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."

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

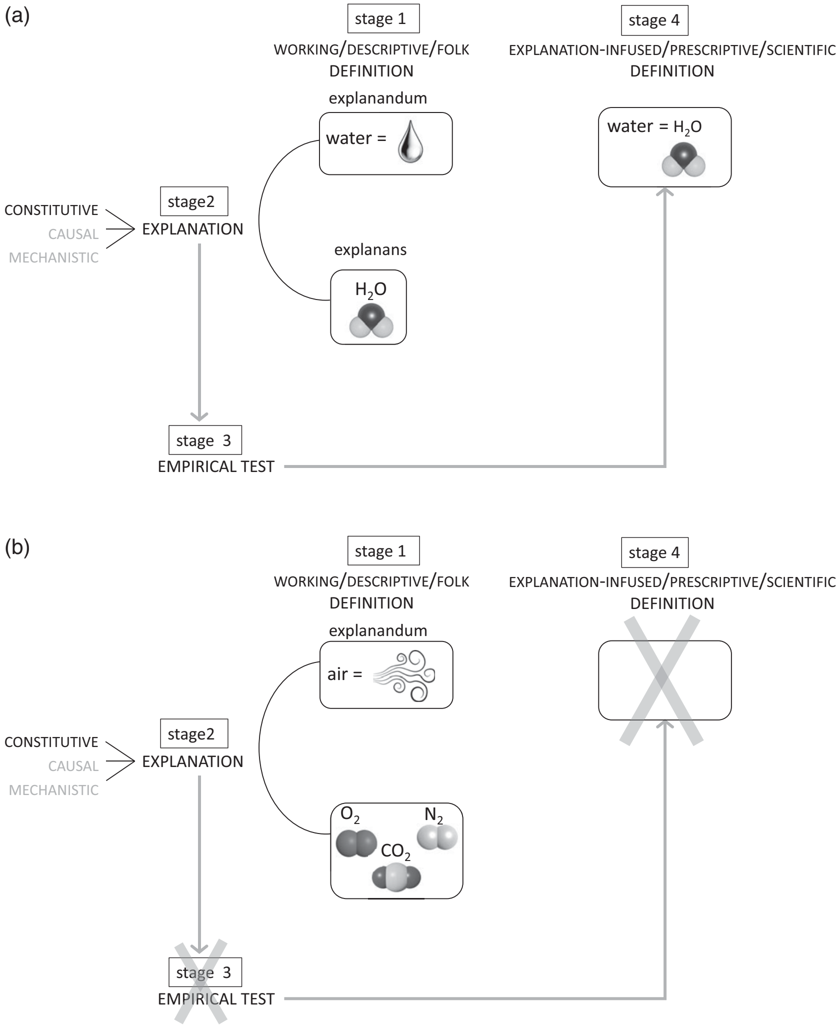

Figure 1.1 Demarcation-explanation cycle: (a) water; (b) air

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

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

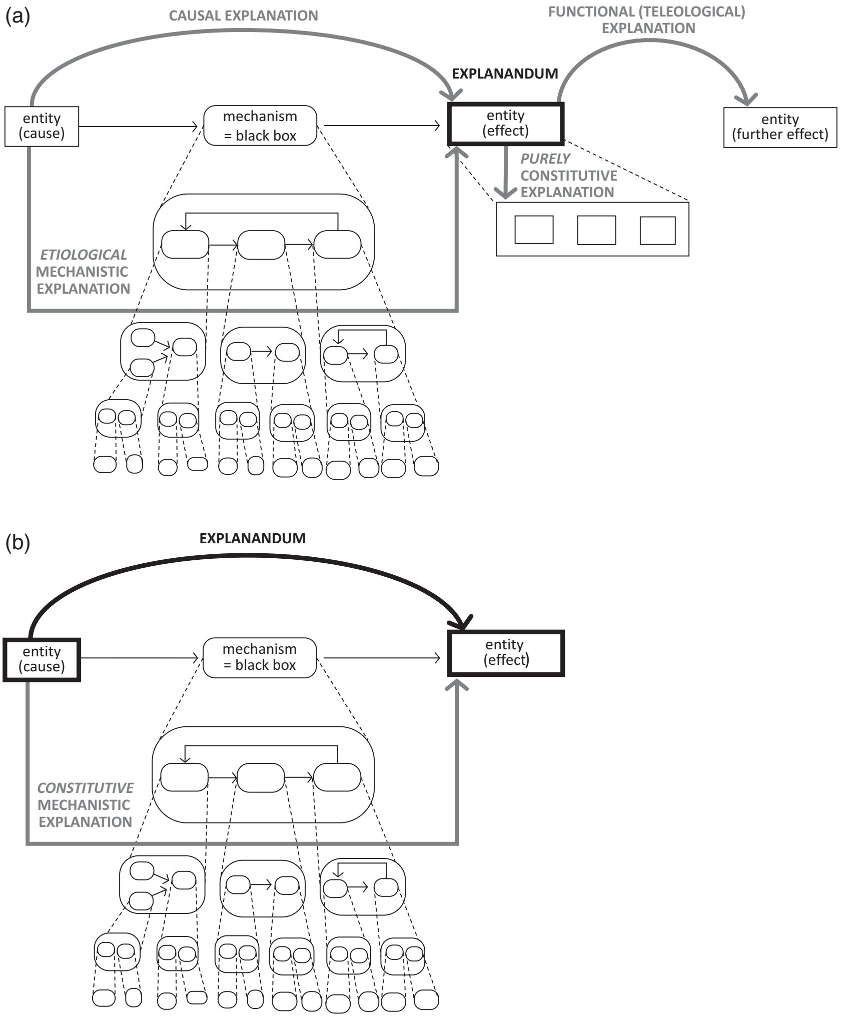

Figure 1.2 Types of explanations: (a) explanandum is an entity; (b) explanandum is a causal relation between entities

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,

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

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).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

explanations not only look downwards (specifying parts; i.e., decomposition) and sideways (specifying causal activities among parts), but also upwards (specifying the organization of parts into wholes; i.e., recomposition) (e.g., Bechtel, 2008; S. Bem & Looren de Jong, 2013).

