ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Stud. Hist. Phil. Sci. 37 (2006) 81–96

Studies in History and Philosophy of Science

www.elsevier.com/locate/shpsa

Mechanistic artefact explanation

Jeroen de Ridder

Delft University of Technology, Department of Philosophy, Jaffalaan 5, 2628 BX Delft, The Netherlands

Abstract

One thing about technical artefacts that needs to be explained is how their physical make-up, or structure, enables them to fulfil the behaviour associated with their function, or, more colloquially, how they work. In this paper I develop an account of such explanations based on the familiar notion of mechanistic explanation. To accomplish this, I (1) outline two explanatory strategies that provide two different types of insight into an artefact's functioning, and (2) show how human action inevitably plays a role in artefact explanation. I then use my own account to criticize other recent work on mechanistic explanation and conclude with some general implications for the philosophy of explanation.

Keywords: Artefact; Technical function; Explanation; Levels of explanation; Mechanisms

1. Introduction

If anything, *technical artefacts* must be susceptible to mechanistic explanation; an explanation that exhibits the causal mechanisms underlying a behavioural regularity. Tailored to the artefact case, a mechanistic explanation consists in showing the causal mechanisms underlying an artefact's behaviour. Since artefacts are designed, created, and used for their ability to exhibit one or a few specific behaviours associated with their function(s), the typical candidate behaviour requiring explanation is the behaviour that corresponds to the artefact's function. Therefore, I will henceforth use the term

E-mail address: g.j.deridder@tudelft.nl (G. J. de Ridder).

0039-3681/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2005.12.009

mechanistic artefact explanation to refer to explanations of how an artefact is able to show the behaviour linked with fulfilling (one of) its function(s).[1] Or, to put it more colloquially, a mechanistic artefact explanation is an answer to the question: How does this thing work?

My goal in this paper is two-tiered. Primarily, I want to give an account of mechanistic artefact explanations. However, I believe that this project generates spin-off that bears on mechanistic explanation in general. Hence, my secondary goal is to look at some recent work on mechanistic explanation and to indicate how it may benefit from insights derived from my account of artefact explanation.

Given these goals, the set-up of the paper follows straightforwardly. I begin with some real-life examples of artefact explanations in the next section. These examples lead to reflections on (1) the structure of mechanistic artefact explanations (Section 3) and (2) the role human action plays in them (Section 4). Taken together, these reflections provide the ingredients for a general account of mechanistic artefact explanation. In Section 5, a comparison of my view with other work reveals how my account exposes an ambiguity that seems to pervade other recent accounts of mechanistic explanation. In Section 6, I conclude with some thoughts on what I dub the *realization-independency* of mechanistic explanation.

Before I present my examples, let me first sketch the expected pay-offs of this project. First, philosophers writing on mechanistic explanation tend to be rather imprecise in their characterization of it. This is definitely true for the original proponents of mechanistic explanation (Railton, 1978; Salmon, 1984), but in my opinion even more recent work (Machamer, Darden, & Craver, 2000; Craver, 2001; Glennan, 2002) still lacks sufficient precision. My account reveals some ambiguities arising out of this imprecision.

Secondly, there is something idiosyncratic about mechanistic artefact explanation. It always involves reference to human action under some appropriate description. This is not supposed to be a very profound assertion; I only wish to draw attention to the trivial point that artefacts do not *do* anything without human agency. They ‘work’ only when we use them and as a result, an explanation of their working must include information about human action. This feature sets artefact explanation apart from mechanistic explanation in the sciences. Natural science aims to describe mechanisms in the world that do not involve human action, while social science focuses on social mechanisms, which do not include physical objects[2] but only (collective) human intentional states and actions. On the face of it, mechanistic artefact explanations involve an unusual explanatory combination of human action and physical objects. Showing how this combination works is therefore a second pay-off.

Thirdly, I conjecture that my account will also bring out another central feature of mechanistic explanation, which, as far as I know, has gone largely unnoticed in recent discussions. It is the *realization-independency* I referred to above.

[1] My project thus does not concern functional explanation. In a mechanistic artefact explanation, the function figures in the explanandum; a functional explanation is traditionally conceived as having a function ascription in the explanans.

[2] At least not in a non-symbolic role, I presume.

2. Artefact explanation in action

As examples, I have selected explanations of two relatively simple pieces of technology. The first example is a paperclip and the second is Thomas Edison's first 'phonograph or speaking machine'. Paperclips have come in many shapes, sizes, and colours for over a century.[3] They all work according to the same basic principle. The example I want to discuss, however, explicitly shows human action in its patent, which makes it especially apt for expository purposes. Take a look at Figure 1.[4] It nicely depicts what one is supposed to do with this paperclip to make it work.

At the risk of belabouring the obvious, let's look a little closer at how a paperclip works.[5] It does not just slide on papers by itself; we have to put it in place by opening the clip and pressing the longer loop against one side of the papers and at the same time flexing it just enough for the smaller loop to slide over the other side. The paperclip in my example actually has to be handled slightly differently, as the patent pictures show vividly. What subsequently makes the paperclip do its job is *springiness*; or rather the force exerted on the papers that results from the paperclip's springing action. Like any material, steel has a characteristic springiness or elasticity. The behaviour resulting from this elasticity is described by Hooke's law: 'As the extension so the force'. More prosaically: $F = -kX$. Up to a limit, the material's elastic limit, beyond which permanent deformation occurs, the more we stretch something, the more resistance it offers to further stretching. This is the key to paperclip behaviour: flexing the loops somewhat out of their normal plane position causes them to exert a force in the opposite direction, thus clamping onto the papers.

