1 Mechanisms and Their Discovery

1.1 Mechanisms and Mechanistic Explanations

Many explanations in biology tell us how phenomena are produced as a result of changes and interactions among the various parts of a physical system. Such systems of parts changing over time are commonly known as 'mechanisms', while the more-or-less detailed descriptions of these systems and the series of changes they undergo are referred to as 'mechanistic explanations'.

If the reader is left wondering what a phenomenon is and how a mechanism explains it, a concrete example may help. Biologists take reproducible sequences of events, whereby exposure of a biological system to a stimulus is consistently followed by a similar kind of response, to be phenomena in need of an explanation (Figure 1). Take sunburns, for example. Prolonged exposure to ultraviolet radiation consistently results in a pronounced redness and swelling of the skin. Since this sequence of events can be consistently reproduced in humans and many other animals, it is unlikely to be just a coincidence. It is something that demands an explanation.

Sunburns belong to a family of related phenomena of great clinical importance known as inflammatory responses. These responses are triggered by a wide variety of harmful stimuli, including radiation, tissue damage, pathogens, and antigens, and typically consist of five directly observable symptoms: erythema (redness), edema (swelling), pain, heat, and loss of function (e.g., immobility). In most cases, inflammatory responses shut down after the threat has been eliminated or the harmful stimulation has subsided, which is highly desirable since a prolonged or chronic inflammatory response has detrimental consequences for the organism.

Much research in immunology and medicine has been devoted to finding out how – that is, by means of which mechanisms – inflammatory responses are produced. Since my goal is simply to give the reader an idea of what a mechanistic explanation may look like, I will focus on one of the better understood mechanisms involving a type of white blood cells known as T-cells. By the end of the 1980s, it had been established that, when an organism is exposed to harmful stimuli, T-cells begin secreting a variety of molecules necessary for mounting inflammatory responses. Thus, one key research question was to find out how T-cells are activated. The molecular underpinnings of T-cell activation turned out to revolve around a couple of regulatory DNA-binding proteins, one of which is the transcriptional factor nuclear factor κB, or NF-κB in short (Sun et al. 1993). In resting T-cells, NF-κB is trapped in the cytoplasm by a protein known as inhibitor of NF-κB, or IκB. When cells are exposed to harmful stimuli (Figure 2A), however, a chain of protein–protein interactions

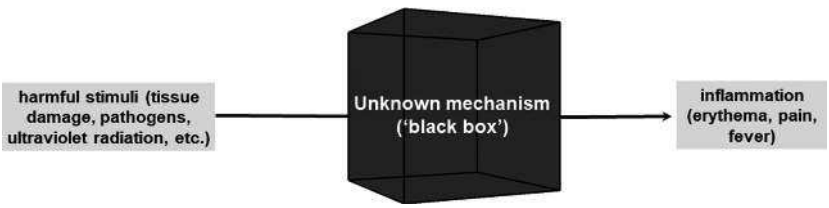

Figure 1 Inflammatory responses as phenomena to be explained

leads to the dissociation of inactive IκB/NF-κB complexes (B). NF-κB is freed (C) and can now move to the nucleus, where it binds specific DNA sequences and enhances the expression of a number of genes, many involved in immune responses (D). This results in a variety of new proteins being manufactured in the cell, such as the cyclooxygenase-2 (COX-2) enzyme, which catalyzes the synthesis of prostaglandins. Once secreted in the bloodstream, prostaglandins promote vasodilation, which is responsible for erythema and edema (redness and swelling of the skin); sensitize spinal neurons to pain; and act on the thermoregulatory center of the hypothalamus to produce fever (E).

