European Journal of Analytic Philosophy EuJAP | 2026 | Vol. 22 | No. 1 | Article 5
https://doi.org/10.31820/ejap.22.1.4

CONCEPTUAL FOUNDATIONS OF *AFFECTIVE* POLARIZATION: TWO NOTIONS OF AFFECT

Aarón Álvarez-González[1] and Aida Roige[2]

[1] University of Valencia, Spain
[2] Universidad Carlos III de Madrid (UC3M), Spain

Original scientific paper – Received: 10/10/2025 Accepted: 02/06/2026

This paper is part of a book symposium on Manuel Almagro's *The Rise of Polarization: Affects, Politics, and Philosophy* guest edited by Miguel Núñez de Prado Gordillo (University of Granada) and Manuel Almagro (University of Valencia)

ABSTRACT

In this paper we argue that the literature on affective states has converged onto two interrelated, but ultimately distinct (and non-derivative) notions of affect: one that characterizes affective states as those with a certain functional profile, and one that characterizes them based on their phenomenology. We argue that, given explanatory gap issues, neither concept can be derived from the other, resulting in a pluralist consensus as to what affect is. We also examine Almagro's proposal to understand affective polarization as a non-individualistic phenomenon and how consistent his usage is with the two presented notions. We conclude that Almagro might have identified a certain form of polarization that, although related, doesn't clearly fit the category of affective polarization.

Keywords: affect; affective states; phenomenology; explanatory gap; emotions; affective polarization.

Correspondence: aaron.estribor@gmail.com; aroige@hum.uc3m.es

1. Introduction

Affective phenomena are an important part of human experience. They inhabit our conscious lives and are commonly summoned in our folk and scientific theories about the mind. Recently appealing to *affect* has become prominent in the explanation of socio-political issues as well. Concretely, the notion of *affective polarization* has been invoked to explain the apparent increase in animosity of many democratic countries' citizens with respect to opposing partisans, especially the United States (Iyengar et al. 2019; Campbell et al. 1960; Hetherington and Rudolph 2015) but also Europe (Reiljan 2020; Wagner 2021).

Affective polarization is defined in Shanto Iyengar's et al. (2019) seminal work as a phenomenon that has two components: *identity* and *affect*. The relevant identity here is *partisanship*, i.e. identification with a political party, as such identity is often adopted in early adulthood and remains stable throughout life, shaping behaviors such as vote choice and how citizens engage with politics (Wagner 2021; Niemi and Jennings 1991). The second component is *affect*: positive towards in-group members, and negative towards out-group members. This treatment of the affective component draws from the long and well-established research on tribal psychology, particularly in-group and out-group affective biases, which are a prominent sort of affective responses so well ingrained in our minds that occur even over trivial or randomly assigned group memberships (see Billig and Tajfel 1973; Dunham et al. 2011; Iyengar and Lelkes 2012).

The methods for measuring affective polarization come from research on social biases. Here, we can distinguish between a person's explicit and implicit biases, depending on whether those get reflected in one's discourse and reflective behavior (in which case they are explicit biases, as when one shows open animadversion towards a partisan of the opposing political party) or whether those covertly affect one's behavior, often outside of one's awareness (in which case they are implicit, as when one inadvertently seats away of the nonpartisan).

A common way to measure explicit biases is by using the so-called "feelings thermometer". The feelings thermometer is a tool aimed to elicit verbal reports from the subjects about how warm or cold they feel towards something—in the case of affective polarization, for example, they could be asked to evaluate their feelings towards in-party and out-party members. The thermometer works by a temperature scale, with lower temperatures corresponding to colder feelings and higher temperatures to warmer ones. The larger the difference between a subject's feeling scores for in-party

versus out-party members, the more affectively polarized that individual is.

Apart from the feelings thermometer, which measures explicit affective attitudes towards in-group and out-group members, other tools are used to capture implicit affect: these behavioral methods are collectively known as "implicit measures". For instance, subjects are tested on economic games such as the trust game or the dictator game and it has been shown that people allocate resources differently depending on whether those go to copartisans or opposing partisans, reflecting a partisan bias in financial allocation (Carlin and Love 2013). Shanto Iyengar and Sean Westwood (2015) also studied the effects of partisan cues when making realistic decisions: in a study, they asked participants to select one of two candidates with similar academic credentials (but different partisan affiliation) for a scholarship. Subjects' decisions showed significant partisan bias. Another recurrent method to capture implicit biases are implicit association tests, based on modifications to the original Race IAT (Greenwald et al. 1998). In the original Race IAT test, one has to put together white faces with good things and black faces with bad things (which would be easy if one feels more positive about White people than Black people), and also put together white faces with bad things and black faces with good things (which would be hard if one feels more positive about White people). The IAT measures the average difference between two sets of reaction times, which shows degree of association. We are thus measuring the valence (either positive or negative) towards members of a particular group, which often dissociates from explicit race prejudice. The Race IAT test can be modified for the political case, to capture affective biases for copartisans and opposing partisans, and indeed, Iyengar and Westwood (2015) did that and found that partisan biases are more prevalent than racial ones.

Both explicit and implicit measures of affective polarization are important insofar as they reveal systematic relations between individuals' affective polarization and socio-political behaviour and events. These effects are not captured by other related notions in the market like the notion of ideological polarization—i.e. divergence of opinion on an ideological spectrum (Reiljan 2020). Thus, we need to invoke affective polarization to capture phenomena like the increase in animadversion or feelings of disliking between Republicans and Democrats between 1960 to 2008, whereas attitudes towards ideological content of copartisans and opposing partisans have remained largely the same (Iyengar et al. 2012).

Recently, Manuel Almagro (2025) has attempted to offer a philosophical conceptual foundation for the phenomenon of affective polarization, its measurements and, in general, the theoretical and explanatory roles to

which it has been put to work. Although we think that Almagro's analysis is illuminating in several respects (for instance, in his emphasis on the role of narratives in reinforcing and establishing political identities), we propose a more classical and, we think, better conceptual foundation. We claim that the literature on affective states, both in cognitive science and philosophy more generally, has converged on two interrelated, but ultimately distinct and non-derivative notions of affect: one that characterizes affective states as those with a certain functional profile, and one that characterizes them based on their phenomenology. We will argue that these two notions already provide us with a solid and pluralistic conceptual foundation that makes sense of the political phenomenon of *affective* polarization, as well as the methods that have been used to study it.

The framework developed here is neutral with respect to many substantive disputes in the philosophy of mind—including disputes concerning the mind-body problem, as well as the metaphysics of consciousness and mental causation—while clarifying the distinct conceptual roles played by phenomenal and functional notions of affect.

Another important aspect of the way in which we understand the notions is that even if they are different and non-reductively related, they share an important assumption about the nature of affects and mental states more generally. Particularly, they are both individualistic in the sense that they talk about mental states of the individual, even if part of what constitutes such states is located outside the skull. In other words: the two notions of affect we present here are individualistic not in the sense of assuming that mental states supervene on intrinsic properties of the individual (as formulated by Fodor 1987, Ch. 2), but insofar they attribute affective states to individuals (as opposed to supra-individual entities, such as public discourse, or non-agential societies or institutions).