Although Thomas Edison's 'phonograph or speaking machine' is a more sophisticated device than the paperclip, its operational principles are still relatively easy to grasp. The patent drawings are shown in Figure 2 (note that the drawings labelled 'Fig. 3' and 'Fig. 4' are alternative recording devices, not shown in the main drawings).[6] The letters in the drawing indicate the principal components: *A* is a cylinder with a helix-shaped groove pattern on its surface. Recording material, according to Edison preferably metallic foil, is fastened on the surface of *A*. Cylinder *A* is secured to shaft *X* (note that *X* has screw thread on one end), which is supported by *P* and *O*. Support *P* also has screw thread on its inside. At one end, the upside of Figure 2, *X* enters tube *L*, which has an elongated slot in it. *L* can be rotated by device *M* and because *X* has a pin (*2*) secured to it that is sticking through the slot in *L*, *X* can also rotate. With the action of the screw thread on *P*'s inside, *X* will start moving towards *O*. Component *B* is a mouthpiece and *G* is its diaphragm, which has a hard pin-like indenting point at its centre poised exactly perpendicular to the surface of *A*. In its operating position, this point indents the recording material. On the other side, a similar assembly is placed: *C* is roughly the same sort of device as *B* but functions as a speaker. Its diaphragm *F* also has an indenting point fixed at its centre, but this one is held in place by a light spring *D* that presses it in the grooves made in the recording material on *A*'s surface.

[3] Those interested in their history and development, as well as that of other everyday artefacts, will enjoy reading Henry Petroski's books, for example Petroski (1992, 1996).

[4] The image is from the online US Patent Databases, http://www.uspto.gov.

[5] The explanation is based on Petroski's account of the history of the paperclip (Petroski, 1996).

[6] Available through http://www.uspto.gov, or http://edison.rutgers.edu/patents/00200521.PDF.

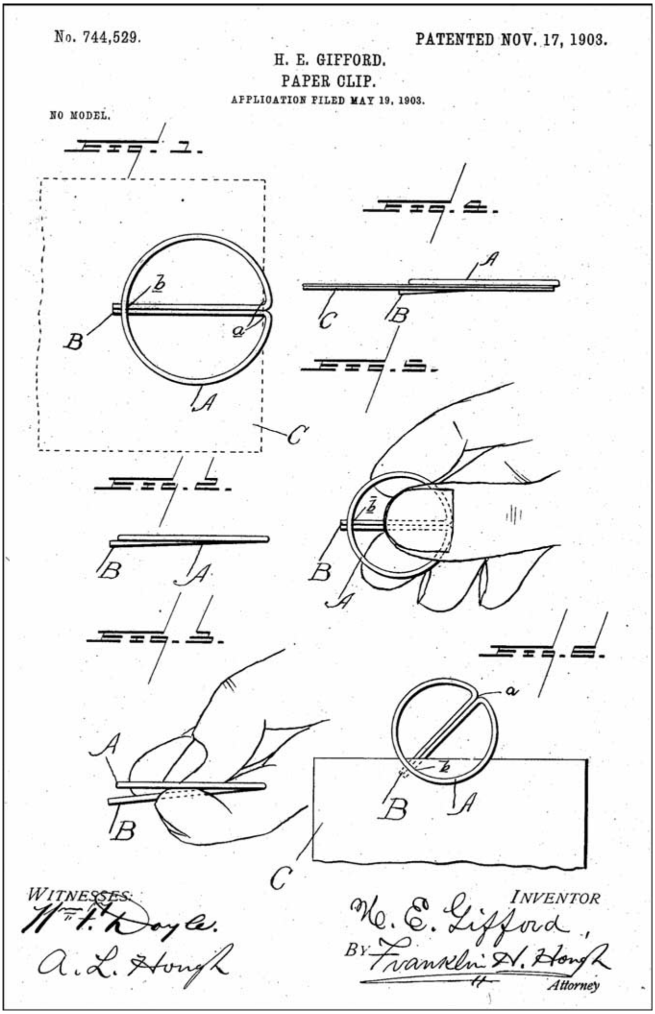

Fig. 1. H. E. Gifford's patented paperclip improvement.

The operation of the 'speaking machine' should now be fairly evident. If someone speaks in the mouthpiece *B* the air vibrations will cause the diaphragm to vibrate so that the indenting point at its centre will make indentations in the rotating recording material on *A*'s surface. As *X*'s rotation makes *A* move in the direction of *O*, the recorded track becomes spiral shaped. The track's pattern of indentations will represent voice movements and strength. Recording can continue until *A* reaches its utmost position at *O*. The recording material can then be removed from *A* and stored for later use, or it can be 'rewound' to its initial position to be played again. If *C* is placed in position, that is, its point placed exactly in a groove of the recorded material and held in place by spring *D*, and *M* makes *A* rotate again, the movements of *C*'s point will make diaphragm *F* vibrate according to the recorded pattern so that the resulting air vibrations replicate the original recorded voice.