Parts are presented as the structural aspect of mechanistic explanations whereas activities are presented as the functional aspect of these explanations (Bechtel, 2005). Parts are structural in that they have a location, shape, and orientation, even if they resist a neat description in these terms (Piccinini & Craver, 2011). Activities are functional in that they are specified in terms of what they do or accomplish, that is, the output parts they produce given a certain input part. In the hangover example, the mechanism of alcohol dehydrogenase takes ethanol as its input and produces acetaldehyde as its output, after which the latter substance forms again the input of the mechanism of acetaldehyde dehydrogenase producing acetate as its output. Explanations that specify activities but leave out structural details are dubbed *functional analyses*. Instead of contrasting the latter with mechanistic explanations, Piccinini and Craver (2011; Craver & Kaplan, 2020) have portrayed them as elliptical or incomplete sketches of mechanistic explanations that may form the first steps towards a complete mechanistic explanation.

In mechanistic explanations and functional analyses, the output of the activities is the explanandum. If the consequences of an explanandum are envisaged, however, we speak of a *functional – in the sense of teleological – explanation* (Mundale & Bechtel, 1996). Functional explanations in psychology and biology, for instance, specify the role that the explanandum plays for an organism's long-term goals or survival or for the species or society as a whole. In the hangover example, it might be speculated that hangovers help to avoid alcohol abuse in the future. Hangovers could alternatively be considered as purely epiphenomenal, defying a functional explanation. Functional explanations can be added to the taxonomy as a fifth type of explanation (see Figure 1.2(a)).

The mereological (i.e., part–whole) view of levels of analysis presented thus far is still compatible with a rough division of levels into three broad super-levels inspired by the levels pioneered by Marr (1982) and others (e.g., Bechtel & Shagrir, 2015): an observable super-level, a mental super-level, and a brain super-level. These levels correspond to strata that are relevant for behavioral and mind sciences. At the *observable super-level*, a system produces an observable output (effect) in response to an observable input (cause). The transition from input to output can be called a process, and is mediated by the mechanism as a whole. At this level, a process is described in terms of its observable input, its observable output, and the relation between the two. Typically, the observable input is called the stimulus, and the observable output is a behavioral or

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

physiological response. The mechanism between input and output is treated here as a black box. At the *mental super-level*, this mechanism is decomposed into submechanisms, which can themselves be described in terms of their inputs, outputs, and interrelations. The intermediate inputs and outputs, which are not observable, are called mental *representations* and the relations or activities among them are called mental *operations* (see more below). Each of the submechanisms at the mental super-level may be decomposed further into even finer-grained submechanisms until, at the final stages of decomposition, they correspond to brain processes situated at the *brain super-level*. In other words, the big black box is recursively decomposed into little black boxes all the way down (i.e., heuristic identity relation between levels; Bechtel, 2008). The three super-levels mentioned here are all situated in the individual. In the social sciences, a fourth, *social super-level* can be proposed, where regular patterns of interactions between individuals are specified (Bunge, 2004).

There is debate about how to understand (a) the relations between (all kinds of) mereological levels, which is an ontological question, and (b) the relations between the scientific theories that occupy the four super-levels, which is an epistemological question. Regarding the ontological question, mechanistic philosophers see inter-level relations as constitutive and therefore identity relations. If, in addition, a view of causation is endorsed in which causes should be separate from and precede their effects, it follows that causal relations are strictly intra-level (Bechtel, 2008, p. 153; Craver & Bechtel, 2007, 2013; Crisp & Warfield, 2001; Romero, 2015; but see Baumgartner & Gebharter, 2016; Krickel, 2017; Leuridan, 2012; Ylikoski, 2013). In line with this view, apparent cases of top-down and bottom-up causation can be recast in terms of mechanistic mediation, that is, hybrids of constitutive and causal relations (Craver & Bechtel, 2007, 2013).

Regarding the epistemological question, approaches to inter-theory relations range between (a) classic (i.e., smooth) reductionism in which higher-level theories (e.g., mental theories) are explained away by lower-level theories (e.g., neuroscientific theories; Oppenheim & Putnam, 1958); (b) new-wave (i.e., "bumpy" and "patchy") reductionism in which higher-level and lower-level theories constrain and inspire each other (Churchland & Churchland, 1992; see Mundale & Bechtel, 1996); (c) the "mosaic unity of sciences" view, in which each science contributes in a non-reductive but still interdependent way (Craver, 2007); and (d) explanatory pluralism that lets many flowers bloom and grants each level full explanatory autonomy (see S. Bem & Looren de Jong, 2013; McCauley, 1996; McCauley & Bechtel, 2001). Mechanistic philosophers profess non-reductionist relations among levels based on the argument that mechanistic explanations span at least two levels (e.g., Bechtel, 2008; McCauley & Bechtel, 2001). Critics are unassuaged by this argument, maintaining that the identity relation

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

between levels inevitably invites some form of reductionism (e.g., Fazekas & Kertész, 2011; Glauer, 2012). Without digging further into the details of this debate, I believe it is useful to separate ontological issues from epistemological ones. According to mechanistic philosophers, who understand the relations between levels in a mereological sense, different levels of analyses do not house different realities, but different ways to parse and look at the same reality. Thus, despite the fact that they assume identity relations between the entities of different levels (i.e., ontological issue), the theories situated at each level can still be granted explanatory individuality, whether in a mosaic with, or independent of, theories on other levels (i.e., epistemological issue). All this is to say that the mechanistic approach adopted in this book does not imply epistemological reductionism.

To take stock, causal explanations are intra-level, with the explanantia situated on the same level as the explanandum. Constitutive and mechanistic explanations cross different levels of analysis. If the individual is taken as the unit of analysis, mechanistic explanations can reside at the mental super-level (i.e., mental mechanistic explanations) or the brain super-level (i.e., neural mechanistic explanations). Let us now consider the ingredients of these two types of explanations in more detail.

1.3.1 Mental Mechanistic Explanations

So far, we learned that mechanisms are made up of parts and activities, and that in the case of mental mechanisms, the parts are representations and the activities are operations (Bechtel, 2008). Mechanisms, moreover, vary in the conditions they require to operate. In the following sections, I will clarify my usage of each of these notions – representations (Section 1.3.1.1), operations (Section 1.3.1.2), and operating conditions (Section 1.3.1.3) – and propose ways in which to organize the variety in each (see Figure 1.3). As it turns out, researchers tend to dichotomize this variety. This has tricked them into binary thinking and the formation of dual-process and dual-system models (Section 1.3.3).