We can now form a general idea about how inflammatory responses are initiated when organisms are exposed to harmful stimuli. What is missing is an explanation of how the responses shut down. It was eventually discovered that NF-κB also enhances the production of its own inhibitor, IκB. The newly synthesized IκB binds NF-κB, trapping it back in the cytoplasm (F). Prostaglandins too curb NF-κB activity by interfering with the signaling pathway leading to the degradation of the inhibitory protein IκB (G). Thus, following stimulation, T-cells are initially activated, resulting in an inflammatory response, then are automatically turned off by means of a negative feedback loop molecular mechanism, which performs a function analogous to that of a common thermostat.

The above explanation is in many respects incomplete. For instance, it is not specified how prostaglandins cause the symptoms of inflammation. Quantitative-dynamic details are also missing. The explanation doesn't say anything about the duration of the inflammatory response or whether cyclical inflammatory responses are generated in response to persistent stimulation. Finally, some familiarity with chemistry is tacitly assumed, such as the notion that protein–protein and protein–DNA interactions are weaker forms of binding in comparison to the covalent bonds holding molecules together. Nevertheless, despite assumed, abstracted, and unknown details, the above narrative and the accompanying diagram should help the reader imagine how inflammatory responses are produced as a result of a sequence of changes and interactions involving various cellular and molecular components of an organism.

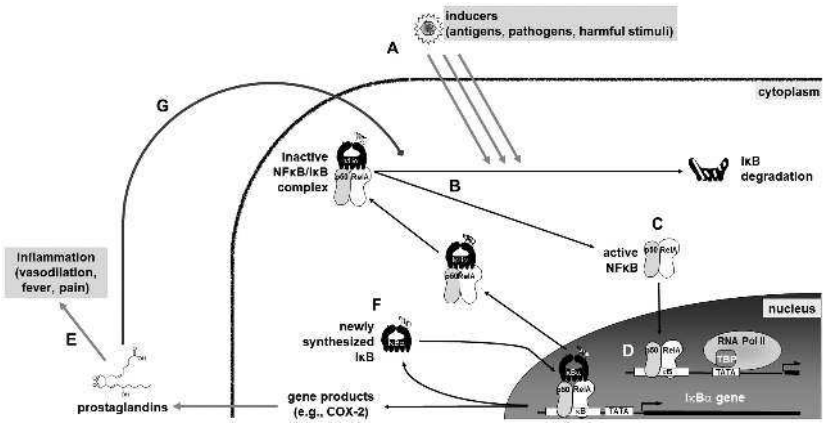

Figure 2 A simplified representation of the NF-κB mechanism and its role in regulating inflammatory responses

1.2 The Discovery of Biological Mechanisms

The elucidation of biological mechanisms often begins with the formulation of hypotheses sketching out possible mechanisms.[1] For instance, experimental results indicating that protein synthesis inhibitors block T-cell activation led some researchers to hypothesize that inflammatory responses rely on a mechanism of genome expression regulation. Initial speculative work is followed by what Lindley Darden (2006b, Ch. 4) describes as a gradual filling in of missing mechanistic details. This usually requires a significant amount of experimental research.

In a first step, a system responsible for producing the phenomenon of interest, or what William Bechtel and Robert Richardson (2010, Ch. 3) call the 'locus of control', is identified. In practice, this amounts to the characterization of an experimental setup in the context of which a phenomenon can be consistently reproduced. For instance, the experimental setup used to elucidate the NF-κB regulatory mechanism consisted mainly of a cell model of T-cell activation. Normal T-cells extracted from the blood of a healthy donor or precancerous ('immortalized') T-cell lines were grown in an artificial medium and stimulated

[1] Peter Machamer, Lindley Darden, and Carl Craver (2000, 18) define a mechanism sketch as "an abstraction for which bottom out entities and activities cannot (yet) be supplied or which contains gaps in its stages. The productive continuity from one stage to the next has missing pieces, black boxes, which we do not yet know how to fill in. A sketch thus serves to indicate what further work needs to be done in order to have a mechanism schema." In contrast, a more detailed mechanism schema, such as the one depicted in Figure 2, is a "truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities . . . When instantiated, mechanism schemata yield mechanistic explanations of the phenomenon that the mechanism produces" (2000, 15, 17).

by the addition of a chemical inducer, such as lipopolysaccharides, a toxin released when the cell walls of certain bacteria are destroyed, and which causes septic shock under physiological conditions.

