Chapter 2
Mechanistic Explanation in Engineering Science

Abstract Explanation already loomed large in Chap. 1 on the explanatory utility of function ascriptions in engineering. In this chapter we take a closer look at the structure of (mechanistic) explanation in engineering. This analysis highlights different meanings that engineers attach to the notion of function, and clarifies the explanatory relevance of this ambiguity, it suggests an extension of the mechanistic program when applied to engineering science and, moreover, contains general lessons on the explanatory power of mechanistic explanations. In explicating the structure of mechanistic explanation, we will also address the question (iii) 'How does artifact x realize its capacity to ϕ?' and the relevance of function ascription in procuring an answer to this question. (we will address this relevance both for type and token-level cases).

Keywords Mechanistic explanation · Function · Engineering · Explanatory power

2.1 Introduction

Use of 'mechanism talk' is ubiquitous in both engineering science (e.g., Chandrasekaran and Josephson 2000; Goel 2013) and philosophical discussions of mechanisms (cf. Levy 2014). Engineered systems, such as pumps, car engines, mouse traps, toilets, soda vending machines, and the like are frequently used in illustrating various aspects of mechanisms and mechanistic explanation. Despite this reference to engineered systems in discussions of mechanisms and mechanistic explanation, focused philosophical analyses of the structure of mechanistic explanations in engineering science are scarce (cf. van Eck 2015a). There is very few philosophical work on engineering mechanisms that does more than (merely) use engineering mechanisms as a loose metaphor, and actually offers sophisticated understanding of what mechanistic explanation looks like in engineering practice. Moreover, although practicing engineers and biologists have been stressing conceptual ties between their disciplines for more than a decade (e.g., Csete and Doyle 2002), this connection has also received scant attention by philosophers, in

D. van Eck, *The Philosophy of Science and Engineering Design*,
SpringerBriefs in Philosophy, DOI 10.1007/978-3-319-35155-1_2

particular with respect to the use of engineering principles in the construction of mechanistic explanations in systems biology (cf. Braillard 2015). In this chapter I aim to make headway on both these issues.

In this chapter I give an outline of the structure of mechanistic explanation in engineering science, and organize this discussion around the usage of different meanings of technical function in engineering practice. I show that depending upon explanatory context, engineers use different conceptions of role function, i.e., *behavior* function and *effect* function, to individuate technical mechanisms and to develop mechanistic explanations. I argue that in order to capture this explanatory diversity, and thus to understand mechanistic explanation in engineering science, the mechanistic concept of role function needs to be regimented into these two domain-specific subtypes of role function when applied to the engineering domain. I illustrate this connection between subtypes of role function and explanatory requests in Sect. 2.2 in terms of token and type-level capacity explanations and in terms of malfunction explanations. The general insight that I take these cases to convey is thus that empirically-informed understanding of mechanistic explanation in engineering science requires sensitivity to this distinction in sub types of role function (van Eck 2015a).

In addition, in Sect. 2.3, I briefly discuss connections between (control) engineering and systems biology, focusing on the usage of engineering principles in the construction of mechanistic explanations in systems biology. Systems biology has adopted engineering tools and principles, in particular from control engineering, to model and explain complex biological systems. These tools are often in the service of characterizing the organization of mechanisms in abstract, truncated fashion. I briefly discuss a case of heat shock response in *Escherichia coli* to illustrate the role of engineering principles in mechanistic explanation in systems biology (cf. El-Samad et al. 2005; Braillard 2015). In this case, again, the distinction between the two subtypes of technical role function proves explanatorily relevant.

In Sect. 2.4, I revisit the engineering cases on capacity and malfunction explanation and argue that they give novel, general insights on the explanatory power of mechanistic explanations. I flesh out the distinctions between the explanatory desiderata of 'completeness and specificity' (Craver 2007) and 'abstraction' (Levy and Bechtel 2013) that are stressed in recent discussions on the explanatory power of mechanistic explanations in terms of these cases and argue that, rather than being in competition, as some authors have it, these desiderata are suitable for different explanation-seeking contexts. Furthermore, I argue that both desiderata fall short in the context of malfunction explanation, since they pull in opposite directions there, and elaborate a novel desideratum that can handle this explanatory context better. This desideratum, I argue, is applicable to both engineering and biological contexts of malfunction explanation.

2.2 Mechanistic Explanation in Engineering Science

2.2.1 *Mechanistic Explanation: Explanation by Decomposition and (Role) Function Ascription*

In this section, we will first have a brief look at the general structure of mechanistic explanation and then apply (and extend) the framework to engineering science. Although there are several accounts of mechanistic explanation on offer in the literature, there is broad consensus on a number of key features:

> All mechanistic explanations begin with (a) the identification of a phenomenon or some phenomena to be explained, (b) proceed by decomposition into the entities and activities relevant to the phenomenon, and (c) give the organization of entities and activities by which they produce the phenomenon. (Illari and Williamson 2012: 123).

Mechanistic explanations thus explain how mechanisms, i.e., organized collections of entities and activities, produce phenomena (Machamer et al. 2000; Glennan 2005; Bechtel and Abrahamsen 2005; Craver 2007). In the literature on explanation in the life sciences, it is now widely recognized that mechanisms play a central role in explaining complex capacities such as digestion, pattern recognition, or the maintenance of circadian rhythms. The idea is that to explain such capacities, one provides a model, or more generally a description/representation, of the mechanism responsible for that capacity.

Role function ascription plays a key role in the (b) decomposition of mechanisms (c) and the elucidation of their organization (Machamer et al. 2000; Craver 2001; Illari and Williamson 2010). As Machamer et al. (2000) write:

> Mechanisms are identified and individuated by the activities, and entities that constitute them, by their start and finish conditions, and by their functional roles. Functions are the roles played by entities and activities in a mechanism. To see an activity as a function is to see it as a component in some mechanism, that is, to see it in a context that is taken to be important, vital, or otherwise significant. (Machamer et al. 2000: 6)

Mechanistic role functions thus refer to activities that make a contribution to the workings of mechanisms of which they are a part, and mechanistic organization is key for the ascription of functions. For instance, in the context of explaining the circulatory system's activity of "delivering goods to tissues", the heart's "pumping blood through the circulatory system" is ascribed a function relative to organizational features such as the availability of blood, and the manner in which veins and arteries are spatially organized (Craver 2001: 64).

