CHAPTER 1

Theory Development and Concepts

Theory development can happen via different paths. Section 1.1 describes one such path: the "demarcation-explanation cycle."[1] This path will turn out to be particularly suitable to describe theory development in the emotion domain. Section 1.2 introduces different types of definitions and ways to evaluate their adequacy. Section 1.3 introduces different types of explanations, and related to this, the notion of levels of analysis. This section also digs deeper into the ingredients of mechanistic explanations such as representations, operations, and operating conditions (related to automaticity). It also briefly pauses to discuss dual-process and dual-system models, different types of rationality, and different usages of the term cognition.

1.1 Demarcation-Explanation Cycle

Scientists develop theories with the aim of explaining, predicting, and/or controlling phenomena (Barnes-Holmes & Hughes, 2013). Although prediction and control are in principle possible without explanation, many agree that explanation is an aim worth pursuing in itself, and that it does have invaluable benefits for prediction and control. "Explanation" is an activity in which an explanandum (i.e., a to-be-explained phenomenon) is linked to an explanans (i.e., an explaining entity or set of entities). To illustrate with a toy example, one type of explanation of the phenomenon of water links it to H_2O. Researchers need to demarcate the explanandum before they can search for an explanans. Rather than being a linear process, however, demarcation and explanation are better understood as alternating activities that can be embedded in a series of cycles.

A first cycle comprises the following four stages (see Figure 1.1(a)). In the first stage, researchers present a provisional demarcation or working definition of the explanandum. If the explanandum is a single entity, the working definition can be a collection of superficial properties.

[1] This path combines elements from Bechtel's (2008) path towards "reconstitution of the phenomenon" with elements from Carnap's (1950) path towards "explication."