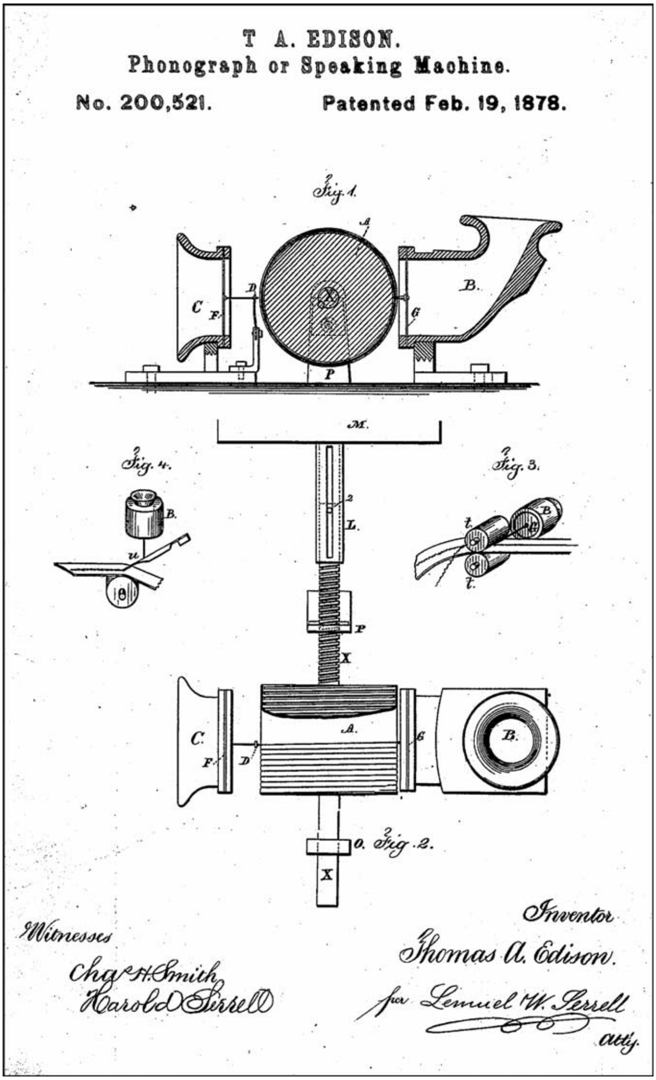

Fig. 2. Edison's original patented phonograph.

3. Two patterns of artefact explanation

I want to look at the examples in the opposite order of presentation for I believe that questions about the role of human actions are better appreciated in the light of a more general pattern of artefact explanation, which is provided by the second example. I pick up the pieces for an account of artefact explanation as I go along discussing my examples in this and the next section.

We should first become a little clearer on the explanandum. By asking: How does this thing work? we express an interest in a certain type of behaviour of the thing in question, typically the behaviour for which it was designed or is commonly used, the behaviour associated with its function. We want to know how it is that the thing has the potential to exhibit that behaviour and thus to fulfil its function. Edison's patent exemplifies such an explanandum; it demonstrates how the phonograph realizes the behaviour for which it was designed: 'to record in permanent characters the human voice and other sounds, from which characters such sounds may be reproduced and rendered audible again at a future time' (Eddison, 1878, p. 1). Note that this description is silent about the means by which the behaviour is to be accomplished. The explanandum is phrased purely in terms of behavioural roles;[7] it only describes the behaviour the phonograph fulfils without referring to the way in which, or the means by which, this behaviour is enabled. This is what the explanans should provide; some story about what is inside the behavioural black box, how the behavioural role is enabled by an underlying mechanism.

There is another thing to note about the explanandum. As several authors in this issue have noted (Franssen, 2006; Kroes & Meijers, 2006; Scheele, 2006; Vermaas & Houkes, 2006), ascriptions of technical functions are entangled in a normative discourse that does not apply to physical behaviour. It makes perfect sense to speak of an artefact's proper technical *function*, but not of its proper *physical behaviour*. Artefacts can exhibit numerous physical behaviours but from a physical point of view there is nothing proper about one particular behaviour as opposed to another. There is a gap of normativity between an artefact's proper function and its (actual) physical behaviour. As a result of this, the explanandum of a mechanistic artefact explanation can be interpreted in two different ways: (1) factually, that is, as being about the artefact's actual workings, or (2) normatively, that is, about how the artefact *ought to* work, if it is to fulfil its proper function. If the artefact is not broken or severely worn, and is used in a proper way in suitable circumstances, that is, does not malfunction, these two interpretations run parallel. Divergence occurs when the artefact malfunctions. I will return to this point in due course and show how the two readings lead to analogously different explanantia.

Now, for the explanans, let us look at the following two examples. First, when describing the mouthpiece Edison glances over the specific details of the device, but does write: '[it] may be of any desired character, so long as proper slots or holes are provided to re-enforce the hissing consonants' (Eddison, 1878, p. 1). Apparently, one of the behaviours of the mouthpiece is to reinforce hissing consonants and Edison says that any implementation will do fine, as long as it shows this behaviour. In effect, he places a behavioural constraint on any proposed mouthpiece implementation by stipulating its role. The second example has to do with cylinder *A*. Here, Edison specifies various implementation details: the number of grooves per inch on its surface, the helical pattern of the grooves, the indenting material on the surface, and *A*'s being secured to shaft *X*. Subsequently, he explicates how these structural features result in various behaviours, together enabling the recording behaviour, as discussed in Section 2. These two examples exemplify two *prima facie* different explanatory strategies; the former working its way 'downward' from the overall behaviour towards the components implementing the required sub-behaviours, and the latter working 'upwards' from the structural features of the components to the overall behaviour. We might say that the first strategy provides *behavioural* or *functional understanding* of an artefact, and the second

[7] Or, alternatively, in *functional* terms, but since the notion of technical function is used throughout this issue with normative connotations, as in 'proper function', I will refrain from using the term 'function' in a non-normative sense here.

structural understanding. Here is a more general characterization. Bechtel and Richardson (1993, p. 18) describe roughly the same strategies, but call them the analytic and synthetic strategy.