There is broad consensus in the literature on mechanistic explanation in the life sciences on the above-mentioned key features of mechanistic explanation, as well as on the importance of (role) function ascription and the functional individuation of mechanisms. And the strong suggestion that one can find in this literature is that the (functional) individuation of mechanisms proceeds in similar fashion in engineering science: frequently, mechanisms of technical artifacts, such as clocks,

mousetraps, and car engines, are invoked as metaphors to elucidate features of biological mechanisms (Craver 2001) and features of mechanisms in general (Glennan 2005; Darden 2006; Illari and Williamson 2012). The mechanistic concept of role function, and its utility in the functional individuation of mechanisms, has likewise been explicated in terms of mechanisms of technical artifacts such as car engines (Craver 2001). At the same time however, rigorous analysis of mechanistic explanatory practices in engineering are few and far in between. This invites the question whether the general framework on mechanistic explanation and mechanism individuation, as it is taken to work in the life sciences, can indeed be applied without significant modifications to engineering and able to provide understanding of mechanistic explanation in this domain.

In this chapter I argue that reference to such technical mechanisms is a loose metaphor and must not be understood as proving insight into mechanistic explanation in engineering science per se (cf. van Eck 2015a). In engineering science, technical mechanisms are not functionally individuated in terms of the concept of role function *simpliciter*. Rather, different notions of engineering function, 'behavior function' and 'effect function', are invoked to individuate technical systems and to explain their workings (van Eck 2015a). In order to capture mechanistic explanatory practices in engineering in well-informed fashion, the general perspective on the functional individuation of mechanisms thus needs to be extended to include both senses of engineering (role) function. In the next section I present the conceptual groundwork for this claim by briefly discussing how these varieties of function are used in mechanism individuation and mechanistic explanation in engineering science.

2.2.2 *Function and Functional Decomposition in Engineering*

Function is a key term in engineering (e.g., Chandrasekaran and Josephson 2000). Descriptions of functions figure prominently in, for instance, design methods (Stone and Wood 2000), reverse engineering analyses (Otto and Wood 2001), and diagnostic reasoning methods (Bell et al. 2007).

Despite the centrality of the term, function has no uniform meaning in engineering: different approaches advance different conceptualizations (Erden et al. 2008), and some researchers use the term with more than one meaning simultaneously (Chandrasekaran and Josephson 2000). This ambiguity led to philosophical analysis of the precise meanings of function involved. Vermaas (2009) regimented the spectrum of available function meanings into three '*archetypical*' engineering conceptualizations of function: *behavior function*–function as the desired behavior of a technical artifact; *effect function*–function as the desired effect of behavior of a technical artifact; *purpose function*–function as the purpose for which a technical

artifact is designed.[1] In the ensuing discussion, the notions of behavior function and effect function are (most) relevant.

Behavior functions are typically modeled as conversions of flows of materials, energy, and signals, where input flows and output flows in the conversion (are assumed to) match in terms of physical conservation laws (Stone and Wood 2000; Otto and Wood 2001). For instance, the function "loosen/tighten screws" of an electric screwdriver is then represented as a conversion of input flows of "screws" and "electricity" into corresponding output flows of "screws", "torque", "heat", and "noise" (cf. Stone and Wood 2000: 364). Since these descriptions of functions are specified such that input and output flows match in terms of physical conservation laws, they are taken to refer to specific physical behaviors of technical artifacts (Vermaas 2009).

Effect function descriptions refer to only the technologically relevant *effects* of the physical behaviors of technical artifacts: the requirements are dropped that descriptions of these effects meet conservation laws and that matching input and output flows are specified (Vermaas 2009). The function of an electric screwdriver is then described simply as, say, "loosen/tighten screws", leaving it unmentioned what the physical antecedents are of this effect. Behavior function descriptions thus refer to the 'complete' behaviors involved, including features like thermal and acoustic energy flows, whereas effect functions refer to subsets of these behaviors, i.e., desired effects.

Engineering descriptions and explanations of the workings of extant technical artifacts and artifact designs are often constructed by functionally decomposing functions into a number of sub functions. The relationships between functions and sets of their sub functions are often graphically represented in functional decomposition models. Like the concept of function, such models come in a variety of 'archetypical' flavors (van Eck 2011). For our purposes, the relevant ones are *behavior functional decomposition*—a model of an organized set of behavior functions, and *effect functional decomposition*—a model of an organized set of effect functions.

The use of (varieties of) functional decomposition is ubiquitous in engineering science in a variety of tasks, like conceptual engineering design (Stone and Wood 2000), failure analysis (Bell et al. 2007), and reverse engineering and redesign (Otto and Wood 2001). Cases in point are, amongst others, reverse engineering explanations which use elaborate behavior functions and functional decompositions, and malfunction explanations which use less detailed effect functions and functional decompositions.

[1]The term 'archetypical' here refers to 'most common'; the three conceptualizations of function are not meant to be exhaustive. For instance, some engineers use 'function' to refer to intentional behaviors of agents (cf. van Eck 2010). In reverse engineering analyses, 'function' refers to actual or expected behavior, without the normative connotation 'desired'.

2.2.3 *Reverse Engineering Explanation (and Redesign): Token Level Capacity Explanation*

In engineering science, reverse engineering and engineering design go hand in glove (e.g. Otto and Wood 2001; Stone and Wood 2000). Consider Otto and Wood's (2001) reverse engineering and redesign method, in which a reverse engineering phase in which reverse engineering explanations are developed for existing artifacts, precedes and drives a subsequent redesign phase of those artifacts. The goal of the reverse engineering phase is to explain how existing artifacts produce their overall (behavior) functions in terms of underlying mechanisms, i.e., organized components and sub functions (behaviors) by which overall (behavior) functions are produced. That is, reverse engineering—mechanistic—explanations give an answer to the question 'How does a particular artifact x realize its capacity to ϕ?'. These explanations of token level capacities are subsequently used in the redesign phase to identify components that function sub optimally and to either improve them or replace them by better functioning ones.

In the reverse engineering phase, an artifact is first broken down component-by-component, and hypotheses are formulated concerning the functions of those components. In this method, functions are behavior functions and represented by conversions of flows of materials, energy, and signals. After this analysis, a different reverse engineering analysis commences in which components are removed, one at a time, and the effects are assessed of removing single components on the overall functioning of the artifact. Such single component removals are used to detail the functions of the (removed) components further. The idea behind this latter analysis is to compare the results from the first and second reverse engineering analysis in order to gain potentially more nuanced understanding of the functions of the components of the (reverse engineered) artifact. Using these two reverse engineering analyses, a behavior functional decomposition of the artifact is then constructed in which the behavior functions of the components are specified and interconnected by their input and output flows of materials, energy, and signals (Otto and Wood 2001). Such models represent parts of the mechanisms by which technical systems operate, to wit: causally connected behaviors of components. Examples of an overall behavior function and behavior functional decomposition of a reverse engineered electric screwdriver are given in Figs. 2.1 and 2.2, respectively.

In the model in Fig. 2.2, temporally organized and interconnected behaviors are described. Components of artifacts are described in Otto and Wood's method in tables, what in engineering are called 'bills of materials', together with a model, called 'exploded view', of the components composing the artifacts. Taken together, these component and behavior functional decomposition models provide functional individuations and representations of mechanisms of artifacts.