This paper has several aims. In the first place, we want to clarify what are the notions of affect at stake in mainstream cognitive science and analytic philosophy, something that has not been done yet to our knowledge. We want to examine what they refer to and how they are related. Secondly, we want to show the fruitfulness of applying them to understand affective polarization. Thirdly, we want to compare the merits of our proposal with the independent proposal of Almagro vis-à-vis the conceptual foundation of affective polarization.

We will proceed by first introducing both notions of affect: the functional notion in section 2, and the phenomenological one in section 3. In section 4 we address the explanatory gap between the two notions and the

possibility, if any, of having a unified account. We then turn to Almagro's book, first reconstructing his view in section 5, to then critically examine it in light of the two classical notions of affect in section 6. We conclude with some thoughts on how Almagro's proposal may have captured largely unnoticed dimensions of polarization, that despite not being—*strictu sensu—affective,* nonetheless enrich our understanding of polarization.

2. The Functional Notion of Affect Emanating from Empirical Cognitive Science

2.1 The Historical Antecedents of the Functional Approach to Affect

Historically, there have been a multitude of terms to describe affective states and processes: passions, appetites, sentiments, experiences, etc. This plurality of terms reflects a plurality of ways of conceptualizing affect, and the mind more generally. A common trend in philosophy was contraposing it to what was "rational", contraposing "cognition" and "emotion", of which the affective was viewed negatively and it was thought to require constraint or guidance by reason.

When psychology began to take shape as a scientific discipline in the late 19th century, the situation didn't change much, and there were a multiplicity of approaches proposing different ways of understanding the mind. This shifted with the advent of cognitive science as its own discipline in the second half of the 20th century. In contrast to psychology's fragmentation about how to best think about the mind, the field of cognitive science emerged, from its very conception, with a unified view of the mind/brain as an information-processing system that can be described at multiple levels of analysis (Marr 1982; Dawson 1998).

Within the classical Marrian (Marr 1982) perspective on cognition, taxonomization of mental states and processes often occurs at the computational level. A computational description involves moving away from the intuitive pre-theoretical or folk psychological notion, identifying what is the information-processing *function* being carried out in terms of inputs and outputs, and what are the constraints for carrying it out (Shagrir 2010). As a result, in the broader physicalist framework of cognitive science, the functional profile of, for example, affect has a central importance when it comes to characterizing it. Indeed, reliance on computational level descriptions to taxonomize mental state types is functionalist in considering that what determines the identity of a mental state is the causally efficacious role it plays in the cognitive system of which it is part (see, e.g., Putnam 1965; Rey 1997; Heil 1998; Levin 2023).

Congruent with such a way of understanding the mind, cognitive science uses a notion of affect that focuses on its functional profile, and thus its relation to the world, other states and behavior—rather than by the way something *feels*. This functionalist notion of affect remains in embodied and extended approaches to cognition: something may be a constitutive part of someone's (an individual agent) mental process even if it is located beyond the skull, if it plays the right functional roles (see, e.g., Clark and Chalmers 1998; Colombetti and Roberts 2015).

Focusing on the functional profile of affective states and processes has proven to be a fruitful framework for understanding it, enabling the recognition of affective processes beyond the situations where folk psychology traditionally confined them. This has allowed us to progress to the point where it is now widely accepted that what were once considered distinct, mutually exclusive categories—thought and affect—are in fact intertwined.[1]

2.2 The Functionalist Characterization of Affect

Cognitive science characterizes affect based on its functional profile, and this is often spelled out in terms of inputs and outputs. Thus, this is the approach we will use in this section to describe the consensus view of what affect is coming from such tradition. What are affect's characteristic inputs and outputs?

The inputs are *appraisals of value* or personal significance a certain thing (understood broadly: e.g., events, objects, properties, changes in our body, situations as represented in our mind) has for us (Moors et al. 2013). These stimuli (e.g., tissue damage in the case of *pain*) are linked, via an appraisal process, to previously stored value. The affective process sets itself in motion when the appraisal links stimuli and value. But what is an appraisal? As Peter Carruthers puts it,

> [A]n appraisal is any mental process (whether associative or judgment-like, whether conscious or unconscious) that links a stimulus (whether real or imagined) to stored personal values (whether innate or acquired), serving to activate them. Some of these values are merely implicit in the "wiring" of innate appraisal mechanisms. This is likely true of the link between

[1] Affective states such as emotions and feelings not only support rational thought but, in some cases, are essential for the operation of higher-order cognition (see, e.g., Damasio's 1994). As a result, the historical contrast between affect and reason has largely been abandoned, as it no longer reflects empirical findings.

> oxygen-deprivation and panic (…). Others are acquired through evaluative learning and are stored in the form of new appraisal conditions in evaluative memory, as when a rat learns that the onset of a light predicts an electric shock, and comes to disvalue the light itself. (Carruthers 2024, 13)

Appraisals of value, thus, put in motion the process that results in emotions, moods and affective states more generally.

What can we say about affect's outputs? Here, there are several outputs occurring in parallel. The consensus is that these outputs are of, at least, three sorts, although some authors may group them differently. One of them are *motor-tendencies*: certain motor plans are activated, whether they are expressive facial behaviour (e.g., dropping one's jaw, closing one's eyebrows) or more complex actions (e.g., stepping away, attacking). These motor plans, if not inhibited, typically result in one's performing the behaviour in question. Part of the empirical task of cognitive science consists in discerning what motor plans are automatically activated by the appraisal, from other behaviours that come further downstream, that may reflect other goals, and be influenced by the affective process, but are not part of the affective process itself.

A second sort of output is *arousal*, or the suite of bodily changes that includes heart rate, breathing rate, sweating, chemicals released into the bloodstream, etc. Bodily arousal is thought to come in degrees: contentment or resignation would be at the low level of the spectrum, and anger and terror at the higher end. Many of these bodily changes occur even before one has a conscious experience of the stimulus (Adolphs and Anderson 2018), and may play a role when it comes to valence. Although every affective state involves a suite of bodily changes, a discussion on the nature of emotions concerns whether such arousal constitutes the core of emotion, or is just an effect or an output (James 1890; Prinz 2004; Scarantino and de Sousa 2021).

A third output of affect is *valence*: valence is an analog magnitude property that can be either positive or negative, and sometimes is described in terms of pleasure and displeasure (when consciously experienced), or reward and punishment. Fear and depressive moods are negatively valenced; joy and optimistic moods have positive valence. Valence is thought to be the "common currency" for decision-making in the mind/brain, serving as a neutral scale allowing for comparisons of qualitatively very different things (e.g., weighing the pleasantness of remaining on one's couch against the displeasantness of going to exercise but with the expectation of future health gains). A source of contemporary discussion concerns the nature of

valence; in particular, whether valence is or not representational, and whether its contents are primarily descriptive, evaluative or imperative (see, e.g., Carruthers 2018, 2023, 2024; Martínez 2022; Martínez and Barlassina 2023). Either way, the functionalist appeals to valence as a dimension or output of affective states.

2.3 Concluding Remarks on The Functionalist Characterization of Affect

In the cognitive scientific tradition, affect is characterized by starting through an *appraisal* of stimuli to stored values, and having as outputs *arousal*, activated *motor plans*, and a *valence* that can be either positive or negative. The notion of affect emanating from empirical cognitive science agrees on as much. When it comes to taxonomizing emotions, moods and felt pleasures, theories diverge on which of these aspects should be given centrality, and how the arrow of the causal or constitutive relationships goes.