1.3.1.1 Representations

Representations have been invoked to cater for the feature of most mental processes that they are directed at something beyond themselves, a feature that philosophers call Intentionality[7] or aboutness (Brentano,

[7] I capitalize the term to indicate the difference with intentional in the ordinary sense, following Searle (1983). The minimal meaning of the term intentional is "directed" (Jacob, 2019). In philosophical usage, *Intentional* refers to the property of a mental state by which it is directed at something beyond itself (Brentano, 1874). In ordinary usage, *intentional* refers to the fact that an agent is directed to (i.e., willing to engage in) an (overt or covert) act (Moors & De Houwer, 2006a).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

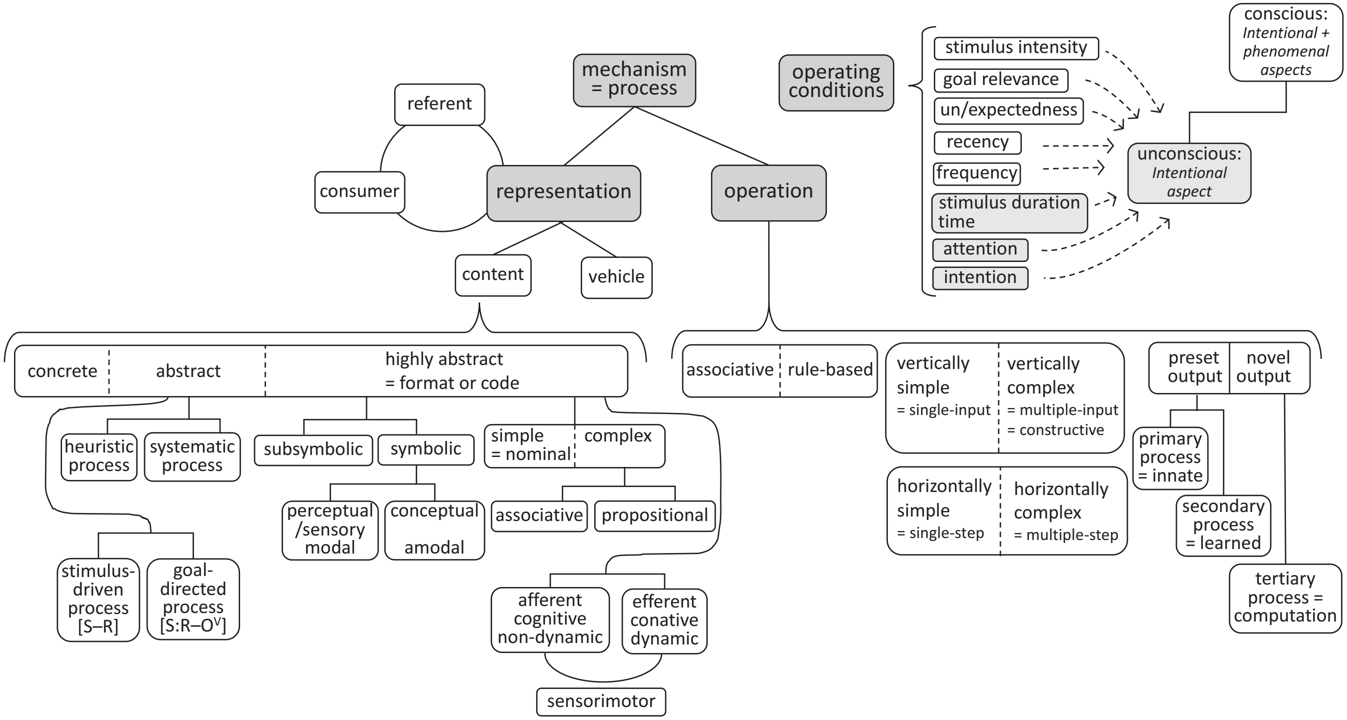

Figure 1.3 Ingredients of mental mechanistic explanations
Note: Full/dotted partitions indicate categorical/gradual distinctions.

1874). One way in which a system can be directed towards something is by forming a representation of it. The notion of representation can be unpacked as being part of a three-place relation, in which it is linked to a referent and a consumer. The referent is an object in the organism's external or internal world. The representation makes the referent available to the consumer or interpretant (Millikan, 1993; Peirce, 1940).

The representation itself is composed of a vehicle and a content. The vehicle is typically identified with the physical brain activity underlying activation of the representation. The content is the referent as represented. This content can be organized along various dimensions. One important dimension is the degree of abstraction. Concrete content can be virtually anything. At the high end of the spectrum, content is so abstract that it slides into what some authors have called the format of representations or the representational code. I will list a number of popular contrasts that have been proposed regarding format: symbolic vs. subsymbolic, conceptual vs. perceptual, simple vs. complex, associative vs. propositional, and afferent vs. efferent. After that, I discuss two dichotomies that concern specific types of representational content: heuristic vs. systematic and stimulus-driven vs. goal-directed.

Let us start with the contrast between symbolic and subsymbolic representations. A symbolic representation is one in which the content is a meaningful entity such as an object. In subsymbolic representations, the meaningful entity is distributed across representations that each refer to a separate feature of the entity. Symbolic representations split further into conceptual and perceptual representations. Conceptual representations are verbal-like or word-like. Perceptual or sensory representations are image-like or picture-like, but they can in principle also be sound-like, smell-like, taste-like, or touch-like.[8] The analogies with words and pictures are not literal – as if there are words and pictures in the head – but rather structural. Perceptual representations have a higher level of detail than conceptual ones and are more vivid, but at the same time, they are schematic in the sense that they allow a form of indeterminacy that actual pictures do not (Barsalou, 1999). For example, the perceptual representation of a tiger has stripes but the number of stripes may remain

[8] The distinction between sensation and perception is not a sharp one. Sensation is used more to refer to the detection of low-level stimulus features by sensory receptors whereas perception is used more to refer to the identification of stimuli based on a combination of sensory stimulus features. Note that the number of senses need not be limited to the "Aristotelian five" listed here (see Macpherson, 2011). Sensation can be organized into (a) external sensation (or exteroception), including vision, hearing, smell, taste, touch, and temperature, and (b) internal sensation, including interoception (internal body state) and proprioception (position of body parts in space, and kinesthesia or movement of body parts).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

indeterminate. Another contrast linked to the conceptual-perceptual contrast is that between amodal and modal representations. Conceptual representations are seen as amodal whereas perceptual representations come in a specific modality, linked to a specific sensory channel (e.g., visual, auditory, olfactory, gustatory, and tactile) (e.g., Barsalou, 2008; Barsalou et al., 2003; Garcia-Marques & Ferreira, 2011).