Such (behavior functional decomposition) models are subsequently used to identify sub-optimally functioning components and so drive succeeding redesign phases (Otto and Wood 2001). The focus here is on the reverse engineering explanation-part of the methodology.

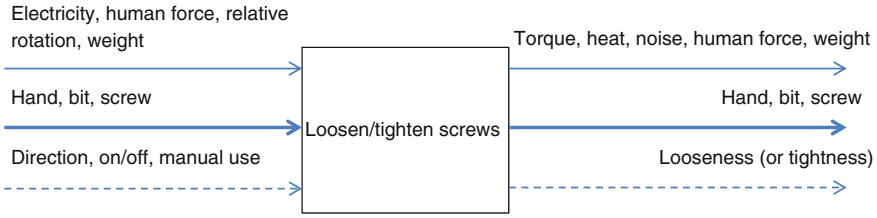

Fig. 2.1 Overall behavior function of an electric power screwdriver. *Thin arrows* represent energy flows; *thick arrows* represent material flows, *dashed arrows* represent signal flows (adapted from Stone and Wood 2000: 363, Fig. 2)

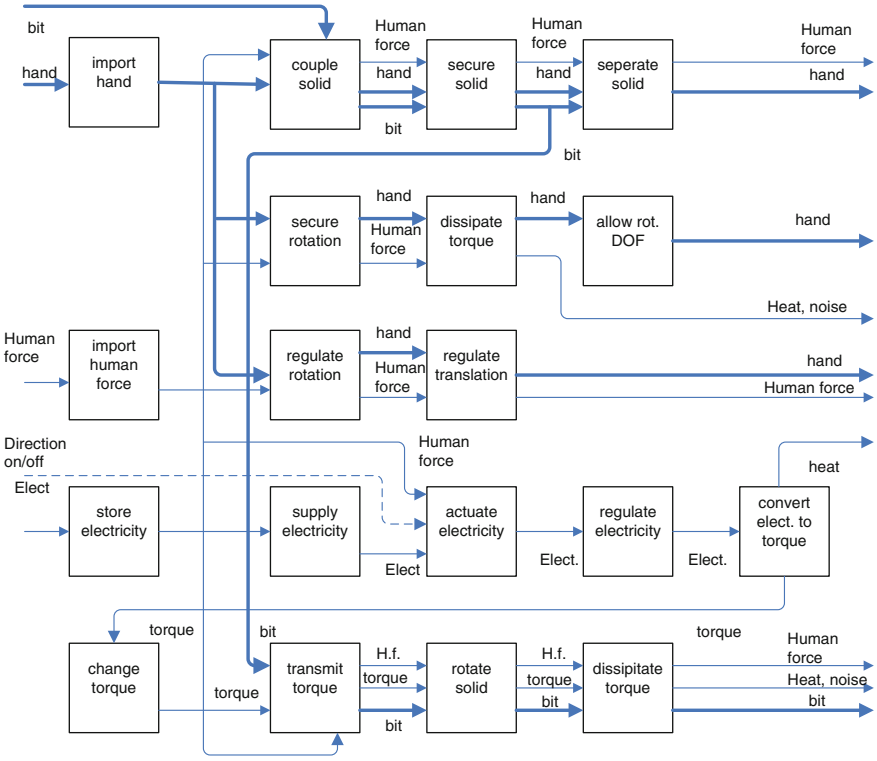

Fig. 2.2 Behavior functional decomposition of an electric power screwdriver. *Thin arrows* represent energy flows; *thick arrows* represent material flows, *dashed arrows* represent signal flows (adapted from Stone and Wood 2000: 364, Fig. 4; cf. Stone et al. 1998, 2000)

In malfunction explanation, this detail in mechanistic models is however not required: engineers take it that less detailed effect functions and functional decompositions there do a better explanatory job (see Chap. 1).

2.2.4 Malfunction Explanation

As we saw in Chap. 1, in malfunction analysis, explanation-seeking questions of the following format arise:

Why does artifact x not serve the expected function to ϕ?

Such questions are *contrastive*: why malfunction, rather than normal function? In the engineering literature, malfunction explanations that answer contrastive questions list different and fewer mechanistic features than reverse engineering explanations which answer questions about normal behavior or function. Such explanations are constructed using effect functions and functional decompositions.

Malfunction explanations in engineering pick out only a few features of mechanisms, i.e., those causal factors—failing components or sub mechanisms—that are taken to make a difference to the occurrence of a specific malfunction, as well as some course grained details of the containing mechanism to understand where the fault is located. Yet, most information about structural and behavioral specifics of malfunctioning components/sub mechanisms, and their containing mechanisms, is left out (Hawkins and Woollons 1998; Bell et al. 2007).[2]

Consider, again, by way of example, the Functional Interpretation Language (FIL) methodology for malfunction analysis and explanation (Bell et al. 2007). In FIL, functions are effect functions and represented in terms of their *triggers* and *effects*. Triggers describe input states that actuate physical behaviors which result in certain (expected) effects. For instance, the function description "depress_brake_pedal"-"red_stop_lamps_lit" of a car's stop light (p. 400). This description is a summary of some salient features of (manipulating) such artifacts; depressing the brake pedal will, if the system functions properly, result in the lighting of the stop lamps.

According to Bell et al. (2007) such trigger and effect representations serve two explanatory ends in malfunction analyses: firstly, they *highlight* relevant behavioral features of a given artifact, i.e., effects, and, simultaneously, provide the means to *ignore* less relevant or irrelevant behavioral features, i.e., physical behaviors underlying these effects; secondly, they support assessing which components are malfunctioning (pp. 400–401).

For instance, the trigger-effect representation "depress_brake_pedal"-"red_stop_lamps_lit" highlights the input condition of a pedal being depressed, and the resulting desired effect of lighted lamps, yet ignores the structural and behavioral specifics of the brake pedal and stop lamps, such as the pedal lever and electrical circuit mechanisms, as well as the energy conversions—e.g., mechanical energy

[2]That is, structural and behavioral characteristics are considered irrelevant in a first round functional analysis of malfunction. After this analysis, more detailed behavioral models of components and their behaviors are used for identifying specific explanatorily relevant structural and behavioral characteristics of malfunctioning components/sub mechanisms (Bell et al. 2007). However, immediately specifying these details in functional models is taken to result in listing a lot of irrelevant details.

conversions into electricity—that are needed to achieve this effect. Such representations only highlight those features that are considered explanatorily relevant to assess malfunctioning systems, and omit reference to physical behaviors/energy conversions by which desired effects are achieved.