3. The Phenomenological Notion

3.1 The Phenomenological Notion of Affect and Some Initial Difficulties

According to the phenomenological notion of affect, affective experience fixes the reference of our affective concepts. Under this approach, affective concepts are phenomenal concepts: concepts used to refer to one's own subjective experience as it occurs (Loar 1990/97). Phenomenal concepts are not exclusively affective. For instance, when I experience red and apply the concept RED to that experience, I deploy a non-affective phenomenal concept. However, in this paper we are going to focus on affective experience and affective phenomenal concepts. Thus, for instance, when I experience anger and apply the concept ANGER to that experience, I deploy an affective phenomenal concept.

One immediate difficulty is that we often use affective concepts in the absence of the relevant experience, as when I attribute anger to someone else. However, this does not undermine the existence of phenomenal concepts (Balog 2009). One may distinguish between basic and derivative uses of phenomenal concepts. Basic uses occur when one refers to one's own presently instantiated phenomenal properties. Derivative uses occur when one refers to past or future experiences, or to the experiences of others, based on assumptions such as the similarity of human minds and the correlation between behaviour and mental states. Under the

phenomenological notion, affective concepts refer to phenomenal affective properties. The concept ANGER, for instance, refers to a particular kind of affective phenomenology. More generally, the concept AFFECTIVE refers to a determinable phenomenal property of which ANGER, JOY, FEAR, and PAIN are determinates.

A classical objection to this approach derives from Wittgenstein's (1953/1965) argument against private language. Since phenomenal properties are accessed privately, it may seem impossible to establish public criteria for the correct application of phenomenal concepts. If language is essentially normative, and normativity requires public standards, then concepts referring exclusively to private phenomenal properties would appear incoherent.

Although influential, this objection is not generally considered decisive today (Balog 2009). Phenomenal concepts continue to play important roles in philosophy of mind and metaethics. In philosophy of mind, they are central to the phenomenal concept strategy, according to which the apparent gap between consciousness and the physical arises from the special epistemic perspective afforded by phenomenal concepts rather than from the existence of non-physical properties properties (see, e.g., Loar 1990; Papineau 2002, 2007). Importantly, the existence of phenomenal concepts has been endorsed by both physicalists (for the phenomenal concepts strategy), like David Papineau (2007), and non-physicalists (as privileged ways of accessing one's *sui generis* phenomenal properties), like David Chalmers (2003). Phenomenal concepts are also important in metaethics. Sentimentalist authors such as David Wiggins (1987), John McDowell (1985), and Justin D'Arms and Daniel Jacobson (2005) argue that at least some evaluative concepts depend constitutively on affective experience. One cannot fully understand concepts such as ADMIRABLE or OFFENSIVE without past or present acquaintance with the relevant affective experiences.

Finally, there are also responses to Wittgensteinian worries. That is, there are alternative theories of meaning that do not require public standards for language to be normative. David Papineau (2011), drawing partly on Ruth Millikan (2000), argues that meaning may be grounded not primarily in publicly checkable rules but in abilities and functions (which explains the normativity of language without dependence on public norms and standards). On this view, phenomenal concepts are concepts whose function is to pick out phenomenal properties by means of the discriminatory abilities of subjects. In this framework for meaning, public checkability is not constitutive, and the locus of the normativity of language is to be found in its functions (e.g., a representation is defective

when its contents depart from the sort of content that representation has the function to represent).

All in all, phenomenal concepts remain a theoretically respectable and influential framework for understanding conscious experience and evaluative thought. The phenomenological notion of affect builds on this framework by treating affective concepts as fundamentally tied to affective phenomenology.

3.2 Brief History of the Notion and its Current Articulation

The phenomenological notion of affect has deep roots in philosophy. One of its clearest articulations appears in Franz Brentano's (1874/1973) account of intentionality. According to Brentano, mental states are characterized by intentionality, that is, by being directed at objects or contents. Importantly, Brentano held that intentionality derives from consciousness itself: mental states are intentional in virtue of being consciously experienced. That is, for Brentano, all intentionality is phenomenal intentionality.

Brentano distinguished three primitive forms of phenomenal intentionality: presentation, judgement, and love/hate. Emotions and desires belong to the latter category. What characterizes these affective states is that they present their objects under an evaluative guise: unlike judgement, which presents things as true or false, affective consciousness presents them as good or bad.

This Brentanian idea has remained influential in contemporary philosophy of mind. Authors associated with analytic phenomenology, such as Uriah Kriegel (2015) and Michelle Montague (2017), argue that evaluative and affective concepts are grounded in affective phenomenology itself. Kriegel (2015), for instance, claims that phenomenal modes of presentation are explanatorily prior to evaluative concepts. According to this view, the concept GOOD derives from the mind's capacity to present objects as good.

Related ideas appear in contemporary sentimentalist metaethics. Sentimentalists maintain that evaluative concepts are conceptually dependent on affective experience. Dispositional sentimentalists hold that evaluative properties are those capable of producing certain affective experiences in certain conditions (see, e.g., Firth 1952), whereas normative sentimentalists claim that evaluative properties are those that merit such experiences (see, e.g., McDowell 1985). In both cases, affective phenomenology plays a constitutive role in fixing evaluative concepts.

Thus, a central commitment of the phenomenological notion is that affective concepts are rooted in affective phenomenology. This does not mean that one experiences the relevant phenomenology every time one deploys an affective concept. Rather, it means that the meaning of the concept depends constitutively on paradigmatic cases of phenomenal acquaintance.

Another important feature of the phenomenological notion is that affective states are often conceived as phenomenally intentional states. Emotions are not mere raw feelings, but conscious states directed at objects or situations under evaluative modes of presentation. Fear seems to be directed at danger and sadness at personal loss. Because they possess intentional content, they are rationally assessable. Fear may be irrational if it is directed at harmless objects. The same goes for sadness if directed at the loss of something of dispensable value, for instance. Contemporary discussions therefore frequently analyse relations of epistemic support and rational coherence between emotions and evaluative beliefs (Helm 2001; Tappolet 2016; Micheli 2010). Consequently, sharp oppositions between affect and cognition are difficult to sustain within this framework.

Finally, the phenomenological notion conceives affects as phenomenally valenced experiences. Affects feel positively valenced (joy, pleasure), negatively valenced (pain, fear, depression), mixed (nostalgia), or affectively neutral. In contrast with some functional approaches, felt valence is here treated as an essential feature of affective states. Moreover, these valenced experiences are closely connected to motivation: negative feelings tend to motivate avoidance or reduction of what produces them, whereas positive feelings tend to motivate approach and preservation.

3.3 Concluding Remarks on the Phenomenological Notion of Affect

According to the phenomenological notion, affective concepts are phenomenal concepts referring to affective phenomenal properties. This framework has played an important role in philosophy of mind and metaethics, particularly in discussions of phenomenal consciousness, intentionality, and evaluative thought.

The phenomenological notion also treats at least some affects as phenomenally intentional and rationally assessable states. This weakens traditional dichotomies between affect and cognition. Furthermore, it regards felt valence as an essential aspect of affective experience and as closely connected to motivation.

Given its theoretical importance and continued influence, any philosophical analysis of affective polarization must engage with the phenomenological notion of affect, whether by endorsing it, rejecting it, or contrasting it with alternative frameworks such as the functional notion discussed in the previous section.