Another contrast has to do with the complexity of the content of representations: Simple or nominal representations hold a single entity (e.g., a cat) whereas complex representations house multiple entities (e.g., a cat and a mat) and relations among these entities. Relations can be of two kinds. Unqualified relations are called "associations" (e.g., the cat and the mat are related but it is not specified how). Qualified relations are called "propositions" (e.g., spatial relations: the cat is on the mat; temporal relations: night follows day; causal relations: allergens cause allergic reactions) (De Houwer, 2014; Moors, 2014c). These are states of affairs that can be expressed in a that-clause (e.g., that the cat is on the mat). This is why propositional representations are often characterized as sentence-like or sentential. Here again, the analogy with sentences is more structural than literal. Propositions are composed of elements that can be recombined (i.e., compositionality and productivity; Fodor, 1981) but they need not be sentences in the head made up of words.

A further contrast is that between afferent (sensory-perceptual/cognitive) and efferent (motor/conative) representations. This contrast can be expressed in terms of a different direction of fit (Searle, 1983): An afferent representation (e.g., perception or belief) has a world-to-mind direction of fit. This means that it is fitting if its content fits with the world, that is, if it is accurate. An efferent representation (e.g., goal) has a mind-to-world direction of fit. This means that it is fitting if the world fits with its content, that is, if it is satisfied. The contrast between afferent and efferent representations also squares neatly with the contrast made in motivation psychology between pure knowledge representations, which are seen as non-dynamic representations, and goals, which are seen as dynamic representations. A dynamic representation typically leads to behavior and its activation accumulates over time (until the goal is fulfilled or overridden by stronger goals), even in the face of obstacles (Bargh & Barndollar, 1996; Bargh et al., 2010). A non-dynamic representation does not lead to behavior and its activation diminishes over time. To illustrate, activation of the goal to have an apple (i.e., the representation of the apple with a value attached to it) leads to behavior to get the apple and does not diminish but rather increases until the apple is obtained or until a more important goal intervenes. The mere thought of an apple (activated by the instruction to think of an apple or a priming procedure in which a picture of an apple is shown), on the other hand, creates a spike in the activation of the representation of an apple, that is, an initial increase that gradually diminishes over time. Some authors have also voted for the existence of sensorimotor representations: combined afferent and efferent representations, also called "embodied" (Barsalou, 2008) or "pushmi-pullyu" representations (Millikan, 1995). These are representations of stimuli that not only contain perceptual features but also information about what can be done with these stimuli, so-called "affordances" (i.e., a term borrowed from Gibson, 1979, but put in a representational jacket here, see Scarantino, 2003).[9]

The distinction between afferent and efferent representations seems to dissolve when the content of the representation is a behavior. This idea stems from James's (1890b, p. 522) ideomotor hypothesis, which states that when an action is carried out, the action command gets bound up with its sensory effects, resulting in bidirectional action–effect links. Once these links are in place, execution of the action conjures up its sensory effects ([R→E]), and perceiving or thinking about the sensory effects is sufficient to put the action in motion ([E→R], i.e., ideo-motor). Thus, for a behavior to occur, the mere thought of the behavior in terms of its immediate outcome is sufficient for it to become executed. No extra fiat is required. Instead, an extra goal to suppress the behavior is required if the person does *not* want to execute the behavior. This may explain why some people avoid standing on the edge of a cliff. If James (1890b) is right, the mere thought of jumping should cause one to jump unless it is suppressed by the goal not to jump. Thus, if the person thinks of jumping but the suppression is temporarily lowered or lifted (e.g., because of inattentiveness), jumping may become a real risk. A similar idea is voiced in W. Prinz's (1990, 1997) common-coding hypothesis, which states that a common representation or code is used to perceive a behavior and to prepare for it. This hypothesis is supported by research showing that the same brain activity (in mirror neurons) occurs when people are instructed to watch someone else carry out a movement as when they are asked to prepare making this movement themselves (e.g., Fadiga et al., 2000; see Hommel et al., 2001).

Let me close with two popular dichotomies that are based on particular types of representational content. A first content dichotomy – central in theories on reasoning and persuasion – is that between systematic and heuristic information processing. An example of systematic information

[9] Gibson (1979) held a non-representationalist view, which places affordances in external objects and not in the minds of agents. Scarantino (2003) and others have taken a more liberal interpretation in which they allow affordances to figure in the content of representations.

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

are persuasive arguments of a speaker; an example of heuristic information is the attractiveness of the speaker (Chaiken et al., 1989).

A second content dichotomy – central in behavior theories – is that between stimulus-driven and goal-directed processes (Heyes & Dickinson, 1990), two candidate mechanisms of behavior causation. In a stimulus-driven process, behavior is caused by the activation of a representation whose content is the association between the stimulus and a response or behavior ([S–R]). In a goal-directed process, behavior is caused by a representation that contains information about the outcome of one or more response options given a certain stimulus ([S:R–O^v]), more precisely, information about the value of these outcomes and about their expectancy (i.e., the probability that these outcomes will occur). As this dichotomy will be a central principle for organizing emotion theories in this book, it will be discussed in more detail in Box 2.1.