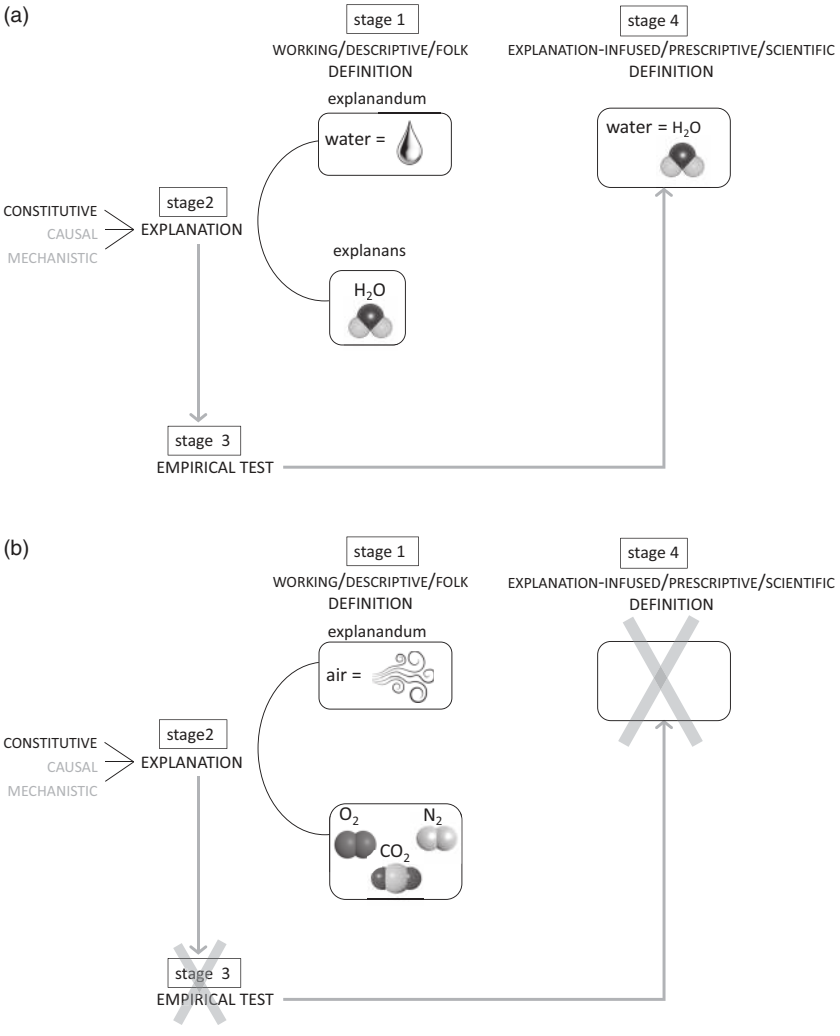

Figure 1.1 Demarcation-explanation cycle: (a) water; (b) air

For instance, water is a transparent, odorless fluid that runs in rivers and falls out of the sky. In the second stage, researchers develop an explanation of some type, in which they link the explanandum to an explanans. In the water example, they discover that the molecular structure of water is H_2O. In the third stage, the explanation is validated by testing it in empirical research. In the water example, researchers take samples of water according to their working definition and they check whether the

molecular structure of these samples is indeed H_2O. If this is sufficiently confirmed, in a fourth stage, the explanans may eventually become part of the definition of the phenomenon, where it replaces the superficial features of the working definition. This definition has now become an explanation-infused definition.[2] Instead of demarcating water as a clear, odorless fluid, it is now equated with H_2O. From now on, water defined as H_2O may figure in new explananda such as the phenomenon that certain substances (e.g., sugar) dissolve in water whereas others (e.g., oil) do not. Note that this new explanandum is no longer a single entity (water), but a regularity between entities (i.e., the mixing of water with other substances and the resulting substance). When new explanations are developed and tested, a scientific theory of water gradually develops.

The entities in science can be understood as sets that have members. This allows us to portray the cycle as follows. Theorists take the working definition of a set as the starting point and develop an explanation in the hope that this will yield a common denominator for the members in this set. If the quest for a common denominator is successful, it forms the basis for the explanation-infused definition of the set.

The demarcation-explanation cycle not only describes (one path towards) theory development in the natural sciences but also in the behavioral and mind sciences, in which all kinds of behaviors and experiences can be targets of explanation. It is especially suitable to describe theory development in the emotion domain, as this domain is still in the stage of figuring out what emotions are. Before we can get our teeth into the emotions, we need to elaborate on the present framework. The following sections discuss types of definitions, types of explanations, and related concepts.

1.2 Types of Definitions and Adequacy

Parallel to what I said about "explanation," "definition" can be thought of as an activity that links a definiendum (i.e., to-be-defined entity) to a definiens (i.e., defining expression) in an identity relation. The demarcation-explanation cycle contains two types of definitions: a *working definition* in Stage 1 and an *explanation-infused definition* in

[2] This corresponds to Bechtel's (2008) "reconstitution of the phenomenon." Several other authors have accepted explanantia at the heart of definitions (e.g., Eilan, 1992; Gordon, 1974; Green, 1992; Reisenzein, 2012; Reisenzein & Junge, 2012; Reisenzein & Schönpflug, 1992; Siemer, 2008). A well-known example is that of "sunburn defined as inflammation of the skin caused by overexposure to the sun" (Gordon, 1978). Note that the credo to avoid conflating explanandum with explanans, although violated in the fourth stage, remains important for the first three stages.

Stage 4. The working definition is often a *descriptive* or *folk* definition, that is, a description of the way in which laypeople understand an entity. The explanation-infused definition is a *prescriptive* or *scientific* definition, that is, a definition in which scientists prescribe how the entity should be understood in scientific discourse (Widen & Russell, 2010).

Another type of distinction pertains to different formats of definitions (J. Lyons, 1977, p. 158). *Intensional* definitions specify the conditions or criteria for a member to belong to a set (i.e., the intension): a single condition that is both necessary and sufficient or a conjunction of necessary conditions that are together sufficient. The conditions are often expressed as properties (Orilia & Paolini Paoletti, 2020). For instance, the set of bachelors has the properties "men" and "unmarried." Note that intensional definitions often do not list all the necessary conditions of a set, but only those that help demarcate the set from specific other sets. The non-mentioned necessary conditions either are implicated in some of the mentioned ones, or they are implicitly assumed. In the bachelor example, the condition "men" implies a bunch of conditions that make the existence of men possible (e.g., that there is a world, and a galaxy) and a bunch of implicit conditions (e.g., that the men are human and that they are adults not babies).