4. The Relation Between the Two Notions

After characterising the two notions it becomes natural to wonder about their relation. In this section we will show how both notions maintain a degree of independence from each other supporting thus our diagnosis that there is a pluralist consensus as to what the available conceptions of affect are.

The first thing to notice is that the two notions are conceptually distinct in the sense that there is no *a priori* entailment between them. It is not conceptually incoherent for someone to endorse the phenomenal notion of affect while rejecting the functional one, or vice-versa. Crucially, neither notion can be straightforwardly deduced from the other; each requires adopting a distinct epistemic stance. For starters, the phenomenal notion refers to seemingly non-relational experiential properties, implies a first-person perspective and is devoid of descriptive links, since phenomenal concepts work as type-demonstrative recognitional concepts (Loar 1990/97). By contrast, the functional notion emanating from cognitive science appeals to the role of affective states and processes in relation to other mental states and processes, which implies a third person perspective. This conceptual independence is a particular instance of the more general phenomenon of the Explanatory Gap between phenomenal and non-phenomenal concepts—a gap that, many have argued, has no solution (see, e.g., Shoemaker 1975; Tye 2006 for discussion).

Despite this gap, the two notions largely overlap in the instances of affect they capture. Both would pick up the paradigmatic examples of affective polarization discussed in the introduction. For instance, your conscious fear of the rise of right-wing political moves would be recognized by both the phenomenal and the functional notions of affect—albeit in a different way.

However, the fact that both notions take as essential to affect differing properties opens the door for them to have different extensions. According to the functional notion, there may be affective states and processes occurring at the sub-personal level. For instance, when choosing among two options, the introduction of a third, less optimal alternative (the decoy)

may subpersonally alter the perceived (valenced) value of the two original options, thus constituting what may be cases of affective processing without the corresponding felt experience (Howes et al. 2016). Conversely, some mental states may involve affective phenomenology without exhibiting the outputs typically associated with affective states in the functionalist notion: certain special emotions like those toward fictional entities, the so-called contemplative emotions (Tappolet 2016), or aesthetic emotions might fall under this category. In any case, we do not need to commit to the existence of actual cases where the extensions of the two notions diverge: we only need to acknowledge that there is no inconsistency in claiming that they could.

Another important aspect between the relation of the two notions is that they are in principle compatible and, hence, one is not in principle forced to choose between them. One can suspend judgement as to what, if any, of these notions is more fundamental. However, we find different trends in the literature trying to reduce one to the other. These trends are again specific cases of more general debates in the philosophy of mind. According to intentionalism (see, for instance, Tye 1995) phenomenal properties supervene on intentional properties, and the latter can be reduced to physical states. Under this paradigm, affective phenomenal concepts would refer to intentional properties ultimately reducible to physical ones. According to non-reductivist accounts about phenomenology and intentionality, like the *phenomenal intentionality research program* (Kriegel 2013), phenomenology is irreducibly prior to intentionality and there is no way of naturalizing the former through the latter. Under this paradigm, affective phenomenal concepts would refer to *sui generis* experiential properties.

Both notions are relatively autonomous, which helps explain why they have played different roles in different domains of science and philosophy. Thus, the functional notion is ubiquitous in cognitive science and empirically informed philosophy. By contrast, as has been already pointed out, the phenomenal notion is more common in classic philosophy of mind, for instance, the analysis of the contents of the conscious mind, and metaethics.

The fact that the two notions are in principle compatible plus the fact that there are no conclusive reasons to get rid of one or the other make possible that there is a stable situation in which, lacking a common and unifying paradigm, both notions are used in contemporary discussions of affective phenomena.

5. Almagro's Notion of Affective Polarization

It seems that the default position among empirical scientists when it comes to interpreting the empirical data is assuming one or both of the two notions just presented are in play. For instance, the default reading of the results in the feeling thermometer is that individuals are reporting their own phenomenal states, through phenomenal concepts, and that (often) they are sincere and in a good epistemic position to know their own internal states. The same goes for the interpretation of overt behavior and implicit measurements: for example, one's behavior in economic games (e.g., inadvertently allocating financial bonuses to co-partisans and monetary penalties to sympathizers of the opposing party) is partly explained by one's affective states understood functionally, as that which causally impacts one's behavior below one's conscious awareness and results on such polarized behavior. Part of our diagnosis in the previous sections is that it is extremely difficult to answer the question of how these two notions relate, given that this is an aspect of the so-called *Explanatory Gap* between the phenomenal and the physical.

In contrast with these readings, Almagro proposes a new interpretation of the data, which he calls the "expressivist diagnosis". He takes the default descriptivist assumptions behind an individual's usage of affective/evaluative terms to lead to what he—echoing what he takes to be a related position in the philosophy of language[2]—calls the "emotivist diagnosis". According to the emotivist diagnosis, affective polarization consists in the combination of one's political identity with a strong affective attachment to it. As Almagro reconstructs the position, the emotivists' diagnosis portrays affectively polarized individuals as akin to football supporters,

[2] In Almagro's own words: "I call this diagnosis "emotivist" because it shares relevant features with what has been called "emotivism" in the philosophy of language, particularly the canonical reception of the metaethical position concerning the meaning of ethical statements attributed to the philosopher Alfred Jules Ayer (1952), among others. According to emotivism, when we make ethical claims such as "Offending others is morally wrong" or "It is good to avoid offending others", we are not expressing our beliefs that things in the world are such and such, but rather our emotional attitudes—our approval or disapproval toward something. In other words, these statements are akin to saying "Boo to offending others!" or "Hurrah for not offending others!", which lack truth-values because they do not express the speaker's beliefs or represent how the world is; they merely convey the speaker's preferences and emotions" (Almagro 2025, 50).
Thus, according to Almagro, a realist interpretation of the empirical results on affective polarization is one that assumes there are certain mental (affective) states explaining people's overt behaviors and their scores in implicit and explicit measures of affect. In contrast to the realist reading, which takes such internal mental affective states as the primary phenomenon of interest, the primary locus of Almagro's analysis is public affective (that is, evaluative) language.

insofar as their attachment to political identities is sustained not by reasons, but by non-rational affective dispositions—positive toward the in-group and negative toward the out-group. This is what Almagro's writes in this regard:

> Given the extraordinary emphasis that the two-dimensional approach places on partisan identity and its capacity to generate intergroup animosity and conflict, it is not surprising that diagnoses from this framework frequently resort to comparisons with the dynamics between fans of rival sports teams. According to a widely accepted diagnosis favored by the two-dimensional approach, affective polarization resembles "two teams fighting over a trophy" (Mason 2018, 4). I call it the emotivist diagnosis. (…)
> According to the emotivist diagnosis, partisanship is a "helluva drug" (Klein 2016), and citizens of polarized democracies are "intoxicated partisans" who "arbitrarily form psychological attachments to their party and blindly support that party in elections, regardless of the candidates' policy positions, priorities, or abilities" (Fowler 2020, 142). (Almagro 2025, 49)

This emotivist diagnosis leads, according to Almagro, to four undesirable consequences: 1) polarized individuals hold political views not on the basis of rational reasons, but on non-rational affects; 2) polarized individuals are insincere in the sense that they do not really *believe* in the political positions they profess, but are merely affectively attached to them; 3) they thus lack, properly speaking, beliefs about the relevant political matters; and, consequently, 4) polarized individuals do not genuinely disagree with each other, in the same way in which sports team supporters cannot be said to truly disagree with each other. All these consequences taken together lead to the claim that affectively polarized individuals are irrational. Almagro thinks this conclusion should be resisted and, hence, that we should dispense with the assumption that personal reports in experiments such as the feelings thermometer *describe* one's affective state, but rather take such reports as primarily performative. Thus, the primary locus of Almagro's analysis becomes public affective (that is, evaluative) discourse, which reflects evaluative commitments.