Let us now turn to the consumer of the representation. The consumer is another part of the system, typically another mechanism, that takes this representation as its input. In philosophy, however, a homuncular interpretation of the consumer proves hard to shake off. There, the consumer is said to have an Attitude[10] towards the content of a (usually propositional) representation. Attitudes vary in mode, with the most common ones being an Attitude of belief and an Attitude of desire. Believing that the train is late is different from desiring that the train is late. To believe it is to judge it as true, to desire it is to want it to come true. A "belief" is the combination of a (propositional) representation and an Attitude of belief. A "desire" is the combination of a (propositional) representation and an Attitude of desire. A belief has a mind-to-world direction of fit (it fits if its content fits with the world); a desire has a world-to-mind direction of fit (it fits if the world fits with its content). It may be noted that beliefs and desires in philosophy show overlap with afferent and efferent representations in psychology. The difference is that in beliefs and desires, the direction of fit is located in the Attitude towards a representation whereas in afferent and efferent representations, it resides in the format of the representation.

So far, a realist picture of mental representations has been drawn, with their vehicle corresponding to actual brain activity, although the precise mapping between representations and brain activity has been left unspecified. However, representations can also be understood in purely

[10] I capitalize Attitude, used here in the philosophical sense, to mark the distinction with attitude in psychology, where it refers to the liking or preference of a person for a certain object, often understood on the mental super-level as the association between an object and a valence label (e.g., apple – positive; Greenwald et al., 2002; but see De Houwer, Gawronski, et al., 2013).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

functional terms as entities that help explain variable input–output relations, without making any commitment regarding the ontological status of these representations. If a response to the same stimulus varies across occasions, it makes sense to posit an intervening entity such as a representation (Bermúdez, 1995; Fodor, 1981; Moors, 2014c). Representations may be nothing but metaphors, as Skinner (1945, 1977) argued, but in this capacity, they do still play an important heuristic role (De Houwer, Barnes-Holmes, & Moors, 2013).

1.3.1.2 Operations

Operations are the activities carried out on representations. Examples of types of operations cited in the literature are associative and rule-based ones (S. A. Sloman, 1996). An associative operation is the activation of an association between at least two representations. A rule-based operation is the application of an abstract rule to representations. To make the distinction intuitive, imagine a person ordering two lemonades at the counter and trying to figure out how much to pay. To solve the problem, she can engage in an associative operation in which she remembers the price that she paid last time. She can also engage in a rule-based operation in which she applies the rule "multiply the price of one lemonade by the number of lemonades ordered." Although intuitive at first sight, the distinction between associative and rule-based operations has turned out to be fairly elusive (Hahn & Chater, 1998; Moors, 2014c). Perhaps it can best be characterized in terms of degrees of abstraction of the representations, rather than as different types of operations, with representations in associative "operations" situated at the more concrete end of the spectrum and those in rule-based "operations" at the more abstract end (for an extensive justification, see Moors, 2014c).

Operations can also be classified with regard to their complexity. It is worth distinguishing between a vertical and horizontal type of complexity. Vertical complexity refers to the number of inputs that an operation integrates simultaneously: Single-input operations take a single input to produce their output; multiple-input operations, also called constructive operations, take two or more inputs to produce their output (Moors, 2010b). The number of inputs can be regarded as independent of the types of operations involved. Indeed, both associative and rule-based operations can be single-input or multiple-input (Moors, 2014c).

Horizontal complexity refers to the number of sequential steps that must be carried out to arrive at an output. Some operations are single-step, others are multiple-step (Logan, 1988). Again, the types of operations involved in the steps is open. They can be rule-based or associative. In the psychology of language, a reduction of steps is called chunking or entrenchment (Hartsuiker & Moors, 2017). In computer science and

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

artificial intelligence, a reduction of steps or of inputs is known as compilation and the reverse movement as decompilation (A. Sloman & Croucher, 1981).

Another distinction has to do with whether the output of the operation was preset or is novel (e.g., Panksepp, 2012). Primary processes are innate. They rely on outputs that were preset during phylogenesis (i.e., the evolution of the species). Secondary processes are learned. They rely on outputs that were preset during ontogenesis (i.e., evolution of the individual). Tertiary processes are computations. They can use raw stimulus input or preset representations but their output is freshly produced during microgenesis (i.e., evolution of some process in real time). Although the preset or novel output of operations is in principle independent of types of operations and complexity, a compelling intuition is that primary and secondary processes suffice with single-input associative operations, whereas computation involves a multiple-input operation, whether it is rule-based or associative. To bake a cake (i.e., a novel entity), you typically have to combine several ingredients (i.e., multiple inputs).

1.3.1.3 Operating Conditions and Automaticity Features

In addition to the representations and operations that fix the nature of a mental process, it is worth pointing at factors that count as conditions under which a process can operate or that influence the strength of a process. Examples are the duration, intensity, goal relevance, and un/expectedness of the input, the amount of attention directed at the input, the recency and frequency of the input, and the goal to engage in the processing of the input. Stimuli that are longer-lasting, more intense, more goal-relevant, more *or* less expected, more attended to, recently and frequently processed, and intended to be processed by the person are more likely to be processed or are processed better. Different taxonomies have been proposed to organize these factors. The social psychological literature, for example, groups factors into the categories of opportunity (e.g., stimulus duration and intensity), capacity (e.g., attentional resources), and motivation (e.g., the goal to engage in the process). I recently proposed a more detailed taxonomy (Moors, 2016), based on the distinctions between current vs. prior factors and between observable vs. mental factors, combining them in a four-field table with (a) current observable factors (e.g., stimulus duration and intensity), (b) prior observable factors (e.g., recency and frequency), (c) current mental factors (i.e., quality of the current representation), and (d) prior mental factors (e.g., quality of a prior representation in working memory).

Only a small subset of the above-listed factors, namely duration/time, attention, and intention, have been linked to the dichotomy between

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

automatic and non-automatic processes. Automatic processes have been characterized as fast, efficient, and unintentional, but also as difficult to counteract and unconscious; non-automatic processes have been bestowed with the opposites of these features (Bargh, 1994; Moors, 2016; Moors & De Houwer, 2006a). While the first three features can be recast in terms of operating conditions (as above), the last two features escape this framing. Saying that a process is automatic in the sense of fast, efficient, or unintentional comes down to saying that this process requires little time, little attentional capacity, or no goal to engage in the process, respectively. However, a process that is difficult to counteract does not operate due to, but despite, the presence of the goal to counteract the process. And whether a process is conscious or unconscious is more aptly considered as a consequence of the presence of other conditions (e.g., time) rather than as a condition for the operation of the process itself (although it can be a condition for the operation of subsequent processes; Moors, 2016).