Extensional definitions list the members within a set (i.e., the extension). Intensional and extensional definitions are reciprocal: A set with the intension "all integers between 2 and 7" fixes the extension to {3, 4, 5, 6}. Conversely, a set with the extension {3, 4, 5, 6} leaves room for several intensions, of which a simple one is "integers between 2 and 7" and a more complex one could be "integers that subtract 7, 6, 5, 4, and 3 from 10." A complete extensional definition is only possible for finite sets. For infinite sets, the most one can do is give a sampling definition in which a few prototypical members are listed.

A special type of extensional definitions, which I call divisio definitions, specify the subsets within a set.[3] Divisio definitions not only help to demarcate a set, similar to intensional and extensional definitions, but also to organize the variety within a set. Sets can often be partitioned in more than one way. The set {3, 4, 5, 6} can be split on a low level into subsets that correspond to each of the members ({3},{4},{5},{6}). On a higher level, it can be split into the broad subsets of small ({3, 4}) and large numbers ({5, 6}), but also into the broad subsets of even ({2, 4}) and odd ({3, 5}) numbers. The way in which theorists partition a set thus involves an element of choice.

[3] The term was originally used by Cicero (*Topics*, V. 28; cited in Ierodiakonou, 1993).

The sets, subsets, and members that science is interested in qualify as *types* (i.e., abstract entities) that can be exemplified or instantiated by *tokens* (i.e., concrete entities in space-time; Wetzel, 2018). It could be argued that when members are understood as types, they are in fact subsets of tokens. For this reason, I will continue to talk about "divisio definitions" instead of "extensional definitions."

In principle, both working definitions and scientific definitions can take on an intensional format (i.e., a list of properties) and a divisio format (i.e., a list of subsets). While scientific definitions strive for completeness and precision, working definitions are first approximations. This is why working definitions will often be partial or incomplete.

The scientific definitions in Stage 4 can be evaluated in terms of their adequacy using meta-criteria such as similarity, fruitfulness, and simplicity, to name the most important ones (Carnap, 1950). I first discuss what these criteria entail in the case of intensional definitions before turning to divisio definitions.

In the case of intensional definitions, the similarity meta-criterion entails that the extension of the scientific definition bears sufficient overlap with the extension of the working definition. This means that the scientific definition should tie in with common sense (Green, 1992; Scarantino, 2012b). For instance, the members of the scientific set "water" should show substantial overlap with members of the folk set "water."

The fruitfulness meta-criterion requires that a set allows for scientific extrapolation, that is, the generalization of discoveries about one exemplar to other exemplars in the set (Griffiths, 2004a; Scarantino, 2012b). Scientific extrapolation is only possible when the set is homogeneous in a non-superficial way. Exemplars must share a deep similarity such as a common constitution, a common causal mechanism, or even a common function. If the set is too heterogeneous, not enough generalizations can be made from one exemplar to another. According to this criterion, "diamond" is an adequate set because all its members are constituted by one mineral whereas "jade" is inadequate because its members can be constituted by two different minerals: jadeite and nephrite. Discoveries for jadeite may not generalize to nephrite.

The meta-criterion of simplicity or parsimony, finally, requires that the conditions in a scientific definition be few. Demarcating the set of water using H_2O as the only condition is simple. In fact, the simplicity meta-criterion is hard to separate from the fruitfulness meta-criterion. The ideal is to find a simple common ground among the members of a set, not a complex disjunction of several partially common grounds as this would again hamper extrapolation. This can be captured in the term "fruitfulness-annex-simplicity meta-criterion" but for ease of communication

I will continue to use the term fruitfulness and treat the simplicity meta-criterion as part of it.