In a second step, variables describing the experimental setup are targeted by experimental interventions in the hope of demonstrating that specific changes in the experimental setup and the physical systems of which it is composed result in changes in the phenomenon under investigation. For example, the IκB inhibitor was shown to be part of the regulatory mechanism of T-cell activity based on experimental evidence demonstrating that mutations in the sequence of IκB result in a prolonged activation of T-cells following stimulation. Carl Craver (2007, Chs. 2–3) gives an excellent analysis of the role that causal relevance plays in the identification of putative mechanistic components. In particular, he points out that causal relevance is demonstrated by means of controlled experiments, thus establishing a connection between the scientific practices involved in the elucidation of mechanisms and James Woodward's (2003) interventionist account of causation.

Finally, a mechanism must eventually be 'recomposed' in order to show how it generates the phenomenon (Bechtel 2011). A mechanism's recomposition can be physical, for instance, an in vitro reconstitution experiment, or conceptual, such as a computer simulation or simply a narrative or diagram such as those illustrated earlier. One way or another, the goal is to demonstrate that components organized, acting, and having the properties described in the mechanistic explanation can produce, and ideally are sufficient to produce the phenomenon under investigation.

1.3 Experimental Methodology First

In scientific practice, hypothesis and experimentation proceed in tandem. The initial characterization of an experimental setup and the choice of variables targeted for intervention are motivated by existing hypotheses about the mechanism responsible for a phenomenon. Conversely, experimental results, such as the ability or failure to consistently reproduce the phenomenon, as well as the fortuitous discovery of causally relevant factors, play a key role in the formulation of new mechanistic hypotheses.

This Element, however, focuses almost entirely on experimental inquiry.[2] There are two reasons for this unusual choice. The first has to do with the fact that most philosophers approach experimental practice from the perspective of

[2] Perhaps a clarification is needed here. What I mean by 'experimental' or 'empirical inquiry' is experimental research in science. Experimentalists approach hypothetical explanations with an attitude of 'suspended belief' until adequate supporting evidence is produced. Being cautious, however, is not the same as being skeptical. Experimental research is a sustained effort to produce

famous discovery episodes that have left a profound mark on the development of science. Major discoveries are brought under philosophical scrutiny in an attempt to reconstruct the reasoning involved in the formulation of particularly fruitful hypotheses and the design and execution of famous experiments testing such hypotheses. Lindley Darden (1991; 2006b), William Bechtel (2006; 2008), Carl Craver (2007), and Marcel Weber (2005) analyze specific cases from biochemistry, genetics, cell biology, molecular biology, electrophysiology, and neuroscience, examining in detail some of the most famous discoveries of mechanisms in biology. Nevertheless, science can also be approached from the less glamorous, but equally relevant perspective of the basic methodological principles governing everyday 'normal science' research, as Kuhn would put it. The latter crystallized over the past four centuries into an autonomous set of practices governing experimental design independently of any specific explanatory hypotheses. Thus, one rationale for approaching mechanistic discovery from the standpoint of experimental methodology is that it has the potential of providing a novel and, I think, more generally applicable perspective on mechanisms and their discovery.

The second rationale is intimately linked to the immediate goal of this Element, which is to elaborate a metaphysical account of mechanisms. A number of important publications have recently brought mechanistic metaphysics under systematic scrutiny, most notably Stuart Glennan's (2017) *The New Mechanical Philosophy*. Nevertheless, these works also reveal a disconnect between two lines of philosophical inquiry. One deals with epistemic issues related to the discovery of biological mechanisms and the intelligibility of mechanistic representations. The other relates to a metaphysical inquiry into how mechanisms relate to issues such as ontology, causation, laws, and levels of organization. The main object of interest here is no longer discovery, but rather the final product of scientific research, namely the narratives and diagrammatic representations typically found in biology textbooks, which are analyzed in order to gain insights into what these epistemic products presuppose and entail from a metaphysical point of view. Despite the individual contributions of each project, the two remain disconnected inasmuch as it is not clear how the relatively modest and down-to-earth discovery strategies identified by the epistemic project can justify the more audacious claims associated with the metaphysical project.

