

DOI: 10.1017/psa.2024.56

This is a manuscript accepted for publication in *Philosophy of Science*.

This version may be subject to change during the production process.

Biological Mistake Theory and the Question of Function

Authors

David S. Oderberg (d.s.oderberg@reading.ac.uk), Jonathan Hill (j.hill@reading.ac.uk),

Christopher Austin (christopherja@gmail.com), Ingo Bojak (i.bojak@reading.ac.uk),

François Cinotti (f.m.cinotti@reading.ac.uk), Jonathan M. Gibbins

(j.m.gibbins@reading.ac.uk)

Conflict of interest

None.

Funding statement

The authors gratefully acknowledge the financial support of the John Templeton Foundation (#62220). The opinions expressed in this paper are those of the authors and not those of the John Templeton Foundation.

Abstract

The making of mistakes by organisms and other living systems is a theoretically and empirically unifying feature of biological investigation. Mistake theory is a rigorous and experimentally productive way of understanding this widespread phenomenon. It does, however, run up against the long-standing ‘functions’ debate in philosophy of biology. Against the objection that mistakes are just a kind of malfunction, and that without a position on functions there can be no theory of mistakes, we reply that this is to misunderstand the theory. In this paper we set out the basic concepts of mistake theory and then argue that mistakes are a distinctive phenomenon in their own right, not just a kind of malfunction.

Moreover, the functions debate is, to a large degree, independent of the concept of biological mistakes we outline. In particular, although the popular selected effects theory may retain its place within a more pluralistic conception of biological function, there is also need for a more forward-looking approach, where a robust concept of normativity can be an important driver of future experimental work.

1. Introduction

Organisms get things wrong: they miss their prey, get lost, confuse food and poison, fall into traps, take bait, move too late to escape predators, and misidentify threats. Getting things wrong – making mistakes – also occurs at sub-organismic and supra-organismic levels. Antibodies are fooled by pathogens.¹ Ribosomes can cause the misfolding of proteins (Shcherbakov et al. 2019). Shoals of fish swim into nets. Pods of whales beach themselves. The dissimilarities across kinds and levels are, of course, huge, but common themes are apparent.

Mistakes involve various kinds of activity such as mislocation, misidentification, and mistiming.² Organisms³ need to act at the right time in order to achieve their ends of survival, reproduction, nutrition, self-maintenance, and so on. They have to find what they need in the right place as well. If they mistake what they need or aim for with something they do not, the results can be disastrous. Organisms often need to find or make the *right kind* or *amount* of something (food, water, shelter). They have to make the right *trade-offs* between different objectives: feeding too long might lead to becoming food oneself; acting *too* cautiously might mean not getting food at all. To speak in an Aristotelian way, organisms have to spend much of their lives acting in the right way, at the right time and place, in order to achieve the right objectives in the circumstances at hand.⁴

We know this is true for humans: mistakes are a daily part of our lives. What is often overlooked – not unknown, but not emphasised or systematised by biologists and philosophers of biology – is that mistake-making, and the potential for it, are a daily part of *all* life. We may think of mistakes as a distinctively human phenomenon involving

¹ The classic example is antigenic mimicry, which is thought to be at the root of many autoimmune diseases; see (Wildner 2023), (Oldstone 1998), where it is clear that it is not only the self antigens that fool the antibodies in many cases, but the antigens, such as viruses, themselves.

² For discussion of some specific kinds of mistake-making across different scales and systems, see (Oderberg et al. forthcoming).

³ For ease of exposition, we will usually refer only to organisms, all the while implying that mistakes can be made at sub- and supra-organismic levels, as suggested.

⁴ Aristotle, *Nicomachean Ethics* III.7, 1115b17 (Crisp ed. 2004, 50).

uniquely human characteristics⁵ such as language, self-consciousness, rational thought, free will and responsibility. Yet the general and informal concept of mistake-making is simply *getting something wrong*, not getting something wrong verbally, or self-consciously, or due to carelessness. In fact, it is not even essential to the general concept of mistake-making that the mistake be *correctible* then and there, or in the future, or that it can be learned from or conditioned out of an organism. A highly generic concept of mistake is available – one with the potential to unify biology theoretically while at the same time generating novel experimental hypotheses of interest to the practising biologist.