Theorists must strike a balance between similarity and fruitfulness even though there are no guidelines for how to establish their relative weights (Swartz, 1997). If the folk set is heterogeneous at the outset, a trade-off between these meta-criteria is inevitable. Maximizing similarity comes at the cost of fruitfulness and maximizing fruitfulness comes at the cost of similarity. Take again the folk set "jade," which is composed of the minerals of jadeite and nephrite. If the scientific definition keeps both minerals on board, this would ensure maximal similarity at the expense of fruitfulness. If the scientific definition keeps only one mineral on board and throws out the other, this would ensure maximal fruitfulness at the expense of similarity. In between these extreme forms of prioritizing similarity or fruitfulness, more subtle forms can be identified.

One moderate form of prioritizing similarity over fruitfulness consists in giving up the quest for a classic intensional definition (with one condition that is both necessary and sufficient or a conjunction of necessary conditions that are jointly sufficient) and turning instead to a cluster-type definition. Simply put, a cluster-type definition is a weak form of intensional definition in which the status of the conditions is relaxed from necessary to typical (Boyd, 1999, 2010; Searle, 1958; Wittgenstein, 1953). For instance, the conditions used to demarcate the set of lemons are typical instead of necessary: oval (some lemons are round), yellow (some lemons are green), and acid (some lemons are bitter). Members belong to the set when they show more or less resemblance with a prototype (Rosch, 1999), understood as an average of all members of the set (Posner & Keele, 1968) or a salient member (Kahneman & Miller, 1986; see Russell, 1991). More formally, cluster-type definitions can be expressed as a disjunction of sets of jointly sufficient properties (Longworth & Scarantino, 2010). The set of lemons has the properties "oval, yellow, and sour" or "oval, yellow, and bitter," or "round, yellow, and sour," and so on. Thus, cluster-type definitions still count as intensional definitions but they are more complex than their classic counterparts and they may hamper smooth extrapolation. Cluster sets are common in science. In addition to lemons, other popular examples are biological species, games, art, and mental disorders. Proponents of this approach argue that the cost for fruitfulness, although in principle increased, remains low in practice. The fact that a strict intensional definition has not been found for lemons does not bother people who need to buy lemons to make lemonade. If it tastes and smells like lemon, it will do.

Moderate forms of prioritizing fruitfulness over similarity, on the other hand, consist in trimming the folk set to a smaller or larger degree.

For instance, when the folk set "fish" turned out to contain not just cold-blooded vertebrates that have gills throughout life (like guppies and sharks) but also a small number of warm-blooded species that breathe through lungs (like dolphins and whales), the latter were trimmed off from the scientific set of fish. The case discussed above in which nephrite is thrown out of the set of jade is more radical in that much more from the initial set is lost. Another solution to handle heterogeneity in this case would be to split the folk set into two equally valid subsets. In this way, more can be rescued from the folk set than just a single subset.

The most radical form of prioritizing fruitfulness over similarity consists in the elimination of the set altogether. If the quest for a common ground turns out to be unsuccessful, scientists may conclude that the set cannot reach a scientific status. Take the example of air (see Figure 1.1 (b)).[4] Just like water, air was once thought to be a fundamental building block of nature. The working definition of air contained superficial features such as that it is a transparent, odorless gas that fills our lungs and the sky. Scientists discovered that all members of the set of air are composed of varying molecules such as oxygen, nitrogen, and carbon dioxide. The lack of a stable common denominator led them to conclude that air is not an adequate scientific set (at least not in chemistry). The question of whether the folk set "emotion" is more like "water," "fish," "jade," or "air" is one that I will be considering later in this book.

Turning to the case of divisio definitions then, the similarity meta-criterion entails that the scientific definition carves up the set in a similar way to the working definition. The fruitfulness meta-criterion stipulates that subsets should be created on the basis of simple criteria that allow for extrapolation between the members of each subset. For instance, a scientific divisio definition with subsets solid, fluid, and gasiform H_2O is similar to the working divisio definition with subsets ice, running water, and steam. The partitioning is fruitful because it is based simply on temperature differences and allows extrapolation within each of the resulting subsets.