Secondly, such trigger-effect descriptions support comparing normally functioning technical systems with malfunctioning ones (Bell et al. 2007). Trigger-effect descriptions support assessing whether the expected effects in fact obtain, and, if not, which and how components are malfunctioning (Bell et al. 2007). A normally functioning artifact, say the car's stop lights, has both a trigger and an effect occurring; the brake pedal is depressed and the stop lights are lit. Trigger-effect descriptions support analysis of two varieties of malfunction. First, a trigger may occur, yet fail to result in the intended effect. Say, the brake pedal is depressed, yet the stoplights are not on. Second, a trigger may not be occurring, yet the effect is nevertheless present. Say, the brake pedal is not depressed, yet the stoplights are on (see Bell et al. 2007). Such analysis of the actual states of triggers and effects allows one to focus on the most likely causes of failure (Bell et al. 2007). Say, if the pedal is depressed and the lights fail to ignite, first likely causes to investigate may be whether the electrical circuits in the lights are broken or the 'on/off' connection between the brake and electrical circuitry (connected to the lamp) is damaged. On the other hand, if the pedal is not depressed and the lights are lit, a first likely cause to investigate may be whether the 'on/off'connection between the brake and the electrical circuitry is damaged. To support more detailed malfunction analyses, functions are often decomposed into sub functions in FIL. An example of a functional decomposition of a two-ring cooking hob is given in Fig. 2.3.

The usage of effect functions and functional decompositions in FIL is the optimal choice given that function descriptions are used to black-box or suppress reference to unwanted behavioral and structural details. Effect function descriptions only highlight the relevant difference making properties with respect to malfunctioning artifacts, whereas more elaborate behavior function descriptions include irrelevant details such as, say, the thermal energy generated when lamps are lit.

Effect function descriptions also prove the optimal choice in the third explanation-seeking context that we consider: type level capacity explanation.

2.2.5 *Abstraction, Generality, and Type Level Capacity Explanation*

Explanatory models specified in terms of behavior function descriptions, which typically are represented by operations-on-flows (e.g. Hirtz et al 2002; Otto and Wood 1998, 2001; Pahl and Beitz 1988), as in the reverse engineering case, are fairly precise and complete when measured against models solely specified in terms of effect function descriptions, which typically are represented by verb-noun pairs (e.g., Bell et al. 2007; Deng 2002; Kitamura et al. 2005). The omission of details in

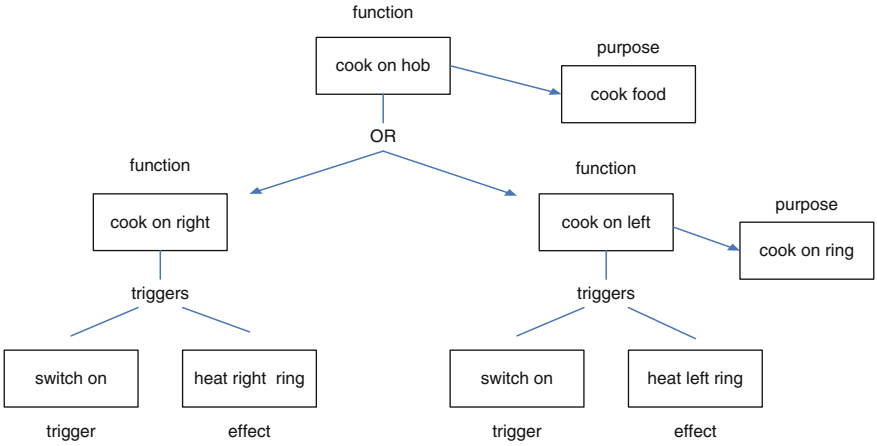

Fig. 2.3 Effect functional decomposition of a two-ring cooking hob (adapted from Bell et al. 2007)

explanatory models has important advantages, as discussions of abstraction and generality make clear (Weisberg 2007; Levy and Bechtel 2013); it makes such abstract models suitable for describing and explaining a larger class of technical systems, i.e., for type level capacity explanation rather than capacity explanation of individual tokens (as in the reverse engineering case). The Functional Concept Ontology (FCO) method for design and design knowledge management gives a good illustration of this point (Kitamura et al. 2005).

In a nutshell, the method uses knowledge bases in which, amongst others, functional descriptions of types of extant technical systems are archived, as well as part-whole relations between functions and sets of sub functions that compose 'upper level' functions. Functional descriptions in this method are descriptions of effect functions (van Eck 2011). The part-whole relations are 'enriched' with specifications of general technological principles by which sets of sub functions compose or achieve 'upper level' functions. These technological principles are called 'ways of achievement' (Kitamura et al. 2005). An example of an effect functional decomposition of a type of heavy duty stapler is given in Fig. 2.4.

By solely specifying effect functions and abstract, general technological principles, and omitting details about the precise manner in which materials, energies, and signals are processed, i.e., by not referring to behavior functions, such models are useful to capture the operation of types of mechanisms rather than individual tokens mechanisms. They focus on common features across token systems only, and omit reference to material energy and signal conversions that may differ across these token systems. They can be invoked to explain complex capacities of types of technical systems, here a type of heavy duty stapler, and such explanations are constructed using effect functions and functional decompositions.

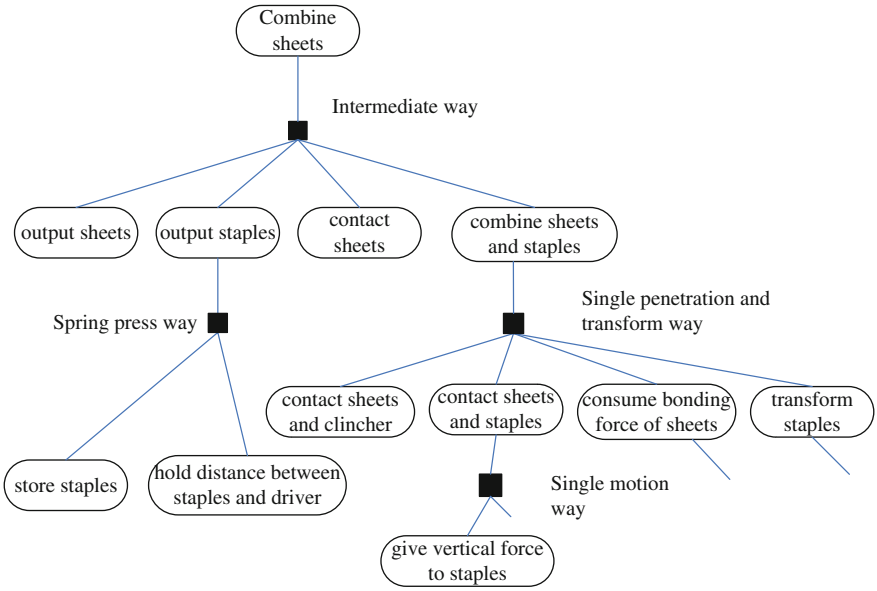

Fig. 2.4 Effect functional decomposition of a stapler. Functions are described in *ovals*, *black squares* refer to ways of achievement (adapted from Ookubo et al. 2007, p. 9, Fig. 3b.)