> **Top-down strategy**: take the behaviour to be explained and decompose it into more basic sub-behaviours, reiterate this step if possible, it should become clear how the complex behaviour being explained is realized by simpler behaviours in a specific spatiotemporal configuration, and for all the sub-behaviours, indicate which component(s) take(s) care of them.
>
> **Bottom-up strategy**: name the structural components of the artefact and give information about their physicochemical make-up and spatial configuration, show how their physicochemical features and configuration result in various behaviours and then describe how these behaviours, in their spatiotemporal configuration, together make up the behaviour to be explained.[8]

A couple of remarks about these two strategies are in place. First, the top-down strategy allows for decompositions of the overall behaviour that are still completely phrased in behavioural role terms, black-boxing implementation details. However, at some level the proposed implementation will nonetheless come into play implicitly since it dictates how a specific sub-behaviour must be decomposed. To see this, look at Figure 2 again. Whereas the big drawings on the patent show how to record sounds by indenting metallic foil, the smaller ones to the left and right exemplify other possible sub-behaviours for the recording function: moving a thread and impressing the shape of the thread on a roll of paper or depositing more or less ink on a roll of paper. The general picture emerging is that an overall behaviour is decomposed until the sub-behaviours necessitate a choice for one of possibly many functionally equivalent implementations. This choice then determines how the sub-behaviour in question is decomposed. This procedure repeats itself at lower levels of sub-behaviours until all the structural components have been ascribed their roles in the overall behavioural decomposition.

Secondly, both strategies mention decomposition, but not in quite the same sense. The first strategy is based on behavioural decomposition, the second on structural decomposition, that is, decomposition into physical parts. These two do not necessarily carve up the artefact into the same parts. One sub-behaviour can be implemented by more than one structural part and one structural part can play more than one behavioural role. If we think of the behavioural decomposition as producing a set of sub-behaviours B and the structural decomposition as producing a set of physical parts S, the relation between the two decompositions can be described as an *n:n* mapping between the elements of B and S, and if each sub-behaviour is implemented by a separate part, the mapping degenerates to a *1:1* case. Figure 3 provides a schematic example of this mapping; the b_is represent sub-behaviours and the s_is structural parts. The dotted lines specify the mapping between the two. Roughly speaking, the top-down strategy lays down the behavioural

[8] To anyone familiar with the literature on functional analysis, these two strategies may seem identical with Cummins's analytical and subsumption strategies (Cummins, 1975, pp. 758–760). This is not quite true, for (1) my strategies are both self-contained as explanations of artefact behaviour, while his are incomplete without one another and (2) it seems in principle possible to conduct both of Cummins's strategies wholly in functional, or behavioural, terms, whereas my bottom-up strategy intrinsically includes a description of the underlying physical implementation.

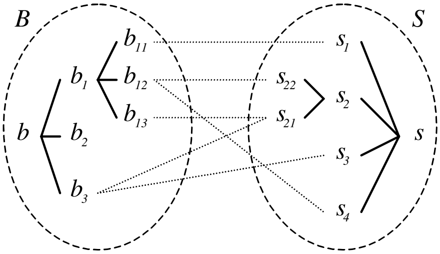

Fig. 3. Illustration of mapping between behavioural and structural decompositions.

decomposition and then points out where the structural parts fit in, while the bottom-up strategy first lays down the structural decomposition and then assigns behavioural parts their place therein. The strategies' starting points differ, but they both use both decompositions because both their end products include the mapping between the two decompositions. So although each explanatory strategy is self-contained, the two decompositions are complementary, in the sense that (1) the behavioural decomposition 'prescribes' which of the many behaviours of the structural parts are of interest and (2) the structural decomposition 'prescribes' to which of the physical parts the various sub-behaviours from the behavioural decomposition may be ascribed.

Thirdly, both strategies allow for multilevel explanations: behaviours can be broken up into sub-behaviours, parts can be broken up into parts of parts, and so on. A natural endpoint for the decomposition process, and for the explanation, seems to be the smallest physical components that make up the artefact (by which I mean the stuff one finds on assembly drawings of the artefact, not elementary particles) and the sub-behaviours they implement. It might be possible to continue explaining properties of elementary components, but then we leave the domain of engineering and enter that of natural science.

The obvious question now is how the two strategies are related. The first thing to note is that the top-down strategy does not seem to make much sense for the paperclip case. Paperclip behaviour is so basic that it is insusceptible to further decomposition. In a case like this the top-down strategy would amount to an assertion that this paperclip-shaped piece of metal exerts a force when its loops are bent out of position, which is hardly explanatory since this simply *is* paperclip behaviour and nothing is decomposed and thereby made more intelligible (I am ignoring the human action part of the explanation for now). At the same time, the bottom-up strategy makes perfect sense; the explanation of the paperclip I gave earlier embodies it. This suggests that the top-down strategy only conveys explanatory information for artefacts incorporating some degree of complexity, for example by having more than one behavioural and/or structural part.[9] For simple artefacts, the behavioural decomposition consists of only the overall behaviour or it is wholly determined by the specific implementation chosen so that it becomes impossible to provide a behavioural decomposition without at the same time describing the physical implementation. For example, the

[9] This shows immediately when we consider a slightly more complex 'species' of paperclips: the little two-legged clamps held together by a coiled spring. For such paperclips the top-down strategy makes more sense since this type of paperclip has a few components with different behavioural roles.

choice for a particular species of paperclip directly determines how its overall behaviour is decomposed, if at all, cf. note 9.