Once a set has reached the status of a scientific set, it can be called a scientific or investigative kind (Brigandt, 2003; Griffiths, 2004a). Some scientific kinds are called natural kinds. A natural kind not only requires a common denominator that allows for extrapolation, but also that the common denominator be natural, as is captured in the aphorism that natural kinds carve nature at its joints. Natural kinds are typically contrasted with arbitrary or conventional kinds, in which the members are held together by a common feature that is not natural but resides, at least

[4] I owe this example to Jim Russell.

in part, in the minds of the people making the classification. Examples are the set of weeds and the set of pet animals. The differences between weeds and cultivated plants or between pets and other animals cannot be easily captured in natural terms. Weeds and pets can nevertheless be considered as investigative kinds in certain scientific disciplines such as domestication science (Griffiths, 2004a). The question of whether emotion is a natural kind or a conventional kind has gathered some interest among emotion theorists. It is good to realize, however, that the debate about emotions as natural kinds is complicated by the fact that some scholars have stretched the meaning of natural kinds and use it as synonymous with scientific kinds. Such an extension of meaning is based on the ideas that (a) "natural" is not synonymous with "material" but can also be "mental" and (b) "natural" does not need to equate with a "natural essence" (as per a classic intensional definition) but can also include a "cluster of natural features" (as per a cluster-type intensional definition) (for discussions see Barrett, 2006a; Boyd, 1999; Griffiths, 2004a; Scarantino, 2012b; Scarantino & Griffiths, 2011).

1.3 Types of Explanations and Levels of Analysis

Explanations come in various types. Three types will turn out to be relevant for present purposes: constitutive explanations, causal explanations, and mechanistic explanations (see Figure 1.2). I illustrate these types with the hangover example. A *constitutive explanation* specifies the constituents or components of a phenomenon. For instance, a hangover is comprised of a headache, nausea, and a dry mouth. This constitutive explanation is not yet a definition because the presence of these components is not sufficient to demarcate hangovers from other phenomena. Indeed, a headache, nausea, and a dry mouth may also occur when someone has the flu. To demarcate hangovers from viral infections we probably need a *causal explanation*, in which a hangover is linked to excessive drinking the night before. In such an explanation, a phenomenon is explained by pointing at an antecedent cause. A *mechanistic explanation* specifies the detailed steps of the mechanism that mediates between the cause and the explanandum. Drinking allows alcohol to flow into the bloodstream, part of which is transformed by the liver into acetaldehyde (via a mechanism called alcohol dehydrogenase) and further into acetate (via a mechanism called acetyl dehydrogenase). This causes the contraction of blood vessels in the brain, ending up in a headache, and so on.

The nature of these three types of explanations is best understood if we place them within a levels-of-analysis framework. Levels can be distinguished on the basis of several criteria (e.g., scientific disciplines, strata