Starting from the premises that all processes require an input of sufficient quality to get launched, and that many mental processes take a representation as their input, I have proposed that the quality or activation level of this input representation is the proximal factor that determines the occurrence and strength of these mental processes (Moors, 2016). A first threshold of activation must be exceeded for the representation to serve as the input to an unconscious process; a second threshold must be reached for the representation to become conscious and serve as the input to a conscious process. In addition, I proposed that the various factors listed above (e.g., stimulus duration, stimulus intensity, attention, frequency, recency) feed into this proximal factor, and that they do so in an additive way. If stimulus duration is reduced, for instance, an increase in stimulus intensity or attention may compensate so that the total activation level is sufficient to launch the process and/or to make it conscious.

A few more words about consciousness are in order. Being conscious of a process requires being conscious of the input and output of a process as well as of their interrelation. These inputs and outputs must be representations and they must be situated on a high level of analysis. It is unlikely that people can be conscious of raw stimulus input and of low-level mental processes. A person can become conscious of the fact that watching advertisements influences her urge to go shopping, for instance, but not of the many detailed representations that go into this process on a lower level of analysis. On a final note, philosophers sometimes use the terms personal-level vs. subpersonal-level processes to refer to conscious vs. unconscious processes.

As mentioned above (see Section 1.3.1.1), the representations involved in mental processes allow for Intentionality, the characteristic of being directed at or about something (Brentano, 1874; Searle, 1983). Both

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

conscious and unconscious mental representations have Intentionality. In conscious representations, moreover, this Intentional aspect is combined with a phenomenal aspect. The phenomenal aspect consists of the qualia or the non-representational content of experience, the aspect of experience that remains after the Intentional aspect is stripped away (Block, 1995; Searle, 1983). For instance, it is what seeing red or having an itching toe feels like, when all there is to know about redness and itches is removed. Sensations like redness and itches have a felt aspect (i.e., there is "something it is like" to have them; Nagel, 1994) that defies any verbal description. Qualia may not be confined to sensations, but also apply to conscious verbal and even abstract thoughts. Some theorists believe such thoughts can only give rise to qualia in an indirect way, however, via the mental images they conjure up (J. J. Prinz, 2010). While unconscious representations are supposed to have Intentionality without phenomenality, some authors have ventured the existence of conscious mental entities that have phenomenality but lack Intentionality, such as objectless sensations and positive or negative feelings (e.g., Reisenzein, 2012; but see Brentano, 1874).[11] Finally, the first-order consciousness discussed so far must be distinguished from second-order consciousness or the state of being conscious of one's first-order conscious entities (Block, 1995; Wegner & Bargh, 1998). Non-human animals are typically assumed to be capable of the former but not the latter type (see Heyes, 2008).

1.3.2 Neural Mechanistic Explanations

The parts and activities in neural explanations correspond to neural representations and neural operations. Neural representations have sometimes been identified with populations or patterns of firing neurons (Bechtel, 2001). In the domain of perception, for instance, the neurons in cortical area V4 code for color whereas those in cortical areas MT and V5 code for motion. In other domains, additional criteria are used to demarcate working parts such as patterns of connectivity (e.g., Bechtel, 2005,

[11] There is debate about how to cash out the distinction between Intentional and phenomenal aspects of consciousness and about how to relate both aspects to one another. So far, I have equated the Intentional aspect with the content of representations, and the phenomenal aspect with non-representational content. The phenomenal aspect can be absent but if it is present, it is supervenient or dependent on the Intentional aspect, like the icing on a cake (Byrne, 2001). Another proposal is that the Intentional aspect depends on the phenomenal aspect in the sense that the phenomenal aspect is what gives the Intentional aspect its meaning (Natsoulas, 1981), as can be illustrated by the argument that the abstract thought of a circle remains meaningless until it is injected with phenomenal experience. Still another proposal is that both aspects are mutually dependent and interwoven (Eilan, 1998).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

p. 316; Mundale, 1998). Identifying operations in the brain turns out to be more challenging, however (Bechtel, 2005). In connectionist models (e.g., Rumelhart et al., 1986), the only operation allowed is the firing or activation of the neurons. Yet some scholars have argued that this level of characterizing operations may be too low (Bechtel, 2005).

1.3.3 Dual-Process/System and Multiple-Process/System Models

The various dichotomies listed so far have all been used as grounds for splitting the realm of processes into two exhaustive subsets. Dual-process models have been based, for instance, on formats of representations (e.g., modal vs. amodal; associative vs. propositional), contents of representations (e.g., heuristic vs. systematic; Chaiken et al., 1989; stimulus-driven vs. goal-directed; Balleine & Dickinson, 1998), types of operations (e.g., associative vs. rule-based; S. A. Sloman, 1996), operating conditions (e.g., automatic vs. non-automatic; see Moors & De Houwer, 2006a), and brain locations (e.g., subcortical vs. prefrontal cortical).

Many dual-*process* models have mapped two or more dichotomies onto each other, which has turned them into dual-*system* models (e.g., J. S. B. T. Evans, 2003; Hofmann et al., 2009; Kahneman & Frederick, 2005; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004). System 1 houses processes that represent heuristic information (in reasoning models) or information on stimuli and responses (i.e., stimulus-driven process in behavioral models), represented in the form of associations (e.g., handsome – reliable; snake – flee), activated via associative operations, in an automatic way, and implemented by subcortical brain areas. System 2 houses processes that represent systematic information (in reasoning models) or information about outcomes of responses (i.e., goal-directed process in behavioral models), in propositional format, handled by rule-based operations, in a non-automatic way, and implemented in prefrontal cortical brain areas.