Contrary to the emotivist diagnosis, Almagro thinks that we should shift towards what he calls "the expressivist diagnosis" (Almagro 2025, 7). According to the expressivist diagnosis, individuals do not report their feelings in tasks like the feeling thermometer. Rather, what they do is *voicing their commitments*—that is, making linguistic commitments towards certain ways of interpreting the social and political world.

Almagro is not assuming that a person has a pre-existing internal mental state corresponding to what he voices, but rather, that the person has a disposition to make abstract linguistic commitments. In this sense, Almagro argues that the phenomenon of affective polarization is not something that must be explained individually (for instance, in terms of subjects' dispositions to identify with certain views or dislike members of the out-group) but rather, it is a higher-order phenomenon that has to do with the distribution of affects at the societal level. Individuals are best understood as nodes in this social structure of commitment-making, supporting it. They may or may not instantiate certain phenomenal feelings and/or functional profiles, but, crucially, they are committed to the *abstract* political views they embody when employing public affective (i.e. evaluative) language in tasks like the feeling thermometer.

> At first glance, this idea might seem to be already covered by the two-dimensional view, and to some extent, it is. However, there's a crucial distinction to be made here. If affective polarization includes not just identity and emotions but also an attachment to specific discourses and narratives, then citizens' political statements—even those concerning their feelings and emotions—may reflect their allegiance to particular ideologies, identities, and parties, rather than merely reporting their personal emotional experiences. This suggests that emotions alone are not the only relevant factor in characterizing affective polarization. I will expand on this point later, but for now, let's introduce this fifth dimension.
> Voicing our Attachments: Citizens' speech about their opinions, feelings and emotions about polarized topics voices their level of attachment to certain narratives, ideologies, and parties. But it's been qualified that this could be a result of a mismatch between abstract and concrete judgments that polarized people tend to exhibit. (Almagro 2025, 55-6)

This expressivist shift avoids the pitfalls of the emotivist diagnosis because it does not depict individuals as just affectively (and non-rationally) attached to a certain political identity. Rather, individuals and their affects are part of a social structure of affective oppositions (e.g. in-group, out-group) informed by certain *narratives* to which these individuals grant *high credence* (i.e. degree of subjective probability) and they *perform* in their lives generally and in the feeling thermometer task particularly. In this sense, Almagro thinks, affectively polarized individuals are at least as rational as non-affectively polarized individuals. Affectively polarized individuals believe in their political cosmovision, at least in the abstract,

on the basis of reasons narratively articulated[3] and, hence, what they do when they affectively dislike the out-group is not basing their political views on those affects, but rather commit to those political views through their use of affectively charged language, thereby genuinely disagreeing with one another.

> Now we are in a position to further elaborate on the expressivist diagnosis. According to it, affectively polarized groups are in genuine disagreement, in the sense that they support and adhere to opposing discourses, which are associated with different policies and ways of living. The adherence to such discourses can be rational, in the sense that people supporting them are not necessarily epistemically defective and have reasons and arguments to defend their positions. They can also be sincere, in the sense that they don't deliberately misrepresent their beliefs. Their political statements might report their professed beliefs in the abstract, but they still serve an expressive function. Their support for certain ideas, as well as the expression of certain emotions, is an indication of how convinced they are of the truth of certain discourses, mostly in the abstract. Thus, their claims don't merely express their emotions or preferences; these claims express how attached they are to certain narratives and identities. (Almagro 2025, 61)

Thus, Almagro's recommendation is replacing the *two-dimensionalist framework* (which is the theoretical basis for the emotivist diagnosis) with a *multi-dimensionalist framework* (which is the theoretical basis for the expressivist diagnosis) that allows for genuine disagreements by virtue of incorporating not only affect and identity, but also narratives in which the individuals have high credence and through which they articulate their commitments (for further discussion of this aspect of Almagro's framework, see Pedrini and Marchetti 2026, in this issue).

[3] Narratively articulated in the public discourse. As Almagro puts it: "The emphasis shifts away from individual rationality and toward the effectiveness of the narratives and strategies used to manipulate and divide public opinion. This approach provides an alternative way to argue that the rise of polarization is a non-epistemic phenomenon, underscoring that it is more about the design of persuasive strategies than about individual cognitive processes" (Almagro 2025, 134).

6. Why We Should Keep the Interpretation of Empirical Results as Reflecting the Mental States Captured by Our Two Notions

In the last section, we presented Almagro's proposal to characterize affective polarization. Now, we are going to argue that we should keep the conceptual foundations of *affective* polarization closely tied to our two notions instead.

6.1 Why One Cannot Avoid the Metaphysics of Affect When It Comes to Analyzing Affective Polarization

Almagro explicitly claims that his expressivist diagnosis frees the theorist from having to get involved in the metaphysics of the phenomenon of affective polarization. In Almagro's own words:

> Another essential feature of expressivism is *its commitment to avoiding the metaphysical question of the domain of discourse it analyzes*. We do not need to appeal to moral or aesthetic facts to understand what we are doing when we say that something is morally right or beautiful. Similarly, we do not need to appeal to phenomenal events to understand the information conveyed when people claim that we have warm or cold feelings toward a certain issue, or that they feel animosity toward those who belong to opposing groups. (Almagro 2025, 62, our emphasis)

Almagro seems to be claiming not only that reports of certain feelings toward a particular issue or group (such as animosity) do not necessarily correspond to the subjects' underlying mental states, but also that the presence of such states is not what matters for understanding the linguistic behaviour of polarized subjects. Rather, what matters, on his view, is the discursive commitment to which subjects bind themselves through their utterances, considered independently of the psychological states that may or may not accompany them. He maintains that there is no need to appeal to mental states in order to understand the content of what is being expressed. Consequently, Almagro opts to sidestep altogether the metaphysical implications of the discourse he analyzes. The real question for Almagro is, then, whether one can plausibly make sense of people's reports without presupposing that there is something that they are *meant* to represent, namely: the mental states instantiated in the reporters.

Almagro's expressivism construes reports of affective attitudes primarily as *performative commitments* to act or respond in certain ways. We believe this is mistaken. If such utterances were primarily performative commitments,

we would expect a comparatively robust connection between the attitudes expressed and the subject's subsequent behaviour. Yet this is often not the case—as we know to happen empirically (see, e.g., Green et al. 2007; Kawakami et al. 2009; Lee et al. 2023), and as Almagro himself acknowledges in discussing his father's case. Individuals frequently report egalitarian attitudes, while failing to act accordingly; other times, they manifestly hold hostile or "cold" attitudes toward a group while failing to behave consistently with the practical commitments that such attitudes would seem to involve.