Both precision and generality are, as in other scientific domains, important in engineering: precise models offer in-depth understanding of the manner in which specific technical systems work and thus offer the means to adjust specific details in redesign phases in order to improve system functionality; more abstract and general models explain how types of technical systems operate. Such models are useful in (re) design contexts where predominantly knowledge on functional organization drives the initial design phase, and component-solutions are not considered in the initial phase of function specification, so as to consider different solution variants for these functional organizations (van Eck 2015b).

Since these desiderata of precision and generality are difficult to meet with single models, behavior functions and functional decompositions are used when precision is required and effect functions and functional decompositions are used when generality is needed. In engineering design, specific notions of function and functional decomposition are tailor-made depending upon the explanatory and/or design task at hand.

2.2.6 *Capturing Mechanistic Explanation in Engineering Science: Pluralism About Mechanistic Role Functions*

The upshot of these three cases is that explanations in engineering (as in every science of course) are constructed relative to explanatory objectives and,

importantly, that the level of detail included in these explanatory models hinges on specific concepts of technical function. This latter feature marks a relevant distinction with the manner in which role function ascription and mechanism individuation is understood in the literature on mechanistic explanation in the life sciences. Engineering scientists simplify or increase the details of explanations—functional decompositions—depending on the explanatory purpose at hand, and these adjustments are made using specific concepts of technical function (compare e.g., Figs. 2.2 and 2.3, or Figs. 2.2 and 2.4). In the context of reverse engineering explanation of complex capacities of token technical systems, elaborate or 'complete' descriptions of mechanisms are provided, in terms of behavior functions and functional decompositions, to answer the question how a specific technical system exhibits a given overall behavior. In malfunction explanation, less elaborate 'sketches' of mechanisms are provided in terms of effect functions and functional decompositions, referring only to some mechanistic features, namely those difference making factors that mark the *contrast* between normal functioning and malfunctioning technical systems. Finally, when explaining complex capacities of types of technical systems, abstracting away from specific details of individual token cases, effect functions and functional decompositions are invoked. So, depending upon explanatory context, mechanisms are individuated in different ways using different conceptualizations of function in engineering science. Function ascription thus again proves highly relevant, both for type and token level capacity explanation and for malfunction explanation. Importantly, neither function conceptualization in itself accommodates both ways in which mechanisms are functionally individuated in engineering science. Behavior and effect function ascriptions are invoked to individuate mechanisms in different ways depending on the task at hand.

However, this distinction in functional individuation, and its reliance on different subtypes of function, is blurred in a perspective that understands mechanism individuation and mechanistic explanation in terms of mechanistic role function ascription simpliciter. The concept of mechanistic role function, an activity that makes a contribution to the workings of a mechanism of which it is a part, admits of two interpretations in the context of engineering science: behavior function on the one hand and effect function on the other. So the point is that in order to arrive at empirically informed understanding of explanatory practices in engineering, and at consistency of the general structure of mechanistic explanation with these practices, regimenting the concept of role function into domain-specific engineering concepts of behavior and effect function, i.e., sub types of role function, is needed.[3]

[3]Note that behavior and effect descriptions of function describe, in different ways, the contributions of components to mechanisms of which they are a part. The distinction between behavior and effect function thus is not to be conflated with the distinction between a mechanism description and a description of a mechanisms' overall activity. Neither is the behavior-effect function distinction to be conflated with the distinction between 'isolated' and 'contextual' descriptions of an entity's activity (Craver 2001): isolated descriptions describe activities without taking into account the mechanisms in which they are situated; contextual descriptions describe activities in terms of the

I now briefly consider another facet of the relationship between mechanistic explanation and engineering that has received little sustained analysis: the usage of engineering principles in the construction of mechanistic explanations in systems biology. Here we will see again that the distinction in subtypes of role function is relevant; the manner in which biological mechanisms are individuated in engineering terms, hinges on specific engineering conceptualizations of function. Specifically, effect function descriptions are used to describe and explain biological mechanisms in abstract, truncated fashion.

2.3 Explanation by Effect Functional Decomposition: Where Engineering and Systems Biology Meet

2.3.1 Engineering and Mechanistic Explanation in System Biology: The E. coli *Heat Shock Case*

Although philosophy, it seems, is only recently picking up on the fruitful cross-talk between engineering and systems biology (cf. Braillard 2015), engineers and systems biologists alike have been stressing the conceptual ties for more than a decade (Hartwell et al. 1999; Lazebnik 2002). With biological data about complex biological systems exploding during the last twenty years or so, due to (functional) genomics projects and the like, opportunities to understand complex biological systems in far greater detail became available. Yet cashing out that promise also signaled the need for new tools that enabled massive data analysis and integration in order to build explanatory models of these complex systems with a scale and complexity hitherto unknown. Here is where, amongst others, engineering tools came in. For instance, decomposition and control principles governing the construction of engineering systems are now being used to characterize complex biological systems (Tomlin and Axelrod 2005).

A case in point is research by El-Samad et al. on the mechanism(s) to counter heat shock in *E. coli* (El-Samad et al. 2005; cf. Tomlin and Axelrod 2005; Braillard 2015). Heat shock response is a widely conserved response of cells to cope with environmental stress brought about by unusual increases in temperature, involving the induced expression of heat shock proteins. Such temperature increases can damage proteins by breaking down their tertiary structures. Heat shock proteins come in two varieties and mitigate this effect in two different ways: molecular chaperones do so by refolding denatured proteins and proteases by degrading denatured proteins. If the response is sufficiently swift and massive, cell death can

(Footnote 3 continued)

mechanistic contexts in which they are situated and to which they contribute. Both behavior and effect functions are of the contextual variety, describing contributions of components to the mechanisms of which they are a part.

be prevented by protein repair and/or removal of damaged proteins. The response needs to be tightly controlled in the sense that it is only activated in case of heat shock, since the response is highly energy consuming and would make too high energy demands if heat shock proteins would be produced all the time. Cells thus must maintain a delicate balance between the protective effect of heat shock protein production and the metabolic cost of overproducing these proteins. In *E. coli*, the RNA polymerase cofactor σ^{32} promotes the transcription of heat shock proteins. After heat shock stress—temperature increase—σ^{32} activity increases, resulting in the transcription of specific heat shock gene promoters, which initiate transcription of genes, which in turn encode specific heat shock proteins—chaperones and proteases. This heat shock protein expression, when appropriate, prevents cell death. This mechanism uses both feed forward and feedback loops that process information about temperature and the folding state of proteins in the cell. σ^{32} activity is crucial in all this and depends on a feed forward mechanism that senses temperature and controls σ^{32} transcription, and feedback regulatory mechanisms that register the folding levels of proteins (levels of denatured cellular protein) and degrade σ^{32}. These regulatory feedback mechanism are crucial to ensure that σ^{32} synthesis, activity, and stability is brought back to normal levels after a sufficient number of heat shock proteins have been produced and the threat to cell death is averted.