Dual-system models have met with serious criticism (Keren & Schul, 2009; Melnikoff & Bargh, 2018a, 2018b; Moors, 2014c; Moors & De Houwer, 2006b). In brief, it has been argued that assumptions of alignment should not be made a priori but should be investigated empirically. This empirical research is complicated by the fact that some dichotomies (e.g., associative vs. rule-based operations) resist a clear definition, thereby making it nearly impossible to diagnose them (Hahn & Chater, 1998; see Moors, 2014c, for a review). Other dichotomies, such as that between automatic and non-automatic modes of processing (and perhaps also that between associative vs. rule-based operations), are gradual in nature instead of binary, thereby allowing only for relative conclusions. Keeping these caveats in mind, empirical research does provide evidence

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

for *non*-alignment between several dichotomies. For example, several studies have shown that goal-directed processes can be relatively automatic (Aarts & Dijksterhuis, 2000). Other work has shown that people can learn to categorize stimuli based on two sources of information (e.g., angle and density) that follow a complex rule even though participants are unable to articulate the rule (Hélie, Roeder, & Ash, 2010; Hélie, Waldschmidt, & Ash, 2010; Kovacs et al., 2021). Still other research has shown that rule-based reasoning can be fast (Newman et al., 2017). These findings have led some scholars to propose single-system models in which dissociations are understood in terms of complexity rather than in terms of qualitatively different systems (e.g., Kruglanski & Gigerenzer, 2011; Osman, 2004).

Some scholars have proposed triple-system models (e.g., Leventhal & Scherer, 1987; Panksepp, 2012; Panksepp & Watt, 2011).[12] Panksepp (2012), for instance, distinguished between three layers of organization in the brain. The first layer located in subcortical areas houses innate (i.e., primary) processes, mostly stimulus-driven ones, which are supposed to be triggered automatically. The second layer houses basic learning (i.e., secondary) processes such as those involved in classical and operant conditioning. Learning does not start from a blank slate, but builds further on innate processes. Learning processes may range in complexity and there is debate about the extent to which they involve computation. Once installed, however, deployment of innate and learned knowledge is assumed to happen via single-input associative processes that are automatic. The third layer is located in neocortical regions and houses computations (i.e., tertiary processes), which are assumed to rely on complex rule-based operations that are non-automatic. Here again, the alignments are assumed a priori and may not survive empirical testing.

1.3.4 Rationality

The two systems in dual-system models have also been aligned with another dichotomy: that between rationality and irrationality. This dichotomy does not concern properties of the processes or their operating conditions, but points at an outsider evaluation relative to certain standards (Davidson, 1985b). The presence of such standards reveals a normativist approach (as distinct from a descriptivist approach, e.g., Elqayam & Evans, 2011). Before turning to the alignment, let me explain a few basic distinctions.

Rationality comes in different shapes. Theoretical or epistemic rationality refers to the accuracy of an entity to represent the external

[12] Some scholars have also proposed multi-process or multi-system models with four types of processes (e.g., Conrey et al., 2005; Sherman, 2006).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

world. Practical rationality or adaptiveness refers to the degree to which an entity satisfies goals and leads to well-being. Afferent representations, such as perceptions and beliefs, have a mind-to-world direction of fit. They are evaluated in terms of how well they fit with the external world, that is, their accuracy or theoretical rationality. Efferent representations, such as desires and goals, have a world-to-mind direction of fit. They are evaluated in terms of how well the world fits with them, that is, how well they are satisfied. The satisfaction of goals must be done by subgoals and ultimately by behavior. Thus, it is subgoals and behaviors that are typically evaluated in terms of how well they satisfy goals or well-being, that is, how practically rational they are.

In addition to the afferent and efferent representations and behavior that are produced as the outputs of processes, rationality can also be judged for the processes themselves (Elster, 2010). While the rationality of the outputs of processes is judged based on their *actual* fit (accuracy in the case of theoretical rationality; satisfaction in the case of practical rationality), the rationality of processes is judged based on their *potential* to produce a fitting output, irrespective of the output itself. Rationality in the output-sense and the process-sense may dissociate. Indeed, a reasoning process that uses the right logic may still produce a false belief whereas one that uses the wrong logic may still produce a correct belief. Likewise, a process of behavior causation designed to satisfy goals (i.e., a goal-directed process) may fail to do so whereas one that is not so designed (e.g., a stimulus-driven process) may still accidentally satisfy goals.

One of the reasons for these dissociations is that the rationality of human thought and behavior is never Olympian, but always bounded by the information that is available to the individual, along with the opportunity, capacity, and motivation of the individual to process this information (Bechtel & Richardson, 2010; Simon, 1983). Individuals facing a decision lack information about all possible outcomes, and especially all possible long-term outcomes, of their behavior. Rationality can also be judged from a global or enlightened point of view, which depends on the best evidence that is actually or conceivably available (and hence still not Olympian) (Salmela, 2008). A decision process is rational according to an enlightened standard if the individual makes use of the best available information, but irrational if the individual does not use this information, for instance, due to a lack of opportunity, capacity, and/or motivation.[13]

A few additional distinctions are worth making regarding practical rationality. For one thing, the goals and well-being can be those of one

[13] It could be argued that opportunity does not belong in this list because a lack of opportunity implies a lack of access to information.

individual or those of a community of people. The former type of practical rationality can be called prudential rationality; the latter type can be called moral rationality.[14] Orthogonal to the beneficiary of the rational behavior (individual, community), it can also be specified who does the evaluating: the individual or the community. Taken together, the behavior of an individual can be evaluated by an individual or a community to be good or bad for the individual or the community. Finally, well-being can be understood in terms of the satisfaction of short-term goals (local rationality), long-term goals (ultimate rationality), or an optimal mix of the two (Lemaire, 2021). These are just a few distinctions that highlight the versatility and complexity of the notion of rationality.

Turning back to the alignment of dichotomies, System 1 is typically mapped onto irrationality and System 2 onto rationality. I would even argue that the apparent deviations from rationality observed in daily life are what motivated the creation of dual-system models in the first place. It is when people talk or act dumb that explanations arise in terms of a dumb system taking over. People tend to think that dissociations on the observable super-level of analysis match with dissociations on the mental and neural super-levels and they disregard the possibility that the same mechanism may produce different outputs given different inputs and different other conditions. Yet as several scholars have argued, even double dissociations (e.g., Lieberman et al., 2004) are ultimately unsuitable to settle debates between dual-system models and alternative, single-system models (Chater, 2003; Keren & Schul, 2009).