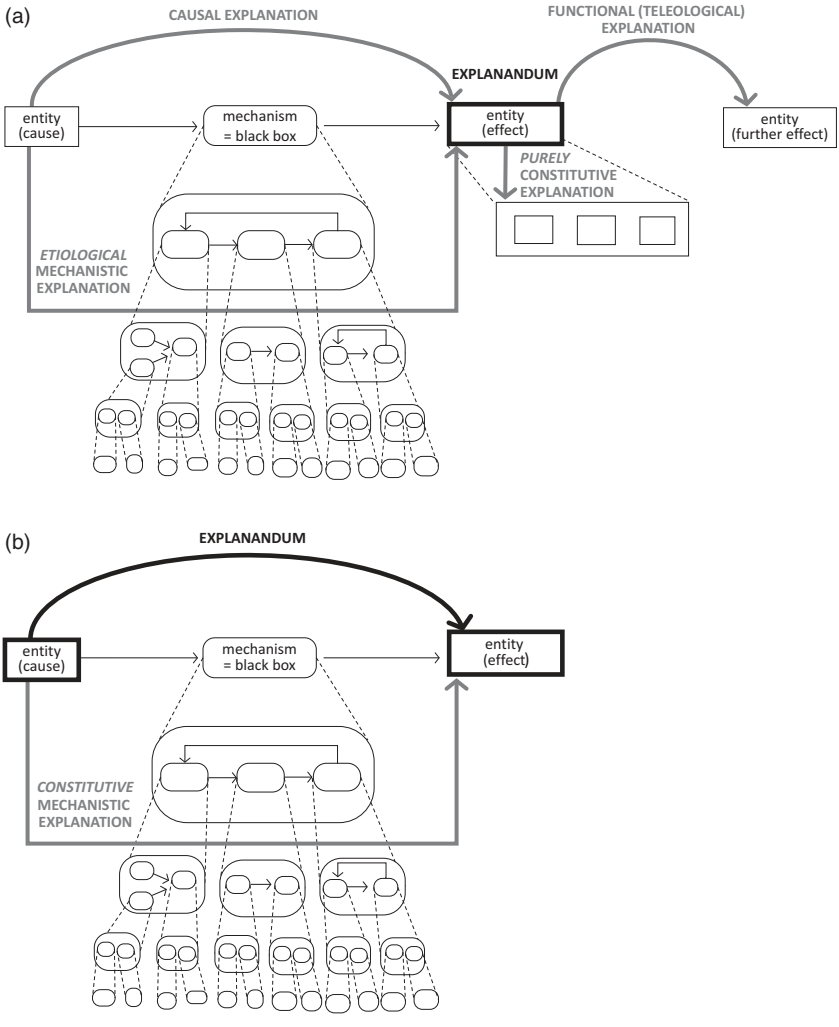

Figure 1.2 Types of explanations: (a) explanandum is an entity; (b) explanandum is a causal relation between entities

across nature, mere aggregates, size, and complexity; see Bechtel, 2008; Craver, 2015). I follow the proposal of mechanistic philosophers of science (e.g., Craver, 2015) to distinguish levels on the basis of mereological (i.e., part–whole) relationships: Level A is lower than level B if the entities at level A are parts of the entities at level B.

In a causal explanation, the explanantia are causal factors situated at the same level of analysis as the explanandum (Craver & Bechtel, 2007,

2013). In constitutive and mechanistic explanations, the explanantia are parts. Constitutive explanations specify the parts of the explanandum, whereas mechanistic explanations specify the parts of the mechanism that mediates between the cause and the explanandum. Thus, mechanistic explanations start from and build on causal explanations in that they specify the mechanisms at a lower level of analysis that mediate between the causal entities (specified in the causal explanation) and the explanandum (Craver, 2013).

In the case in which the explanandum is itself a causal relation between entities (and not a simple entity), explanations that specify the parts of the mechanism mediating between the two entities count as constitutive explanations, strictly speaking. Craver and Tabery (2019; Salmon, 1984) treat the latter type of explanation as a subform of mechanistic explanations, calling them *constitutive mechanistic explanations* (Figure 1.2(b)), next to the subform of *etiological mechanistic explanations* (i.e., which correspond to what I called mechanistic explanations simpliciter so far; Figure 1.2(a)). This leads to an extension of the taxonomy of explanations into four types: purely constitutive ones, causal ones, etiological mechanistic ones, and constitutive mechanistic ones. The first three are suitable when the explanandum is an entity; the fourth is suitable when the explanandum is a causal relation between entities.

Mechanistic explanations not only specify *parts* but also *activities* that spell out the causal relations between parts. The parts in mechanistic explanations are not like marbles in a bag, but hang together in a causal fashion.[5,6] Minimal descriptions of activities only mention that they are causal; more elaborate descriptions specify that the causal relations are also excitatory or inhibitory, for instance, or that they involve certain types of computations.

In addition to specifying parts and activities, mechanistic explanations also specify the way in which different parts and activities are *organized*. An organization can be linear, describing the linear transition from input to output, but it can also be cyclical, in which case the output of a previous cycle forms the input to a new cycle. In sum, mechanistic

[5] Activities figure in etiological as well as constitutive mechanistic explanations. In purely constitutive explanations, on the other hand, information about activities relating to parts is optional. The parts of an atom (neutron, electron, proton), for instance, are working parts, whereas the parts of a marble statue (head, rump, limbs) are not. Purely constitutive explanations that do report activities are nearly indistinguishable from constitutive mechanistic explanations.

[6] Activities have also been characterized as the manifestations of dispositions (also called powers or capacities; Piccinini & Craver, 2011). Some authors have argued that the task of science is not to uncover the activities themselves but rather these dispositions (Manicas & Secord, 1983).