One way to bridge the gap between the two projects is to stipulate that a metaphysical account of biological mechanisms should remain compatible

evidence for or against explanatory accounts, which is a clear indication that empirical inquiry in science doesn't condone the defeatist attitude typically associated with philosophical empiricism.

with the ontological assumptions of the experimental methodology employed in the elucidation of mechanisms. I think that the immense success of experimental research in biology – and it must be emphasized that biology is an experimental discipline to a much larger extent than physics and chemistry – justifies this demand. I argue therefore that two kinds of considerations should constrain mechanistic metaphysics. The more fundamental ones, amounting to what I call the 'minimal experimental interpretation', are methodological in nature. Two-thirds of this Element is devoted to these considerations. I begin my inquiry with an attempt to define the notion of 'phenomenon', which, surprisingly, received little attention in the mechanistic literature. In Section 2, I defend the view that standard experimental methodology assumes a causal interpretation of measurements in virtue of which causes responsible for differences in measured values of variables can be localized within the spatiotemporal boundaries of physical systems satisfying a given experimental description. This interpretation allows for a definition of phenomena in strictly methodological terms, as data reproduced when experiments are replicated.

An account of phenomena further determines how one construes the relationship between mechanisms and phenomena. In Section 3, I argue that neither of the two main philosophical accounts, the etiological and the part–whole constitutive accounts, is compatible with the demands of experimental research. I reject both accounts in favor of an alternative one according to which the mechanism responsible for a phenomenon is a causal structure that does not allow the variables probed by the measurements involved in the description of the phenomenon to vary independently of one another.

The foundational work conducted in Sections 2 and 3 is meant to provide a minimal framework onto which richer mechanistic ontologies, of the sort typically endorsed by scientists and philosophers, may be grafted. This brings us to the second kind of considerations constraining mechanistic metaphysics, namely those linked to a richer and more diverse set of physical interpretations relying on background knowledge about the physics and chemistry of biological systems. A considerable body of philosophical work on mechanisms targets, directly or indirectly, this kind of considerations. In the philosophical literature, mechanisms are systematically characterized as physical systems composed of spatiotemporally organized parts acting, interacting, or functioning in such a way as to produce, maintain, underlie, or constitute phenomena (Bechtel and Abrahamsen 2005; Glennan 1996; 2002; 2017, Ch. 2; Illari and Williamson 2012; Machamer et al. 2000). These characterizations depict mechanisms as consisting of physical entities acting or playing certain functional roles, or again as parts interacting in accordance to the laws of physics. In Section 4, I review

the main interpretations circulating in the philosophical literature, along with their theoretical and experimental justifications. An exhaustive treatment of these interpretations is beyond the scope of this short Element. The main goals here are to make explicit the ontological commitments associated with each interpretation and to assess their compatibility with the minimal experimental interpretation developed in previous sections. My hope is that the analysis conducted in this section will enable the reader to gain a clearer understanding of how tacit and explicit metaphysical assumptions underpinning philosophical accounts of biological mechanisms relate to scientific theories and experimental practice. Finally, a recapitulation of the main theses defended, including some last-minute clarifications, is provided in the concluding section.

2 What Is a Phenomenon?

A mechanism is invariably characterized as a 'mechanism for a phenomenon'. But what is a phenomenon? Scientists describe phenomena as being both constitutive of empirical reality and objects of explanation, hinting to an overlap between these notions. Logical empiricists took this description at face value, equating phenomena, *explananda*, and observations. Some observations are unaided, and some are mediated by instruments of measurement, but they ultimately all amount to subjective perceptual experiences reported in the medium of language. My task in this section is to introduce an alternative view according to which empirical research in contemporary biology is primarily a matter of conducting controlled experiments in order to generate data structured in such a way as to make possible inferences about the causal structure of the world. Then I argue that, thus understood, data are phenomena to be explained.