El-Samad et al. (2005) constructed a quantitative, mathematical model of the heat shock response in order investigate the dynamical, mechanistic organization that sustains the heat shock response. They came up with an elaborate mathematical model consisting of 31 equations and 7 parameters. To make the model computationally tractable and be able to pose and answer questions about the dynamical, mechanistic organization of the system, the original model had to be trimmed down. As Braillard (2015) stressed, control engineering principles played an important heuristic role in this model reduction, i.e., abstraction, and thus in the discovery of the mechanism' core organizational features that sub serve its overall regulatory behavior. The close analogy between engineered systems and biological ones with respect to functional modular organizations sub serving regulatory processes made this possible. As El-Samad et al. (2005) explain:

> Control and dynamical systems theory is a discipline that uses modular decompositions extensively to make modeling and model reduction more tractable. Because biological networks are themselves complex regulation systems, it is reasonable to expect that seeking similarities with the functional modules traditionally identified in engineering schemes can be particularly useful. (El-Samad et al. 2005: 2737).

In control engineering, decomposition into functional modules (modules defined in terms of their effect-role functions) often begins with identification of the process to be regulated called the 'plant' (cf. Lind 1994), for instance altitude regulation of an airplane or temperature regulation of a thermostat. Modules of the system that contribute to the regulation are described in terms of their contributing functions, the most common of which are 'sensors', 'detectors', 'controllers', 'actuators', and 'feed forward' and 'feedback' signals. For instance, in a simple heating system, the

plant is the temperature regulation process, which is achieved, inter alia, by a sensor module which measures ambient temperature, calculates the deviation from the desired temperature and feeds this information into the thermostat (controller). The thermostat then outputs signals that are send to an actuator (heat fuel valve) that generates an actuation signal (e.g., fuel to furnace) that corrects deviation from the desired temperature. The sensor module again measures the ambient temperature and, if needed, feeds back information on temperature deviations to the controller, and so on.

El-Samad et al. (2005) applied this control engineering perspective to the *E. coli* heat shock response system. In this application, the protein folding task (the refolding of denatured proteins) is taken to be the process to be regulated (plant), the feed forward signal (send by a sensor) is the temperature dependent translational efficiency of $ø^{32}$ synthesis, the controller is the level of $ø^{32}$ activity, chaperones function as actuators of the plant (the actuated plant input is the number of molecular chaperones), and sensors measure plant output (amount of denatured protein), which in turn is fed back to the controller.

This decomposition allowed El-Samad et al. (2005) to construct a simplified model consisting of just 6 equations and 11 parameters in which each equation describes aspects of the behavior of a module. They remark:

> This model provides useful insight into the heat shock system design architecture. It also suggests a mathematical and conceptual modular decomposition that defines the functional blocks or submodules of the heat shock system. This decomposition is drawn by analogy to manmade control systems and is found too constitute a canonical blueprint representation for the heat shock network. (El-Samad et al. 2005: 2736)

What we here thus see is that analogical reasoning with respect to regulation processes and the functional architecture sub serving these processes in engineered and biological systems, led to a functional modular decomposition of a biological system in terms of effect function descriptions that laid bare core organizational features of the system by which it produces regulatory behavior. Engineering tools —modular decompositions specified in terms of effect functions—here serve as a discovery heuristic for a mechanism' core organizational features that sub serve its overall regulatory behavior (cf. Braillard 2015) This usefulness of engineering concepts, i.e., modular decompositions in terms of effect functions, is not specific to the *E. coli* case, but generalizes to a variety of cases (cf. Tomlin and Axelrod 2005) and suggests a general discovery heuristic:

> If the heat shock mechanism can be described and understood in terms of engineering control principles, it will surely be informative to apply these principles to a broad array of cellular regulatory mechanisms and thereby reveal the control architecture under which they operate (Tomlin and Axelrod 2005: 4220).

Analysis of engineering function and explanation has more to offer. In concluding this chapter, I revisit the engineering explanation-seeking contexts from Sect. 2.2 and suggest that these illustrate the complementarity of two allegedly competing perspectives, 'completeness and specificity' (Craver 2007) and 'abstraction' (Levy and Bechtel 2013), on the explanatory power of mechanistic

explanations. And that they pull in opposite directions in the context of malfunction explanation and, hence, that a novel desideratum is required to handle this explanatory context.

2.4 Explanatory Power: Rethinking the Explanatory Desiderata of 'Abstraction' and 'Completeness and Specificity'[4]

According to one influential perspective, the power of mechanistic models is (almost) always increased when these refer to both functional and structural features of mechanisms (Machamer et al. 2000; Craver 2007). On the counterview, mechanistic models have in certain contexts more explanatory traction when reference to structural aspects of mechanisms is suppressed. Models that solely describe functional characteristics, i.e., causal relations between components, explain better how organization impacts the behavior of mechanisms (Levy and Bechtel 2013). The engineering cases presented here allow for a more fine-grained understanding of the relationship between these views: rather than being in competition, they emphasize different explanatory virtues that hold in different explanation-seeking contexts.

I have argued elsewhere that differences between these two (allegedly) competing perspectives on the explanatory power of mechanistic explanations, 'completeness and specificity' (Craver 2007) and 'abstraction' (Levy and Bechtel 2013), essentially boil down to differences in the notions of difference making endorsed in these accounts and that they are in fact not in competition (van Eck 2015a). They are rather suitable for different explanation-seeking contexts. Whereas abstraction dictates that mechanistic explanations should only list the 'primary factors' responsible for the occurrence of system function, 'completeness and specificity' prescribes that in addition to primary ones also 'higher order factors' should be described, which concerns factors influencing the precise manner in which a system function occurs or those sub serving the primary factors. The engineering cases gives an empirical illustration of this 'complementarity view'.

2.4.1 Malfunction Explanation: Local Specificity and Global Abstraction

In the context of reverse engineering explanation presented here, i.e., token level capacity explanation, engineers take details to matter: elaborate behavior functional decompositions, and related component models, are constructed to describe the mechanisms of specific artifacts, via the breaking down of artifacts

[4]This section draws on van Eck (2015a).

component-by-component and assessing the effects of single component removals on their overall behaviors. This perspective agrees with the 'completeness and specificity' view on mechanistic explanations. In the model of the reverse engineered electrical screwdriver in Fig. 2.2, for instance, both factors that make a difference to the occurrence of the screwdriver's overall behavior are listed, such as 'supply electricity' and 'convert electricity to torque', as well as factors that affect the way in which this behavior is manifested, such as 'dissipate torque' into 'heat' and 'noise' flows, and 'allow rotational degrees of freedom' (the latter concerns controlling the movement of materials along a specific degree of freedom (Stone and Wood 2000), here appropriate hand positions for correct functioning of the screwdriver).