Secondly, the two strategies go together very well. Partly, they overlap in both producing the mapping between structural and behavioural decompositions, and for the other part, they are complementary. Roughly, the bottom-up strategy picks up where the top-down strategy stops. The output of the latter is an overview of all the sub-behaviours and the components that exhibit them. The former strategy then produces the structural, physicochemical, underpinning for the behaviours by showing which properties of the components enable them. Given this close cooperation, one might wonder if the two strategies should not simply be combined into one strategy. I think the answer is negative, for such a combined strategy would place too rigorous a requirement on artefact explanation. It seems incorrect to maintain that a good artefact explanation must always provide the structural underpinning delivered by the bottom-up strategy; it is often perfectly adequate to answer 'how does this work' questions by decomposing the behaviour and asserting that the components exhibit the identified sub-behaviours. In particular, the bottom-up strategy seems to be appropriate primarily in an engineering or design context. Engineers must know about the properties of the materials and components they use in design and they should be able to give an account of how these properties result in the relevant behaviours. In most everyday contexts however, people are not interested in the gory details, but only need a rough behavioural understanding of how a complex behaviour is produced.[10] Behavioural understanding is also particularly important in the early stages of engineering design, when engineers considers various possible conceptual design, without (yet) worrying about technical implementation details.

4. What about human action?

So far, I have said nothing about the role of human action in artefact explanation. The paperclip example can shed more light on this role. Let me first draw a sketchy picture of artefact use in general (cf. Houkes, Vermaas, Dorst, & De Vries, 2002). People use artefacts to accomplish goals and they do so by manipulating them according to more or less standardized use plans that involve sequences of actions. By doing something with an artefact, a change in the obtaining state-of-affairs is brought about that moves the artefact user closer to a desired state-of-affairs. In this picture, we need to find out how it is that actions with artefacts contribute to realizing a desired state-of-affairs.

In the paperclip case, the action involved is the slight bending of the loops by pressing the longest loop against the papers and then sliding the clip on. Although humans typically perform this action, it is beyond question that the paperclip would still work were a robot or zombie to perform the same action. This shows that the *intentionality* of the actions performed with artefacts does not really matter for our understanding of how an artefact works. The only thing that matters is the physical movements going on, regardless of whether humans, zombies, or robots are responsible for bringing these movements about. A distinction from action theory between an action *qua* intentional action and an action *qua* physical event can clarify this point. When I beckon someone to come over, my hand moves along a trajectory. This physical movement of my hand is the action *qua* physical

[10] Vermaas & Houkes (2006) capture an analogous point with their notion of cloaking.

event, or *intrinsic event* associated with my action of beckoning (Ruben, 2003, p. 44). To the extent that human actions are involved in artefact use, the intrinsic events associated with these actions alone suffice to explain the artefact's operation. Obviously, intrinsic events do not suffice to explain why an artefact was used or designed in the first place, but those are different explanatory tasks. To understand the physical operations of the artefact, 'how it works', we need only see what sort of physical processes are going on. Although it is surely of practical value to know which of these processes are typically brought about by human actions, it does not matter for the artefact's proper operation if in fact they are brought about by a robot or a skilful monkey. The intrinsic event is doing the explanatory work here and not the intentional aspect of the action.

However, there is more to say about how actions and physical artefact behaviour go together in an artefact explanation. Both explanatory strategies from the previous section rely on the, so far unspecified, notion of behaviour. Saying that an object can be made to behave in certain ways is to claim that it possesses certain physical *capacities*, namely those that give rise to the behaviours in question. If a doorbell rings when someone presses it, it must have the capacity to ring when pressed. Sugar dissolves in water because it has the capacity to do so. Objects have capacities in virtue of their physical make-up, that is, the configuration and constituent materials of their components and ultimately the properties of their constituent materials; this is what the bottom-up strategy exposes. Because a paperclip is made of bent metal wire, it has a certain springiness, which gives it its paper-clipping capacity. Capacities can be expressed by conditional statements: if an object x has the capacity to F it must be true that x Fs if G-ed, provided appropriate background conditions obtain and no counteracting influences disturb the process. The first proviso is indispensable because behaviours may only occur under more or less special circumstances; the second because of the always present possibility of pre-empting influences, cf. Mumford (1996, pp. 90–91) for the two sorts of conditions.[11]

With this conditional analysis, we can see clearly what role human actions—or, more precisely, the intrinsic events associated with human actions—play in an artefact explanation. They provide the antecedent events on which manifestation of certain capacities is conditional. Human actions 'stimulate' the artefact in appropriate ways for it to respond by exhibiting its behavioural capacities. I bend the loops of the paperclip out of position and it exerts a force in the opposite direction that holds a set of papers together. Somebody speaks into the mouthpiece of Edison's phonograph, the diaphragm starts to vibrate, and the indenting point indents the recording material. By manipulating an artefact you trigger one of its (sub-)capacities to occur; the intrinsic event associated with your action provides G, which makes the artefact (or one of its components) F.