2.1 Data

In order to understand what data are and how they relate to phenomena, it is useful to consider once again the example of inflammatory responses. As stated in the previous section, these responses are biological outcomes characterized by the occurrence of five symptoms – erythema, edema, pain, heat, and loss of function – following harmful stimulation, such as tissue damage or exposure to ultraviolet radiation. Based on this example, a phenomenon may be viewed as a 'black box' causal process initiated by a set of stimuli and terminating in a set of symptoms (Figure 1).

Each of the inputs and outputs of the phenomenon is treated as a variable whose values are given by measurements involving a specified technique. Some variables have a clear physical interpretation, which is to say that we know

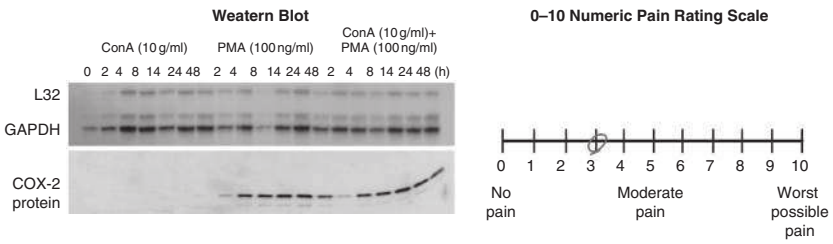

Figure 3 Data

something about the physical nature of that which is measured or, if we prefer, the factor to which the variable refers. Other variables don't have a physical interpretation. A Western blot detects proteins (Figure 3, left). In contrast, a pain assessment test measures pain, yet it is not clear what pain is beyond 'that which is reported on the occasion of a pain assessment test' (Figure 3, right). Erythema assessment stands somewhere in between. It measures several unspecified physiological changes, which include an increase of blood flow in the skin. As a rule, the more that is known about the causal structure of the world and the inner workings of the measurement technique, the clearer the physical interpretation. If little is known or if disputes arise about the correct interpretation, the only practicable option is operationalization. In this case, scientists distinguish and define different variables according to the techniques used to measure them, although it should be understood that there are no reasons to assume a one-to-one correspondence between operationalized variables and features of physical reality (Bridgman 1927, Ch. 1).

As a first approximation, data are measured values of variables. For instance, erythema is measured by visual assessment and assigned values ranging from 'no erythema' to 'violet erythema with edema'. In humans, pain intensity is measured by eliciting a physical response such as stating or pointing to a value on a numerical scale. In rats, pain is typically assessed by the Randall–Selitto test, which involves a measure of the threshold pressure at which the animal withdraws an inflamed paw. Body temperature values are given by thermometer readings. Loss of function assessment varies depending on the locus of inflammation. For instance, the severity of rheumatoid arthritis is measured by a mobility test, such as the fingertip-to-palm distance during maximal finger flexion.

While measurements do not preclude conventionalist elements, such as the choice of a particular scale, most experimental scientists assume that data consists of physical effects informative of the causal structure of the world (Trout 1998, 56–57). From a bare-bones methodological point of view, this simply means that differences in measured values reflect differences in the causal structure of the world. If this working hypothesis is correct, comparisons

between measurements can be used to localize and identify causal difference-makers. The general strategy is to repeat measurements under circumstances that are kept constant in some respects and deliberately changed in other respects; the former are part of a 'control' condition, the latter of the 'test' condition.