By contrast, if these utterances are understood in the more ordinary way, namely as reports of occurrent mental states, the mismatch between reported attitudes and subsequent behaviour becomes considerably less puzzling. Reports of mental states are not primarily evaluated in terms of whether they successfully guide conduct, but in terms of whether they accurately describe the subject's psychological condition at a given moment. Affective states, especially fleeting or context-sensitive ones, may or may not translate into congruent behaviour depending on the interaction of multiple motivational and cognitive factors.

This difference can be illuminated in terms of direction of fit. Commitments, much like intentions, plausibly exhibit a world-to-mind direction of fit: they aim at bringing the world into conformity with the attitude expressed. Reports of affective states, by contrast, exhibit a mind-to-world direction of fit: they are correct insofar as they accurately represent the subject's current psychological condition. If this is right, it is far less surprising that individuals sometimes report transient feelings of hostility or distance that fail to structure their overall behavioural dispositions than it would be if they had genuinely undertaken practical commitments corresponding to those attitudes.

Consider, for example, an individual who reports feeling "cold" toward a social group in the abstract while nevertheless expressing warmth and sympathy toward a particular neighbour belonging to that same group shortly thereafter. Such a case is not especially problematic if affective utterances are understood as reports of occurrent mental states. Reporting one's psychological condition at a given moment does not normally commit one to maintaining that condition over time; one is not thereby normatively constrained to continue feeling the same way in future contexts. By contrast, performative commitments seem to carry a stronger normative and diachronic dimension. To undertake a commitment is, at least under minimal assumptions of rationality and practical consistency, to bind oneself in ways that extend beyond the immediate moment of utterance and to sustain a comparatively stable orientation toward future

conduct. If this is correct, then the observed instability between reported affective attitudes and subsequent behaviour is considerably easier to accommodate on a descriptive interpretation of affective reports than on Almagro's performative account.

Another important drawback of Almagro's framework is that it is too tied to explicit linguistic behaviour. Affective polarization extends beyond linguistic practice and occurs as well in non-discursive behavior, such as choices in economic games or willingness to embrace certain individuals to one's own family. Explaining this overall phenomenon requires appealing to affective states, and is something that the expressivist account, by itself, cannot adequately provide.

Given the foregoing, different interpretations of affective reports appear to carry different metaphysical commitments, some of which may be more abductively adequate to the phenomenon than others. If affective reports are understood primarily as public commitments constituted through the social practices surrounding language, then affective polarization would seem to be, fundamentally, a social or collective phenomenon. By contrast, if such reports are understood as descriptions of individuals' psychological states, then affective polarization would appear to be grounded, at least in part, in the mental life of individuals.

Either way, we contend that an adequate conceptual foundation for affective polarization cannot be provided without engaging, at least to some extent, with the metaphysics of the phenomenon. This is not merely a theoretical concern. If affective polarization indeed poses a threat to democratic life, as many authors have argued, then understanding its underlying nature becomes essential for designing effective interventions. In particular, it matters whether affective polarization is primarily a phenomenon emerging at the collective level of social practices and discursive commitments or, instead, one rooted in the affective states of individuals.

Almagro may believe that metaphysical questions can be bracketed because his primary concern is practical—namely, how to intervene in the phenomenon. However, successful intervention presupposes at least a minimal understanding of what exactly is being intervened upon. For instance, if affective polarization is ultimately grounded in individuals' affective mental states, then any effective intervention must, directly or indirectly, bring about changes in those states. Conversely, if the phenomenon is primarily constituted by public commitments embedded in social practices, then interventions targeting discursive norms and collective dynamics may be more appropriate. In either case, metaphysical

considerations are not dispensable, since they shape both the explanatory framework and the kinds of interventions that can plausibly succeed.

Indeed, even though Almagro explicitly distances himself from taking a metaphysical stance, his position appears to presuppose that affective polarization is not essentially an individual phenomenon but a collective one. But if this is correct, then Almagro is engaging in metaphysics after all: his expressivism appears to endorse or imply a negative metaphysical claim, i.e. that affective polarization is not essentially related to *individual* mental states. Indeed, he states this quite explicitly:

> Polarization, on the other hand, is a social rather than an individual phenomenon: *There cannot be polarized individuals in isolation,* whereas there can be individuals who hold extreme views in isolation. (Almagro 2025, 103, our emphasis)

According to this, affective polarization is not appropriately predicated of individuals but of groups. That is, mental states of individuals are not sufficient for affective polarization and this, we think, is in tension with the very notion of *affect*, as it is understood according to our two notions. How can *affective* polarization be about *affects,* when according to Almagro what matters most is the linguistic performance of evaluative commitments? It seems that in his expressivist diagnosis affective states are not doing real explanatory work. We contend this is misguided for two reasons: it departs from our pre-theorethical intuitions, which back up the two notions, and it is methodologically problematic. In the next two subsections we develop these two points in turn.

6.2 Intuitive Considerations in Favor of the Two Notions

It seems almost a truism nowadays that affective polarization has increased in Western democracies. This suggests that there is a folk grasp of what affective polarization looks like, that prima facie allows people to recognize it if present. Indeed, probably something like such folk perceptions motivated the empirical study of polarization, renovating its general interest. It is thus worth knowing whether Almagro's proposal respects the intuitions that fueled the study of polarization in the first place. We think it does not. Recall that, according to Almagro, affective polarization is a phenomenon predicable of groups but not of individuals. This, however, clashes with our intuitions about the phenomenon. Consider the following vignettes.

> *The polarized brain in a vat.* Amber was born as a brain in a vat. The machines feed her information about what seems to be

> the political structure of her world, which seems to be divided between Republicans and Democrats. As a result, Amber became highly involved in the Republican party and developed strong animosity towards what she thinks are the members of the Democrat party. Of course, there is nothing in reality that corresponds to Amber's representations about the political structure of her world.

What would we say about Amber's case? That Amber is polarized is, at least, intelligible, even though she is alone in the world. So it would seem that we get some support that it is perfectly conceivable to have isolated individuals who are polarized. One could resist that conclusion by arguing that we may not have very strong intuitions about how to judge this far away skeptical scenario, or that our intuitions are not reliable because they exploit problematic (internalist) assumptions about how language, and representations more generally, work. For instance, what does "Democrat" mean in a Matrix-like scenario?

Even more, given his Wittgenstenian commitments, Almagro would probably discard this as a case of polarization on two grounds: (1) that there cannot be a language in isolation, given that language needs public checkability (and without language there is no affective polarization, since many of its purported dimensions, like narratives, are missing), and that (2) that the brain in a vat case presented lacks a social structure, and there cannot be affective polarization without it.

Thus, we need to consider another vignette that neutralizes these defects of the first one, while still showing that it is possible to have affectively polarized individuals in isolation:

> *The Republican who fell asleep before the apocalypse.* John the Republican falls asleep. A moment after John's falling asleep, the apocalypse occurs. All the inhabitants of the planet Earth die apart from John. When John wakes up he thinks about how much he hates Democrats and that he is going to start a campaign to damage the reputation of the Democrat leader. After thirty minutes of brushing his teeth, taking a shower, getting dressed, and so on, John leaves his house and discovers the painful truth: everything is destroyed, not only Democrats, but also Republicans, his family, etc.