Such primary and higher order details matter given that the reverse engineering explanation ultimately is in the service of redesign purposes: identifying components that function sub-optimally in a reverse engineered artifact and subsequent optimization in redesigned artifacts. The manner in which a particular technical system exhibits a given piece of behavior then becomes important. For instance, in an empirical example of the reverse engineering of an electric wok and its subsequent redesign, structural features of components affected the precise manner in which temperature distribution across the wok's bowl was manifested, and modifications of these features were needed to optimize temperature distribution across the bowl; the electric heating elements of the wok, such as a bimetallic temperature controller, were housed in too narrow a circular channel and optimized in the redesign phase (Otto and Wood 1998).

The abstraction perspective, on the other hand, is suitable in the context of type level capacity explanation. There the omission of details concerning the precise manner in which materials, energies, and signals are processed sub serves the description and explanation of the workings of (multiple) types of technical systems, rather than specific token systems. Such models only require the specification of primary factors that affect the occurrence of specific complex capacities. For instance, the capacity of heavy duty staplers to 'connect sheets' (cf. Fig. 2.4). Higher order details are not needed, since these are or might be specific to particular tokens systems.

At first glance, it seems that the abstraction perspective is also better suited to capture malfunction explanation. In that context, as we saw, engineers advance the maxim that 'less is more' when it comes to adequate explanations. Closer inspection however reveals that in this explanatory context 'abstraction' and 'completeness and specificity' pull in opposite directions (van Eck 2015a).

To see this, consider that in order to understand how a malfunctioning component or sub mechanism makes a difference to the *occurrence* of a specific system level malfunction, one needs to know how the failing component or sub mechanism is situated within a mechanism that underlies normal functioning. That is, malfunctions are identified against a backdrop of normal mechanism functioning (cf. Thagard 2003; Moghaddam-Taaheri 2011). This is required to explain the contrast drawn in the explanandum—why malfunction, rather than normal function.

This also happens in FIL, in which function descriptions and functional decomposition models in terms of trigger-effect descriptions are used to specify normal functioning, and to provide the context against which to assess specific malfunctions, such as a trigger that occurs yet fails to result in an expected effect—say, a cooking hobs' switch that is on but does not result in the heating of a ring (Bell et al. 2007). Such contrastive factors that explain the contrast drawn in the explanandum, i.e., make the difference, between malfunction and normal function are primary ones that underlie the occurrence of the specific system-level malfunction in question. Say, in the above example, the electrical circuitry connected to the ring that is damaged as a result of which the ring does not heat, and food cannot be heated. Also the details on normal functioning that are needed to understand why the factor(s) cited in the explanans, e.g., a broken electrical wiring, is a contrastive one, concerns primary factors that underlie normal functioning. Since fact and foil in the contrastive explanandum concern the occurrence of malfunction and function, respectively, the factors needed to understand which part(s) of the mechanism malfunction and which ones function normally should be primary ones as well. Information on the precise manner in which mechanisms normally manifest their functions is irrelevant here. Knowing that rings of cooking hobs normally heat when switches are thrown is sufficient to understand that when this trigger-effect relation does not obtain, a malfunction occurs.

Also, it suffices to describe properly functioning parts of mechanisms in abstract fashion, i.e., in terms of functionally characterized components and their functions, since their job is only to highlight where in the mechanism a malfunctioning component or sub mechanisms is located. Listing structural features, such as size and shape, is irrelevant here for what matters is knowing what these components/sub mechanisms (normally) do. I here label the constraint to specify common features of functioning and malfunctioning mechanisms in terms of functionally characterized components and their functions, '*global abstraction*'. However, the contrastive factor(s) that makes the difference to the occurrence of a specific system-level malfunction often will have to be described in more elaborate fashion and its description will, in addition to functional characteristics, also refer to structural features. The manner in which a component is, say, broken or worn often does make a difference to the occurrence of a system level malfunction. A rupture in the electrical wiring of the cooking hob, for instance, which leads to failure of the ring to heat. Here specificity with respect to structural features is needed as well. I label this constraint to describe both functional and structural characteristics of contrastive difference makers, '*local specificity*' (both to set it apart from 'global abstraction', and from 'completeness' in the sense of specifying both primary and higher order factors; 'local specificity' as I understand it here concerns primary factors only).[5]

[5]This is in keeping with engineering practice. After a first round functional analysis of malfunction, more detailed behavioral models of components and their behaviors are used in FIL for assessing specific structural characteristics of malfunctioning components (Bell et al. 2007).

Malfunction explanations thus require a format in between 'completeness and specificity' and 'abstraction': they require *local specificity* with respect to descriptions of malfunctioning components/sub mechanisms and *global abstraction* with respect to descriptions of the mechanisms in which the component/sub mechanism failures are placed. This analysis extends current thinking about the explanatory power of mechanistic explanations by spelling out a novel desideratum for malfunction explanations. The lesson is that in this context, explanations that contain local specificity and global abstraction are better than either complete or abstract mechanistic explanations. And, as we saw, in the context of engineering science, depending on the richness that is required of explanations, specific concepts of technical function and functional decomposition are invoked. The examples of reverse engineering explanation/token level capacity explanation analyzed here use behavior functions and functional decompositions, whereas malfunction explanations and type level capacity explanations are procured in terms of effect functions and functional decompositions.

A further question emerges: is 'local specificity and global abstraction' a desideratum only for malfunction explanations of technical systems, or does it also apply to malfunction explanations in other scientific domains, like biology? I argue below that explanations of biological malfunctions also best exhibit 'local specificity and global abstraction'.

2.4.2 Malfunction Explanation in Biology

Also in the case of explaining biological malfunction, I take it that explanations that are locally specific and globally abstract are the optimal ones. Consider, for instance, impaired blood circulation in the circulatory system.[6] Malfunction explanations, of course, should single out those steps—entities engaging in activities—in the circulatory system's mechanism(s) that cause the circulation of blood to be impaired, i.e., make a difference to whether or not impaired blood circulation occurs. In the case of impaired blood distribution, the cause may be that blood transport is disrupted in particular vessels as a result of thrombosis in those vessels. The description of these contrastive factors—damaged vessels due to thrombosis—often will have to be described in elaborate fashion, i.e., in terms of both functional and structural specifics. In our example, it is relevant to know that the damaged vessels fail to perform their function of transporting blood. Yet the manner in which those vessels are damaged, and thus fail to perform their function(s), also makes a difference to the occurrence of impaired blood circulation. When the vessels are only slightly damaged they may still perform their function of transporting blood, so it is relevant to know the nature of the damage, i.e., the manner in which

[6]I adapt this example from Nervi (2010).

structural features of the vessels are deformed. Here, deformations due to thrombosis. Local specificity thus applies to descriptions of such contrastive difference makers.