Let us take stock. What is the ensuing picture of artefact explanation? An explanation of how an artefact works according to the top-down strategy provides a collection of sub-behaviours at various levels, eventually all ascribed to particular structural components of the artefact, and their interactions, that is, how some behaviours provide antecedent or background conditions for others. For some behaviours, there will typically be reference to at least one intrinsic event triggering those behaviours. This explanation can be comple-

[11] I am aware that this paragraph passes over some tantalizingly difficult issues as to the nature of dispositional and categorical properties, capacities, conditionals, and their mutual relations, but a discussion of these issues would lead us too far astray. Therefore I must ask the reader to bear with me and accept this common sense picture of capacities and properties.

mented with another one, bottom-up style. This second explanation goes into the physicochemical properties and configuration of the components, shows how the properties and configuration constitute capacities entailing various behaviours, and then clarifies how the behaviours interact to make up the overall behaviour that needs explanation. Again, in the description of how some behaviours are triggered, intrinsic events will crop up.

Before we broaden our scope to mechanistic explanation in general in the next section, I should return to the normativity of function ascriptions. If an artefact does not malfunction, an explanation along the lines I have sketched should be perfectly satisfactory, but what happens when we want to know how a malfunctioning artefact works or is supposed to work? In other words: What happens when the explanandum is interpreted normatively? As in: How is this thing supposed to work? In such cases, I think the explanans 'inherits' the normative force of the original proper function ascription, that is, the explanans is no longer factually correct, but describes how the artefact *ought to* behave, top-down style, or which properties and capacities it *ought to* have, bottom-up style, given its proper function. So while the pattern of explanation does not change, the explanation should now be read as a normative account of how the artefact ought to behave and which properties and capacities it ought to have. The normative force of this 'ought' derives from the normativity of the original function ascription. To the extent that this original function ascription is justified, the ascriptions of behavioural roles and physical capacities to the artefact components in the explanans will also be justified. Of course, as with function ascriptions to complete artefacts, there are limits on what can justifiably be ascribed to artefact components in an artefact explanation. These limits have to do with our scientific, experiential, or other knowledge of the physicochemical make-up of the components and social practices and institutions. Other contributions to this issue discuss such constraints on function ascriptions (e.g. Scheele, 2006; Vermaas & Houkes, 2006). If, however, we want to know why an artefact malfunctions, the normative 'ought to' explanans is still useful for pinpointing where and how malfunction occurs. Since the behavioural and structural decompositions describe the behavioural roles of the components and their interactions, they also provide a host of information about malfunction—what happens if some component fails to behave as it should. Such information would certainly be helpful for the notoriously difficult task of diagnosing specific instances of malfunction.

5. From artefact explanation to mechanistic explanation

How does this account of artefact explanation relate to other accounts of mechanistic explanation? Recent debate on mechanistic explanation has fleshed out the rudimentary comments by early writers on the topic, but I think my two mechanistic explanatory strategies bring out some shortcomings in other work. I will briefly introduce the mechanicists' project and then show how it is hampered by a failure to recognize the difference between structural and functional understanding.

Mechanicists want to develop the familiar intuition that causal explanation is all about showing how underlying mechanisms and structures produce the event to be explained. The intellectual fathers of this idea, Railton (1978) and Salmon (1984), cashed out the notion of mechanism purely in terms of causality and causal interactions. The 'new mechanicists' think this conception of mechanisms is too narrow; there is more to mechanisms than just causality. Inspired by Simon's (1996) and Wimsatt's (1976) work, they start from the idea that mechanisms are complex systems and then try to develop that notion into the basis for an account of explanation. This project has more supporters than I can discuss here (e.g. Bechtel & Richardson, 1993; Glennan, 1996, 2002; Machamer et al., 2000; Craver, 2001), so I will limit myself to two recent exponents that give a fair impression of the general idea.

A good place to start is Glennan's definition of a mechanism:

> A mechanism for a behaviour is a complex system that produces that behaviour by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations. (Glennan, 2002, p. S344)

This definition intends to capture a number of important aspects of mechanisms. First, that it only makes sense to talk about mechanisms *for behaviours*, not mechanisms *simpliciter*. Mechanisms are relativized to the behaviours they produce. Secondly, mechanisms consist of material parts or objects; they are not just sequences of processes. Thirdly, mechanisms produce their behaviour by the causal interaction of their physical parts. The interactions must be invariant in Woodward's sense (Woodward, 2003), meaning roughly that they should support a relevant range of counterfactuals. Glennan puts his definition to work in an account of mechanistic explanation. To explain a regularity is to give a *mechanical model*:

> a description of a mechanism, including (i) a description of the mechanism's behaviour; and (ii) a description of the mechanism which accounts for that behaviour. (Glennan, 2002, p. S347)

This model is the explanation: (i) is the explanandum and by virtue of showing why (i) is correctly ascribed to the mechanism in question, (ii) forms the explanans. A particular behaviour that needs explanation is explained by a description of the mechanism for that behaviour, which includes a description of how the parts of the mechanism, triggered by an outside event, interact sequentially to produce the behaviour being explained.

Craver (2001) describes how mechanisms of the sort described above enable an account of *multilevel mechanistic explanation*. For behaviour *b* of an item, we can give (1) a contextual description of that behaviour, (2) an isolated description, and (3) a constitutive description. The three descriptions correspond to different perspectives; (1) places the behaviour in a larger system with its own behaviour to which *b* contributes, (2) is a description of *b* in isolation, and (3) shows how *b* is brought about by sub-behaviours of parts of the item. By employing these perspectives, science uncovers hierarchical mechanisms and behavioural roles played by the components of these mechanisms. Craver claims that such role ascriptions to items are important scientific achievements since they allow us to assign items their proper place in the multi-layered picture of how our world works.

I now want to present two related criticisms on these accounts that derive directly from what I have said above about mechanistic explanation. My complaints do not primarily concern what the mechanicists do say, but rather what they do not say—it seems to me that they rely too heavily on hand waving when they talk about describing mechanisms and their behaviours.