In this new scenario, Almagro's possible comeback would be neutralised. John enjoys a language which was publicly acquired and checked (recall that Wittgenstein allows Robinson Crusoe to have a language to the extent

that that language was previously acquired in a linguistic community) and there was a society that John was part of. Now, John seemingly has certain beliefs (that there are Democrats, that he dislikes them, that he did not like his daughter to marry one of them, etc.) and certain consistent affective states (animosity towards Democrats, the desire of not sharing time with Democrats, etc.). According to Almagro, however, John's individual mental states do not suffice for John to be polarized.

Before the apocalypse, John was clearly polarized—Almagro and us would agree on this point—and perhaps he will cease to be polarized after discovering that it has happened. But how should we describe John's state after the apocalypse when he still did not know that it had happened? We would like to say that he was polarized while brushing his teeth while thinking how much he dislikes Democrats. The intuition here is that an imperceptible change for John wouldn't change his status as a polarized individual: John retained all the dispositions and occurrent states that would have explained his counterfactual behaviour had the world not come to an end. Yet, according to Almagro, it is not possible to have polarized individuals in isolation. So not only Almagro fails to cohere with our intuitions without polarization—but he also provides no explanation as to why shall they differ in cases like these.

This is not just some ingenious philosophical exercise. It highlights that our intuitions suggest that individuals' affective states suffice for *affective* polarization to occur. Thus, ignoring them and their explanatory power—as the expressivist account does—is theoretically inappropriate.

6.3 Methodological Considerations in Favour of The Two Notions

Almagro also departs significantly from the notion of affective polarization employed by political scientists. Political scientists measure affective polarization with the empirical tools that come from cognitive sciences (scale ratings, answers to questionnaires, implicit association tests, etc.) under the assumption that the results of these methods should be taken as they are customarily taken in the cognitive sciences, namely: as *measuring mental states of individuals* which are nomologically related to their behaviour.

Yet according to Almagro, when it comes to affective polarization, we should suspend this assumption. Thus, he imposes a huge difference on how we should interpret implicit measurement tools in the context of polarization and other investigative contexts where these tools are also being used. So, for example: while the feelings thermometer is taken as a measurement tool to track both the direction and the intensity of certain

mental states of individuals in sociology and political science (see, e.g., Wilcox et al. 1989; Simonsson 2022), under Almagro's interpretation it would track instead *commitments to act* when it comes to studies on affective polarization. But how are these *affective?* For another example: while the IAT tests are thought to capture individuals' implicit biases in psychology, cognitive science, and political science (see, e.g., Greenwald et al. 1996), when they are used as tools for assessing polarization is not clear what Almagro would consider they track. Within Almagro's framework, what are these implicit measurement tools supposed to measure? Implicit commitments? But given that Almagro puts that much emphasis on *performative* commitments *in the abstract*, this would look like an utterly different sort of beast. Allocation biases or implicit biases are not easily assimilable to, nor explainable by, the idea of making a commitment. This re-interpretation of methodological tools strikes us as unjustified and problematic.

Bearing in mind the above (i.e., the special interpretation of the results in the context of affective polarization), Almagro's position becomes unstable and threatened by a dilemma. On the one hand, Almagro could claim the special status of affective polarization—but, then, he should develop new methods to measure it, ones designed to capture non-individualistic phenomena. Using tools meant to reflect people's internal states to capture something else (for example, a social phenomenon, or a public commitment to act) is as inappropriate as using a thermometer to capture atmospheric pressure. Even if the two attributes were somewhat related, we cannot know what this relation consists in without an independent method designed to capture the second, which he isn't providing.

On the other hand, Almagro can claim that there is continuity between affective polarization and the individualistic phenomena measured by the cognitive science with those methods. But then, given that the customary interpretation is constitutive of the usage of the measurement tool, he would have to face the pressure to keep the individualistic notions presupposed in the customary interpretation.

In a nutshell: Almagro must either accept the individualist implications of the methodological tools he is using, or he has to develop new methodologies to capture the non-individualistic phenomenon he wants to capture. Else, he would be being inconsistent in his treatment of the empirical studies of the field, or he would have to appeal to some sort of exceptionalism for the phenomenon of affective polarization, which he hasn't given any reason to accept.

7. Conclusion

Almagro has identified what could possibly be considered a form of polarization neither entirely affective nor ideological, but rather *discursive*; and that is worth paying attention to and study further. It's likely that the sort of polarization Almagro talks about (more centered around public discourse and commitments of speakers) captures a missing piece of the analysis of polarization: the point of interaction between the ideological and affective polarization. We thus regard *The Rise of Polarization: Affects, Politics and Philosophy* (Almagro 2025) as an important contribution that overcomes the dichotomy between ideological and affective polarization which has predominated the literature until today.

However, we think this should be done without taking away from the classical notions of affect. Maintaining a clear notion of *affect* is of uttermost importance to make sure different researchers working from different fields are actually addressing the same thing in the world. There are already two notions of affect playing this role, we argued: the phenomenological notion and the functional one. The two notions are conceptually independent and cannot be derived from one another, given explanatory gap issues.

The best way to understand the *affective* side of polarization, we contend, comes from these two classical notions from cognitive science and philosophy of mind, which have a proven trajectory of fruitfulness. Both such notions present *affect* as something that is predicated about, or attributed to, *individuals*.

In this sense, these two traditional notions clash against Almagro's insistence that affective polarization is a phenomenon that is primarily located at the social or discursive level, and must be tackled non-individualistically. As we argued in Section 6.2, the claim that affective polarization cannot be predicated of individuals departs from our pre-theoretical intuitions, as illustrated by the thought experiments presented there. It also diverges from the way experts interpret and employ the results generated by the methods used to study polarization. If affective polarization is measured with the methodologically individualistic tools of cognitive science, the presumption is that there is ontological continuity in the phenomenon so measured. If one rejects the continuous nature of affective polarization with the other sorts of affective phenomena measured by the cognitive sciences, the presumption is that the old methods are no longer appropriate and new ones should be developed. Almagro occupies, therefore, an unstable position: he rejects the continuous nature of affective polarization with the other sorts of affective

phenomena measured by the cognitive sciences, but he remains faithful to the old measurement tools.

Nonetheless, it is true that there are aspects of polarization in general that remain theoretically underexplored, and Almagro's analysis brings attention and a much-needed articulation to these. We hope that, in the coming years, this perspective will gain broader recognition, and that discursive or expressive polarization—closely related to, yet distinct from, affective polarization—will emerge as an established object of analysis in its own right.

Acknowledgments

Authors are grateful for the ongoing support and feedback of the two editors of this special issue, as well as for two anonymous reviewers whose comments helped improve the paper. Any remaining misgivings are, obviously, our own.

Funding Information

In Roige's case, work for this paper was funded by the research projects "Cognición Mecanicista: ¿es posible integrar las ciencias de la mente y la computación?" and "Reproche proléptico y emociones morales en contextos de injusticia structural", both financed by the Programa Propio of UC3M. In Álvarez-González's case, work for this paper was funded by a Juan de la Cierva grant from the Ministerio Español de Ciencia, Innovación y Universidades)/Grant Number: JDC2023-050540-I. 2024–2026.

Author Contribution

Equal contribution; authors' names are listed in random order.

REFERENCES

Adolphs, Ralph, and David J. Anderson. 2018. *The Neuroscience of Emotion*. Princeton: Princeton University Press.