And, again, to explain the contrast drawn in the explanandum—why malfunction, rather than normal function—one also needs to know how the failing component or sub mechanism is situated within a mechanism that underlies normal functioning, since malfunctions are identified against a backdrop of normal mechanism functioning (cf. Thagard 2003; Moghaddam-Taaheri 2011). However, descriptions of the relevant properly functioning parts of mechanisms can be given in abstract terms—functionally characterized components and their functions—since their job is only to highlight where in the mechanism a malfunctioning component or sub mechanisms is located. It suffices to know that, say, the cardiac muscle engages in coordinated contraction, that blood is ejected from the ventricles into the aorta and the arterial system, etc. Further detailing of structural specifics, say, the precise shape or size of the cardiac muscle has no added value for locating the fault(s) in the mechanism. So, the desideratum of 'local specificity and global abstraction' is not restricted to malfunction explanations of technical systems, but applies more broadly to malfunction explanations in the biological domain as well.

References

Bechtel, W., & Abrahamson, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421–441.

Bell, J., Snooke, N., & Price, C. (2007). A Language for functional interpretation of model based simulation. *Advanced Engineering Informatics, 21*, 398–409.

Braillard, P.A. (2015). Prospects and limits of explaining biological systems in engineering terms. In P.A. Braillard & C. Malaterre (Eds.), *Explanation in biology* (pp. 319–344). Springer.

Chandrasekaran, B., & Josephson, J. R. (2000). Function in device representation. *Engineering with Computers, 16*, 162–177.

Craver, C. F. (2001). Role functions, Mechanisms, and Hierarchy. *Philosophy of Science, 68*, 53–74.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science, 295*, 1664–1669.

Darden, L. (2006). *Reasoning in biological discoveries*. Cambridge: Cambridge University Press.

Deng, Y. M. (2002). Function and behavior representation in conceptual mechanical design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 16*, 343–362.

El-Samad, H., Kurata, H., Doyle, J. C., Gross, C. A., & Khammash, M. (2005). Surviving heat shock: control strategies for robustness and performance. *PNAS*, *102*(8), 736–2741.

Erden, M. S., Komoto, H., van Beek, T. J., D'Amelio, V., Echavarria, E., & Tomiyama, T. (2008). A review of function modeling: approaches and applications. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *22*, 147–169.

Glennan, S. (2005). Modeling mechanisms. *Studies in the History and Philosophy of the Biological and Biomedical Sciences, 36*(2), 375–388.

Goel, A. K. (2013). A 30-year case study and 15 principles: Implications of an artificial intelligence methodology for functional modeling. *AIEDAM, 27*(3), 203–215.

Hawkins, P. G., & Woollons, D. J. (1998). Failure modes and effects analysis of complex engineering systems using functional models. *Artificial Intelligence in Engineering, 12*(4), 375–397.

Hartwell, L. H., Hopfield, J. J., Leibner, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature, 402*, C47–C52.

Hirtz, J., Stone, R. B., McAdams, D. A., Szykman, S., & Wood, K. L. (2002). A functional basis for engineering design: reconciling and evolving previous efforts. *Research in Engineering Design, 13*, 65–82.

Illari, P., & Williamson, J. (2010). Function and organization: Comparing the mechanisms of protein synthesis and natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences, 41*, 279–291.

Illari, P., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science, 2*, 119–135.

Kitamura, Y., Koji, Y., & Mizoguchi, R. (2005). An ontological model of device function: Industrial deployment and lessons learned. *Applied Ontology*, *1*, 237–262.

Lazebnik, Y. (2002). Can a biologist fix a radio?—Or, What I learned while studying apoptosis. *Cancer Cell, 2*, 179–182.

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of science*, *80*, 241–261.

Levy, A. (2014). Machine-likeness and explanation by decomposition. *Philosopher's imprint, 6*, 1–15.

Lind, M. (1994). Modeling goals and functions of complex industrial plants. *Applied Artificial Intelligence, 8*, 259–283.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 57*, 1–25.

Moghaddam-Taaheri, S. (2011). Understanding Pathology in the context of physiological mechanisms: The practicality of a broken-normal view. *Biology and Philosophy, 26*, 603–611.

Nervi, M. (2010). Mechanism, malfunctions and explanation in medicine. *Biology and Philosophy, 25*, 215–228.

Ookubo, M., Koji, Y., Sasajima, M., Kitamura, Y., Mizoguchi, R. (2007). Towards interoperability between functional taxonomies using an ontology-based mapping. In *Proceedings of the International Conference on Engineering Design (ICED 07)*, 28-31 Aug 2007, Paris, France: 1–12.

Otto, K. N., & Wood, K. L. (1998). Product evolution: A reverse engineering and redesign methodology. *Research in Engineering Design, 10*, 226–243.

Otto, K. N., & Wood, K. L. (2001). *Product design: Techniques in reverse engineering and new product development*. Upper Saddle River NJ: Prentice Hall.

Pahl, G., & Beitz, W. (1988). *Engineering design: A systematic approach*. Berlin: Springer.

Stone, R. B., & Wood, K. L. (2000). Development of a Functional basis for design. *Journal of Mechanical Design, 122*, 359–370.

Stone, R. B., Wood, K. L., & Crawford, R. H. (1998). A heuristic method to identify modules from a functional description of a product. *ASME proceedings*, 1–21.

Stone, R. B., Wood, K. L., & Crawford, R. H. (2000). A heuristic method for identifying modules for product architectures. *Design Studies, 21*, 5–31.

Thagard, P. (2003). Pathways to biomedical discovery. *Philosophy of Science, 70*, 235–254.

Tomlin, C. J., & Axelrod, J. D. (2005). Understanding biology by reverse engineering the control. *PNAS, 102*(12), 4219–4220.

van Eck, D. (2010). On the conversion of functional models: Bridging differences between functional taxonomies in the modeling of user actions. *Research in Engineering Design, 21*(2), 99–111.

van Eck, D. (2011). Supporting design knowledge exchange by converting models of functional decomposition. *Journal of Engineering Design, 22*(11–12), 839–858.

van Eck, D. (2015a). Mechanistic explanation in engineering science. *European Journal for Philosophy of Science, 5*(3), 349–375.

van Eck, D. (2015b). Validating function-based design methods: An explanationist perspective. *Philosophy and Technology, 28*, 511–531.
Vermaas, P. E. (2009). The Flexible Meaning of Function in Engineering, *Proceedings of the 17th International Conference on Engineering Design (ICED 09)*:vol. 2. 113–124.
Weisberg, M. (2007). Three kinds of idealization. *The journal of Philosophy, 104*(12), 639–659.

http://www.springer.com/978-3-319-35154-4

The Philosophy of Science and Engineering Design

van Eck, D.

2016, IX, 75 p. 9 illus., Softcover
ISBN: 978-3-319-35154-4