemporary Theories, and Key Elements*, edited by A. Scarantino, 350–87. Routledge.

Bird, A. and E. Tobin. 2022. "Natural Kinds." In *The Stanford Encyclopedia of Philosophy*. Edited by E. N. Zalta and U. Nodelman. Spring 2023 edition. https://plato.stanford.edu/archives/spr2023/entries/natural-kinds/.

Boyd, R. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61: 127–48.

Briggs, J. L. 1970. *Never in Anger: Portrait of an Eskimo Family*. Harvard University Press.

Burge, T. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1): 73–121.

Carnap, R. (1950) 1963. *Logical Foundations of Probability*. The University of Chicago Press.

Díaz-León, E. 2020. "Descriptive vs. Ameliorative Projects: The Role of Normative Considerations." In *Conceptual Engineering and Conceptual Ethics*, edited by A. Burgess, H. Cappelen, and D. Plunkett, 170–88. Oxford University Press.

DiGirolamo, M. A. and J. A. Russell. 2017. "The Emotion Seen in a Face Can Be a Methodological Artifact: The Process of Elimination Hypothesis." *Emotion* 17 (3): 538–46.

Douglas, H. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138 (3): 453–73.

Eickers, G. 2024. "Scripts and Social Cognition." *Ergo* 10 (54).

Eickers, G and J. Prinz. 2021. "Emotion Recognition as a Social Skill." In *The Routledge Handbook of Philosophy of Skill and Expertise*, edited by E. Fridland and C. Pavese, 347–61. Routledge.

Goddard, C. 1996. "The 'Social Emotions' of Malay (Bahasa Melayu)." *Ethos* 24 (3): 426–64.

———. 2015. "The Complex, Language-Specific Semantics of 'Surprise'." *Review of Cognitive Linguistics* 13 (2): 291–313.

Griffiths, P. E. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. University of Chicago Press.

Hareli, S., S. Elkabetz, and U. Hess. 2018. "Drawing Inferences from Emotion Expressions: The Role of Situative Informativeness and Context." *Emotion* 19 (2): 200–208.

Haslanger, S. 2005. "What Are We Talking About? The Semantics and Politics of Social Kinds." *Hypatia* 20 (4): 10–26.

———. (2000) 2012. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" In *Resisting Reality: Social Construction and Social Critique*, edited by S. Haslanger, 221–47. Oxford University Press.

———. (2006) 2012. "What Good Are Our Intuitions? Philosophical Analysis and Social Kinds." In *Resisting Reality: Social Construction and Social Critique*, edited by S. Haslanger, 381–405. Oxford University Press.

Levisen, C. 2019. "Biases We Live By: Anglocentrism in Linguistics and Cognitive Sciences." *Language Sciences*, 76 (November 2019): 101173.

Lindquist, K. A., T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett. 2012. "The Brain Basis of Emotion: A Meta-Analytic Review." *Behavioral and Brain Sciences* 35 (03): 121–43.

Kripke, S. A. 1980. *Naming and Necessity*. Harvard University Press.

Mun, C. 2021. *Interdisciplinary Foundations for the Science of Emotion: Unification without Consilience*. Springer International.

Ngo, N. and D. M. Isaacowitz. 2015. "Use of Context in Emotion Perception: The Role of Top-Down Control, Cue Type, and Perceiver's Age." *Emotion* 15 (3): 292–302.

Putnam, H. 1975. 'The Meaning of "Meaning"'. In *Mind, Language and Reality: Philosophical Papers, Volume 2*, edited by H. Putnam 215–71. Cambridge University Press.

Pust, J. 2024. "Intuition." In *The Stanford Encyclopedia of Philosophy*. Edited by E. N. Zalta and U. Nodelman. Fall 2024 edition. https://plato.stanford.edu/archives/fall2024/entries/intuition/.

Rottenberg, J., J. J. Gross, and I. H. Gotlib. 2005. "Emotion Context Insensitivity in Major Depressive Disorder." *Journal of Abnormal Psychology* 114 (4): 627–39.

Russell, J. A. 2009. "Emotion, Core Affect, and Psychological Construction." *Cognition & Emotion* 23 (7): 1259–83.

Sabini, J. and M. Silver. 2005. "Why Emotion Names and Experiences Don't Neatly Pair." *Psychological Inquiry* 16 (1): 1–10..

Scarantino, A. 2012. "How to Define Emotions Scientifically." *Emotion Review* 4 (4): 358–68.

Strawson, P. F. 1963. "Carnap's Views on Constructed Systems versus Natural Languages in Analytic Philosophy." In *The Philosophy of Rudolf Carnap*, edited by P. A. Schilpp, 503–18. Open Court.

Thomasson, A. L. 2020. "A Pragmatic Method for Normative Conceptual Work." In *Conceptual Engineering and Conceptual Ethics*, edited by A. Burgess, H. Cappelen, and D. Plunkett, 435–58. Oxford University Press.

Touroutoglou, A., K. A. Lindquist, B. C. Dickerson, and L. F. Barrett. 2014. "Intrinsic Connectivity in the Human Brain Does Not Reveal Networks for 'Basic' Emotions." *Social Cognitive and Affective Neuroscience* 10 (9): 1257–65.

Uchida, Y. and S. Kitayama. 2009. "Happiness and Unhappiness in East and West: Themes and Variations." *Emotion* 9 (4): 441–56.

Wierzbicka, A. 1994. "Emotion, Language, and Cultural Scripts." In *Emotion and Culture: Empirical Studies of Mutual Influence*, edited by S. Kitayama and H. R. Markus, 133–96. American Psychological Association.

———. 1999. *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press.

———. 2009. "Overcoming Anglocentrism in Emotion Research." *Emotion Review* 1 (1): 21–23.

———. 2014. *Imprisoned in English: The Hazards of English as a Default Language*. Oxford University Press.

Wikforss, Å. 2008. "Semantic Externalism and Psychological Externalism." *Philosophy Compass* 3 (1): 158–81.