Almagro, Manuel. 2025. *The Rise of Polarization: Affects, Politics, and Philosophy*. London: Routledge. https://doi.org/10.4324/9781003400448

Balog, Katalin. 2009. "Phenomenal Concepts." In *The Oxford Handbook of Philosophy of Mind*, edited by Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter, 292–312. Oxford: University of Oxford Press.

Brentano, Franz. 1874/1973. *Psychology from Empirical Standpoint*. London: Routledge.

Carruthers, Peter. 2018. "Valence and Value." *Philosophy and Phenomenological Research* 97 (3): 658–680.

Carruthers, Peter. 2023. "On Valence: Imperative or Representation of Value?" *The British Journal for the Philosophy of Science* 74 (3): 533–553.

Carruthers, Peter. 2024. *Human Motives: Hedonism, Altruism, and the Science of Affect*. Oxford: Oxford University Press.

Carruthers, Peter. 2025. *Explaining our Actions*. Cambridge: Cambridge University Press.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19. https://doi.org/10.1093/analys/58.1.7

Colombetti, Giovanna, and Tom Roberts. 2015. "Extending the Extended Mind: The Case for Extended Affectivity". *Philosophical Studies* 172: 1243–1263. https://doi.org/10.1007/s11098-014-0347-3

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Grosset/Putnam.

D'Arms, Justin, and Daniel Jacobson. 2005. "Two Arguments for Sentimentalism." *Philosophical Issues* 15 (1): 1–12. https://doi.org/10.1111/j.1533-6077.2005.00050.x.

Dawson, Michael R. 1998. *Understanding Cognitive Science*. Cambridge: Blackwell.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in The Philosophy of Mind*. Cambridge, MA: MIT Press

Green, Alexander R., Dana R. Carney, Daniel J. Pallin, Long H. Ngo, Kristal L. Raymond, Lisa I. Iezzoni, and Mahzarin R. Banaji. 2007. "Implicit Bias Among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients." *Journal of General Internal Medicine* 22 (9): 1231–1238. https://doi.org/10.1007/s11606-007-0258-5

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology*, 74 (6): 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Helm, Bennett W. 2001. *Emotional Reason: Deliberation, Motivation, and the Nature of Value.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511520044

James, William. 1890. *Principles of Psychology.* Henry Holt.

Kawakami, Kerry et al. 2009. "Mispredicting Affective and Behavioral Responses to Racism." *Science* 323: 276–278. https://doi.org/10.1126/science.1164951

Kriegel, Uriah, ed. 2013. *Phenomenal Intentionality.* Oxford: Oxford University Press.

Kriegel, Uriah. 2015. *The Varieties of Consciousness*. New York: Oxford University Press.

Kent M. Lee, Kristen A. Lindquist, and B. Keith Payne. 2023. "Constructing Explicit Prejudice: Evidence from Large Sample Datasets." *Personality & Social Psychology Bulletin* 49 (4): 541–553. https://doi.org/10.1177/01461672221075926

Levin, Janet. 2023. "Functionalism." *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/sum2023/entries/functionalism

Loar, Brian. 1990/97. "Phenomenal States." *Philosophical Perspectives* 4: 81–108; reprinted with revisions in *The Nature of Consciousness*, edited by Ned Block, Owen Flanagan, and Güven Güzeldere, 597–616. Cambridge, Mass.: MIT Press.

Heil, John, ed. 1998. *Philosophy of Mind: A Contemporary Introduction*. New York: Routledge.

Howes, Andrew, Paul A. Warren, George Farmer, Wael El-Deredy, and Richard L. Lewis. 2016. "Why Contextual Preference Reversals Maximize Expected Value." *Psychological Review* 123 (4): 368–391. https://doi.org/10.1037/a0039996

Marr, David. 1982/2010. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, Mass: MIT press.

Martínez, Manolo. 2011. "Imperative Content and the Painfulness of Pain." *Phenomenology and the Cognitive Sciences* 10 (1): 67–90.

Martínez, Manolo, and Luca Barlassina. 2023. "The Informational Profile of Valence: The Metasemantic Argument for Imperativism." *The British Journal for the Philosophy of Science.* https://doi.org/10.1086/726998

McDowell, John. 1985/1998. "Values and Secondary Qualities." In *Ethical Theory* 1, edited by James Rachels, 210–226. New York: Oxford University Press.

Micheli, Raphaël. 2010. "Emotions as Objects of Argumentative Constructions." *Argumentation* 24: 1–17. https://doi.org/10.1007/s10503-008-9120-0

Montague, Michelle. 2017. "A Contemporary View of Brentano's Theory of Emotion." *The Monist* 100 (1): 64–68. https://doi.org/10.1093/monist/onw019.

Moors, Agnes, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. "Appraisal Theories of Emotion: State of the Art and Future Development." *Emotion Review* 5: 119–24.

Papineau, David. 2002. *Thinking About Consciousness*. New York: Oxford University Press.

Papineau, David. 2007. "Phenomenal and Perceptual Concepts." In *Phenomenal Concepts and Phenomenal Knowledge*, edited by Torin Alter and Sven Walter, 307–37. New York: Oxford University Press.

Papineau, David. 2011. "Phenomenal Concepts and the Private Language Argument." *American Philosophical Quarterly* 48 (2): 175–184.

Pedrini, Patrizia, and Jacopo Marchetti. 2026. "Performative Contingency and *Affective* Polarization." *European Journal of Analytic Philosophy* 22 (1): (SI3) 1–29. https://doi.org/10.31820/ejap.22.1.2

Prinz, Jesse. 2004. *Gut Reactions.* Oxford: Oxford University Press.

Putnam, Hilary. 1965. "Brains and Behavior." In *Analytical Philosophy* vol. 2, edited by R. J. Butler, 24–36. Oxford: Blackwell.

Rey, Georges. 1997. *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Cambridge, Mass.: Wiley-Blackwell.

Scarantino, Andrea, and Ronald de Sousa. 2021. "Emotion." *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), edited by Edward N. Zalta, https://plato.stanford.edu/archives/sum2021/entries/emotion/

Shagrir, Oron. 2010. "Marr on Computational-Level Theories." *Philosophy of Science* 77 (4): 477–500. https://doi.org/10.1086/656005

Shoemaker, Sidney. 1975. "Functionalism and Qualia." *Philosophical Studies* 27 (5): 291–315.

Simonsson, Otto, Olivier Bazin, Stephen D. Fisher, and Simon B. Goldberg. 2022. "Effects of an 8-Week Mindfulness Course on Affective Polarization." *Mindfulness* 13: 474–483. https://doi.org/10.1007/s12671-021-01808-0

Tappolet, Christine. 2016. *Emotions, Values and Agency*. Oxford: Oxford University Press.

Tye, Michael. 2006. "Absent Qualia and the Mind-Body Problem." *Philosophical Review* 115 (2): 139–168.

Tye, Michael. 1995. *Ten Problems of Consciousness.* Cambridge, MA: MIT Press.

Wiggins, David. 1987. *A Sensible Subjectivism*. Oxford: Oxford University Press.

Wilcox, Clyde, Lee Sigelman, and Elizabeth Cook. 1989. "Some Like it Hot: Individual Differences in Responses to Group Feeling Thermometers." *Public Opinion Quarterly* 53 (2): 246–257. https://doi.org/10.1086/269505

Wittgenstein, Ludwig. 1953/1965. *Philosophical Investigations.* New York: The Macmillan.