But how can we make the alignment between ir/rationality and System 1/2 more intelligible? A first source of this alignment is the obvious connection between heuristic content and irrationality and systematic content and rationality. Buying a car because it has the best cost-benefit ratio is more rational than buying it because the salesperson is handsome. A second source is the widely (but often implicitly) assumed trade-off between automaticity and rationality (reminiscent of the better-known trade-off between speed and accuracy), which is tied to the complexity and hence flexibility of processes (e.g., Strack & Deutsch, 2004; see Moors et al., 2017). In short, complex processes are supposed to be non-automatic, or less automatic, than their simple counterparts because integrating more inputs (i.e., vertical complexity) or going through more steps (i.e., horizontal complexity) takes more time and effort. At the same time, complex processes are more flexible than their simple counterparts

[14] Moral rationality is thus one way to cash out morality. The alignment between morality and the well-being of a group of people is grounded in an extrinsic, relational view of moral values (see Rodogno, 2016).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

because the more sources of information are taken into account, the better the output can be attuned to these sources. For instance, a goal-directed process that takes into account the outcomes of response options allows for more flexibility than a stimulus-driven process in which this information is absent. More flexible processes, in turn, are more likely to produce accurate (i.e., theoretically rational) and hence adaptive (i.e., practically rational) outcomes (see Box 2.1).

The alignment of the two systems with the rational-irrational dichotomy has recently come under fire. Not all forms of complexity require more time and effort, or the differences may be negligible. Eventually, it is an empirical question to know just how much complexity can be handled under poor operating conditions (see Moors, 2014c; Moors et al., 2017).

1.3.5 Cognition

Now is perhaps a good time to turn to the multifarious meaning of the term cognition. Cognition is a contrastive notion: It takes on different meanings depending on the entities with which it is contrasted (Moors, 2007, 2009). When contrasted with the body, cognitive means mental. For scholars who believe that the realm of the mental is exhausted by representational entities (e.g., Brentano, 1874), cognitive is also synonymous with representational. Thus, cognitive processes are representation-mediated processes. For scholars who leave room within the mental for representational (i.e., Intentional) as well as non-representational (i.e., purely phenomenal) entities, cognitive also refers to representational but they use it to mark the boundary with the phenomenal part of the mental.

The term cognitive has also been used to point at a specific content or format of representation, to a specific type of operations, and to non-automatic processes. Thus, when contrasted with emotional representations, cognitive representations refer to representations with cold, non-valenced content. When contrasted with motivational or conative (i.e., dynamic or efferent) representations, cognitive representations refer to pure knowledge (i.e., non-dynamic or afferent) representations. When contrasted with perceptual or sensory representations, cognitive representations refer to conceptual or propositional representations. When contrasted with associative operations, cognitive operations are understood as rule-based. And when contrasted with automatic processes, cognitive processes are understood as non-automatic. A final meaning of cognitive is when it is used to refer to the mental super-level and contrasted with the observable and brain super-levels. In this book, I will specify the meaning of cognition I have in mind if the contrasting category is not obvious.

On a final note, scholars who take representations as the bearers of information are called representationalists. Throughout the history of

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

cognitive science, several scholars have explored non-representationalist alternatives (e.g., Gibson, 1979; Hutto & Myin, 2017; Stich, 1983; Wakefield & Dreyfus, 1991; see discussions by Bechtel, 1998a, 1998b). They maintain that cognition and Intentionality are possible without representations (see more in Chapter 6).

In the coming chapters, I apply the demarcation-explanation cycle to the emotion domain. As mentioned, theories in this domain are still trying to figure out what emotions are. That is, they try to find an adequate scientific definition for the set of emotions, or – if this turns out to be impossible – to replace it with sets that promise to be more fruitful. The sober fact that the first cycle has so far not led to a consensual scientific definition of emotion has not stopped researchers from moving on to further cycles in which emotions figure in other explananda, such as the influence of emotions on attention, perception, memory, judgment, decision-making or behavior, and psychopathology. The focus of this book will nevertheless be on emotion theories concerned with the first cycle.

In psychology, theories are known as evolutionary theories (e.g., Ekman, 1992a), network theories (e.g., Leventhal, 1984), appraisal theories (e.g., Lazarus, 1991; Roseman, 2013; Scherer, 2009b), the goal-directed theory (e.g., Moors, 2017a), psychological constructionist theories (e.g., Barrett, 2006b; Russell, 2003; Schachter, 1964), and social theories (e.g., Mesquita & Parkinson, 2022). In philosophy, theories go by the names of feeling theories (e.g., James, 1890b), judgmental theories (e.g., Green, 1992; Solomon, 1993), quasi-judgmental theories (e.g., Greenspan, 1988), perceptual theories (e.g., Tappolet, 2016), embodied theories (e.g., Colombetti, 2014; Deonna & Teroni, 2012; Griffiths, 2004b; Hutto, 2012; J. J. Prinz, 2004a), and motivational theories (e.g., Scarantino, 2014).

To facilitate the comparison of emotion theories, I organize them in a new typology built around the various stages in the demarcation-explanation cycle. Each of the stages presents questions for which different theories have provided different answers. The typology outlines a multi-axis space in which the axes correspond to the questions that are encountered during the consecutive stages of the cycle. Emotion theories can be placed and compared within this space depending on the answers they have provided to these questions. It may be noted upfront that this exercise is complicated by the fact that theories are not static entities, but evolve continuously. In addition to an analytic approach, in which I try to do justice to the idiosyncrasies of individual theories, I will also adopt a more synthetic approach, in which I try to identify axes that allow drawing fault lines (FL) between larger groups of theories. Chapter 2 identifies axes and fault lines and provides a broad overview

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press

of the possible choices that have been taken by emotion theories (see Table 2.4). Chapters 3–9 discuss emotion theories one by one. To structure my discussion of these theories, I had to create discrete categories again. As a basis for this, I selected one axis, that of the causal-mechanistic explanations of theories, which resulted in the following overarching categories or families: (a) evolutionary theories (including motivational theories), (b) network theories, (c) stimulus evaluation theories (including appraisal theories, judgmental theories, quasi-judgmental theories, perceptual theories, and embodied theories), (d) response evaluation theories (including the goal-directed theory), (e) psychological constructionist theories, and (f) social theories. Prior to my discussion of these theories, I start with two general precursors: Darwin (1872) and James (1890b; a feeling theory).

https://doi.org/10.1017/9781107588882.004 Published online by Cambridge University Press