First, they do not distinguish between the types of understanding conferred by my top-down and bottom-up strategy. This seems largely due to a failure to analyze in any detail different modes of describing mechanisms and their behaviours. For instance, Glennan's characterization of a mechanical model (Glennan, 2002, p. S347, see above) does not stipulate whether the sort of information in a description of a mechanism is to be given in behavioural terms or also in terms of physical parts and their configuration and physicochemical properties. As I have argued in Section 3, those two options lead to either behavioural or structural insight into a mechanism's operations. Craver (2001) does take a stand on the issue; he is clearly talking about top-down behavioural role explanations. As a result his analysis fails to take account of the fact that explanation exclusively in terms of hierarchies of behaviour is impossible in the sense that details about the underlying implementation always permeate a supposedly purely behavioural explanation. First, in determining how particular sub-behaviours are to be decomposed. Because it is usually possible to have more than one functionally equivalent implementation, the choice for a particular implementation dictates how the behavioural role gets decomposed. As an example, think of how Edison's decomposition of the recording behaviour would have looked if he had chosen to use one of the other two devices depicted in his patent. Secondly, implementation details sometimes provide a more plausible rendering of Craver's third, constitutive, mode of description. Namely, in cases where we want to explain 'basic' behaviours. In such cases, explanations mention physicochemical properties in conjunction with the spatial configuration of the mechanism components to explain the behavioural capacities of the components.[12] The elasticity of metal explains the paperclip's capacity to exert a force when its loops are bent out of position, as opposed to some story about lower level behavioural roles of stuff, as Craver's constitutive description has it. By overlooking the role of implementation details in mechanistic explanation, Craver also overlooks the distinction between the two different strategies of mechanistic explanation.

A second conspicuous omission in the accounts of mechanistic explanation under discussion is the role and character of bottom-up mechanistic explanation of behaviour. None of the authors I have been discussing devotes any attention to how the behaviours they take to be part of a mechanism are often explained by capacities of the constituent components of the mechanism, which are in turn explained by underlying structural, that is, physicochemical, and, for a fully general account, perhaps also biological, psychological, social, etc., properties. For instance, Craver states that all there is to learn about a mechanism is the spatial and temporal organization of the entities and activities in it:

> One understands a mechanism by discovering its component entities and activities, and by learning how their activities are spatially and temporally organized ... . Understanding how a mechanism works is just understanding how one activity leads to the next through the spatial layout of the components and through their participation in a stereotyped temporal pattern of activities from beginning to end. (Craver, 2001, pp. 60–61)

[12] It may be objected that such explanations must ultimately be couched in terms of behavioural roles of molecules, atoms, and elementary particles. While this might be a conceivable possibility, it is certainly not beyond doubt that it is in fact feasible. Neither is it clear that this is an adequate interpretation of what low-level physical explanation really does. It would be rather imprudent to have such issues settled beforehand by stipulations deriving from our favoured account of mechanistic explanation.

Taken at face value, this just strikes me as highly implausible. It is certainly true that knowledge of the spatial and temporal organization of mechanism parts is important—my strategies also include such information—but even with this knowledge it makes perfect sense to ask how these spatially and temporally organized behaviours are made possible by the structural characteristics of the objects or materials that make up the mechanism. An answer to this question seems to me a legitimate part of a mechanistic explanation, which adds to our understanding of the mechanism in question.

Hence, in my view, existing work on mechanistic explanation lacks an appropriate sensitivity to the distinction between a top-down and bottom-up strategy of mechanistic explanation. It could also benefit from paying more attention to how mechanism behaviours can be explained by the capacities and structural properties of components, as opposed to just their spatiotemporal relations. I do not believe that my own account comes near to accomplishing this task in its full generality, but I do hope to have demonstrated that it grasps some important features of mechanistic explanation and that it does so better than the other accounts I discussed.

6. Conclusion

To sum up, my account of mechanistic artefact explanation includes two different complementary explanatory strategies. The top-down strategy takes the overall behaviour of the artefact that needs explanation, decomposes it into constituent sub-behaviours, and then ascribes these sub-behaviours to structural components of the artefact. It also describes how the various behaviours interact spatiotemporally to make up the overall behaviour. Human action enters the explanation as intrinsic events, which provide the antecedent conditions for particular sub-behaviour manifestations. This strategy 'black-boxes' details about the underlying physical structures. The second, bottom-up, strategy starts from physicochemical and spatial information about the artefact's components and then shows how these structural features ground behavioural capacities. By discussing how the behaviours resulting from the capacities interact in space and time, it demonstrates how the overall behaviour is brought about. Human action enters this story in the same way as it did in the first strategy. Either strategy provides self-contained explanations and is not intrinsically in need of the other, but the two do naturally favour different contexts; structural understanding as provided by the bottom-up strategy is especially appropriate in engineering contexts, and the top-down strategy's behavioural understanding generally suffices for everyday contexts, but is also important in early stages of engineering design.

It seems to me that my two strategies are equally well applicable to mechanistic explanation outside the domain of artefacts and technology. Nothing makes them exclusively geared towards artefacts as opposed to biological, psychological, sociological, economic, or other mechanisms.