For example, the attempts at standardization transparent in the description of the methods of assessment of erythema and pain intensity are an elementary form of experimental control. Measurement techniques are carefully replicated from one measurement to the next and designed in such a way as to generate perceptually unambiguous outputs meant to eliminate potential disagreements among observers. Standardization aims to ensure that any particular instantiation of the technique and any particular observer can be exchanged with any other without affecting the data. If such exchangeability can be assumed, it is safe to infer that differences in data outputs are due to reasons other than differences in the measurement system. Conversely, comparisons between data generated by different or unstandardized measurement techniques are inconclusive since it is not clear what is kept constant from one measurement to the next. Such comparisons may still be valid, but they require a theory specifying a relationship between variables, such as a common physical interpretation.

The pattern of reasoning illustrated above is an application of the venerable method of difference (Mill 1843, Chapter VIII, § 2), a type of contrastive inference whereby those aspects of a situation kept constant are ruled out as possible causes responsible for differences in effects. Ruling out certain factors is useful inasmuch as it enables researchers to make more-or-less precise claims about the physical localization of the causes responsible for differences in measurements (Bechtel and Richardson 2010). In the case of standardization, the universe is decomposed into two parts: the measurement system, which is assumed to consist of essentially identical copies, and the rest of the universe, a rather large chunk of physical reality containing a yet-to-be-circumscribed object of study. A contrastive inference allows researchers to conclude that, inasmuch as the measurement system is kept constant, the causes responsible for changes in the measured values of a variable lie outside the measurement system; in other words, that something 'out there' is indeed being measured.

2.2 Experimental Models of Phenomena

In more general terms, variations in data are attributed to variations of multiple causes at work in the particular circumstances of each measurement

(Fisher 1935).[3] Fixing these circumstances is expected to reduce variability. In most cases, this is indeed what happens. Data varies less as the spatio-temporal interval between measurements is reduced. A more interesting result, however, is that data obtained by replicating methods and experimental setups also tends to display less variation than data obtained when no such precautions are taken. In some cases, this is true even if measurements are repeated after a long period of time, at different locations, and by different researchers.

Any scientific paper presenting original experimental findings contains a materials and methods section in which techniques and experimental setups are carefully described with the explicit goal of increasing the prospect that subsequent experiments replicating the same methods and circumstances will reproduce similar data. When an experimental setup is described in enough relevant detail as to ensure that highly similar data are systematically obtained over a large number of replication experiments, it is often referred to as an *experimental model of a phenomenon*. For instance, an experimental model of inflammatory responses widely used in basic and clinical research is described as follows:

> UV radiation between 270 and 400 nm, peaking at 310 nm was delivered from 10 fluorescent UV-B lamps, Philips TL20 W/12 (Philips GmbH, Hamburg, Germany), housed in a UV 800 unit (Waldmann GmbH, VS-Schwenningen, Germany). UV-B irradiance (280–320 nm) at the surface of the test areas was measured with a calibrated radiometer equipped with a SCS 280 photodetector (International Light, Newburyport, Mass., USA), and was 2.4 mW/cm^2 at a tube to target a distance of 40 cm . . . Erythema was determined by visual assessment 24 h after irradiation and was graded as follows: 0 = no erythema; 1 = barely perceptible erythema with sharp borders (MED); 2 = light red, marked erythema; 3 = dark red, marked erythema; 4 = violet erythema, edema. (Jocher et al. 2005)

The above procedure, known as the 'ultraviolet-induced erythema assay', consists in inducing artificial sunburns. The inducer of the inflammatory response, or harmful stimulus, is ultraviolet radiation, while a measure of the severity of erythema serves as a proxy for estimating the severity of the response. By specifying, measuring, and ultimately standardizing the inducer it eventually became possible to consistently reproduce inflammatory responses of a desired intensity.[4]

[3] For a historical overview of the development of statistics, see Hacking (1990) and Stigler (1986).
[4] For a history of the ultraviolet-induced erythema assay, see Brune and Hinz (2004). For a philosophical discussion of models, including experimental models, see Ankeny (2001), Ankeny and Leonelli (2011), and Baetu (2014).