I want to conclude with a more general and admittedly imprecise thought that is triggered by my account of mechanistic artefact explanation. If the point made above about top-down behavioural understanding being self-contained is correct and if mechanistic explanation is, as its original proponents certainly seem to have had in mind, a ubiquitous style of scientific explanation, it follows that it can be perfectly acceptable to explain phenomena by describing behavioural roles of constituent mechanism parts without saying anything about how these parts implement their behavioural roles. Hence, there is nothing suspect about high-level explanation that leaves out details about underlying implementation and there is no intrinsic need for explanation to always seek as low a level of causal detail as possible. I propose to call this feature of top-down mechanistic explanation *realization-independency*. We can explain and learn about certain high-level behaviours even when we are largely ignorant of the underlying physical reality. In fact, the truth about the underlying details simply does not matter for the quality of the top-down explanation. They may be metaphysically abstruse beyond any philosopher's wildest fantasies; as long as they satisfy the behavioural-role ascriptions at the higher level the explanation is still good. This is a fascinating result since it implies that any account of explanation which holds that more information about underlying details is necessarily better is misguided.[13] Such a view is not only held by straw men. For instance, Jon Elster subscribes to it when he claims: '[A] more detailed explanation is also an end in itself' (Elster, 1985, p. 5). Michael Taylor seems to be thinking along the same lines: 'A good explanation should be, amongst other things, as *fine-grained* as possible' (Taylor, 1988, p. 96). Railton's (1978) and Salmon's (1984) accounts of causal explanation also seem to be sympathetic to the pro-detail attitude. Against such views, my top-down strategy suggests that there is a legitimate explanatory role for information that does not concern finer grain or more detail. My view here is similar to what other philosophers have claimed about, for instance, explanation in the social sciences (Jackson & Pettit, 1992) or the special sciences in general (Fodor, 1974).

Structural and behavioural understanding of an artefact, or other mechanism, is useful from an explanatory point of view. Structural understanding shows one exactly how a particular mechanism *implements* a piece of behaviour. In an engineering context, such understanding of artefacts is highly useful because it differentiates between the detailed implementation choices to be made in figuring out the details of a new design. Behavioural understanding, however, exhibits how a particular complex behaviour can be created out of simpler behaviours, independent of the particular realization of these behaviours. For artefacts, such understanding is valuable in that it differentiates between possible ways of creating a complex behaviour out of simpler ones. Engineers need this understanding to think of conceptually different ways to create a particular complex behaviour, and 'lay people' may find it useful to get a broad grasp of how an apparently sophisticated piece of behaviour is created out of simpler sub-behaviours, without having to worry about how it is exactly that these sub-behaviours are made possible by the physics of the artefact. These two distinct types of understanding and explanatory strategies are both valuable contributions to our knowledge of technical artefacts or of other mechanisms.

Acknowledgements

Research for this paper formed part of the program The Dual Nature of Technical Artefacts which is supported by the Netherlands Organization for Scientific Research (NWO). My thanks are due to Michael Strevens, Pieter Vermaas, Wybo Houkes, Marcel Scheele, Peter Kroes, and Anthonie Meijers for helpful comments on earlier drafts.

[13] The result also casts doubt on the desirability and usefulness of reductionist projects, but since they seem to have fallen out of grace lately, this is perhaps not very telling.

References

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, *68*(1), 53–74.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, *72*(20), 741–765.

Eddison, T. A. (1878). Phonograph or speaking machine. United States Patent 200,521. 19 February. (Available at http://www.uspto.gov and http://edison.rutgers.edu/patents/00200521.PDF)

Elster, J. (1985). *Making sense of Marx*. Cambridge: Cambridge University Press.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*, 97–115.

Franssen, M. (2006). The normativity of artefacts. *Studies in History and Philosophy of Science*, *37*, this issue. DOI:10.1016/j.shpsa.2005.12.006.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.

Glennan, S. S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*, S342–S353.

Houkes, W. N., Vermaas, P. E., Dorst, K., & De Vries, M. J. (2002). Design and use as plans: An action-theoretical account. *Design Studies*, *23*(3), 303–320.

Jackson, F., & Pettit, P. (1992). Structural explanation in social theory. In D. Charles, & K. Lennon (Eds.), *Reduction, explanation, and realism* (pp. 97–131). Oxford: Clarendon Press.

Kroes, P. A., & Meijers, A. W. M. (2006). Introduction: The dual nature of technical artefacts. *Studies in History and Philosophy of Science*, *37*, this issue. DOI:10.1016/j.shpsa.2005.12.001.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Mumford, S. (1996). Conditionals, functional essences, and Martin on dispositions. *The Philosophical Quarterly*, *46*(182), 86–92.

Petroski, H. (1992). *The evolution of useful things*. New York: Alfred A. Knopf.

Petroski, H. (1996). *Invention by design*. Cambridge, MA: Harvard University Press.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science*, *45*(2), 206–226.

Ruben, D.-H. (2003). *Action and its explanation*. Oxford: Oxford University Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Scheele, M. (2006). Function and use of technical artefacts: Social conditions of function ascription. *Studies in History and Philosophy of Science*, *37*, this issue. DOI:10.1016/j.shpsa.2005.12.004.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Taylor, M. (1988). Rationality and collective action. In M. Taylor (Ed.), *Rationality and revolution* (pp. 63–97). Cambridge: Cambridge University Press.

Vermaas, P. E., & Houkes, W. N. (2006). Technical functions: A drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science*, *37*, this issue. DOI:10.1016/j.shpsa.2005.12.002.

Wimsatt, W. C. (1976). Reductionism, levels of organization, and the mind–body problem. In G. G. Globus, G. Maxwell, & I. Savodnik (Eds.), *Consciousness and the brain: A scientific and philosophical inquiry* (pp. 205–267). New York: Plenum.

Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.