Our unifying mistake-theoretic framework is not intended to supplant existing successful and productive frameworks for structuring biological research. Rather, it complements these by bringing together concepts and experimental proposals that have either been largely ignored or given less prominence than deserved in contemporary discussion. In particular, our mistakes framework is designed to *operationalise* teleological concepts such as purpose, goal-directedness, agency, and the normative evaluations that come with these, such as success and failure, health and disease, flourishing and well-being, and so on. By understanding how organisms get things wrong, we highlight what it is for them to succeed at what they are built to do – flourish, survive, and reproduce. And by regarding biological mistakes as not essentially interest-dependent or investigator-dependent, nor as phenomena that are really ‘just more physics’ with no title to mind-independent and irreducible reality in their own right,⁶ we contribute to the view that living systems have a distinctive ontology that underwrites the treatment of biology as a special science.

Many of the details of mistake theory have been set out elsewhere (Oderberg et al. 2023). The present paper is an exercise in further investigation: to analyse the relation of biological mistakes to the concept of *function*. Although on first inspection one might think that mistakes can be analysed purely in terms of the more general

⁵ Or, if shared by certain other animals, then restricted to us and them.

⁶ Hill et al. (forthcoming) argues specifically for irreducibility.

function/malfunction distinction, this turns out not to be the case.⁷ Developing our theoretical framework for understanding biological mistakes is a prelude to demonstrating its empirical productivity. With the conceptual foundations in place, further work will enable the development of unified approaches to the interrogation of a diverse range of systems by the experimental biologist.

2. Biological Mistakes and the ‘Functions’ Debate

Here is an informal definition of a biological mistake:⁸ A biological mistake is one made by an organism;⁹ an organism makes a mistake just in case it does something that threatens its ability to *act effectively*, as we describe it, in a given environment. That is it. To make a mistake (the qualification ‘biological’ now omitted for convenience) is to get things wrong in one’s environment. Which, of course, raises a host of questions. One is: how does mistake theory tie into the by now hoary and unresolved ‘functions debate’?¹⁰ Is a mistake simply a kind of malfunction? If so, one might wonder what mistake theory brings to the table.

Methodologically, mistake theory as we conceive it is not obliged to have a position in the ‘functions debate’. One should see this as a virtue: the debate has become bogged down in ever more specialised and theory-laden conceptions of ‘function’, to the extent that the protagonists often talk past each other.¹¹ That mistake theory does not entail

⁷ We use ‘malfunction’ and ‘dysfunction’ synonymously, as do most writers. (That two terms stand for the same phenomenon in this context is gratuitously confusing, but we work with what has been handed down in the literature.)

⁸ For a more detailed and technical account, see (Oderberg et al. 2023) and also (Hill et al. 2022).

⁹ All the while recalling – *mutatis mutandis* for parts, sub-systems, species, and populations which belong to organisms or to which organisms belong.

¹⁰ For an excellent overview of that debate, see (Garson 2016).

¹¹ Contrast the presupposition that a correct definition of function must be historical in nature since it has to emerge from all-encompassing evolutionary theory (‘Nothing in biology makes sense except in the light of evolution’, as Theodosius Dobzhansky (1973) notoriously put it) with the assumption that evolutionary theory has nothing to tell us about function, which is a purely contemporaneous concept – that a

one particular definition of function might also be considered a virtue for allowing the theory to be adaptable to a variety of accounts. In particular, we will discuss the popular selected effects theory of function later, arguing that whatever problems that theory may independently face, it is not strictly ruled out by mistake theory as long as the selected effects theorist and the mistake theorist acknowledge that they are not asking the same questions.

A key constituent of the concept of mistake-making is that of *effective action* in an organism's environment. The notion of effective action is quite generic: it is not tied conceptually to the concept of 'the function of trait T', for instance. Moreover, whereas in the functions debate the common view is that organisms themselves have no function – this being thought solely as a characteristic of traits (Garson 2019, 20) – we take it that effective action is a holistic feature of organisms as well as of parts, sub-systems, collectives, and any entity capable of the minimal agency¹² necessary for making a mistake. Now one could, in light of the Aristotelian tradition for example, speak of effective action as a kind of function as well – *ergon* in the sense of doing that which promotes and enhances organismal flourishing.¹³ In that generic normative sense, every organism has a function, or perhaps a complex of highly general functions – to flourish, survive, and reproduce.¹⁴ In general, however, we prefer the term 'effective action' so as to put some distance between mistake theory and the functions debate. At the same time, however, we help ourselves occasionally to the term 'function' in our broad sense of 'effective action' in order to emphasise that to the extent that function does make a central appearance in mistake theory, it is normatively characterised. We will say more about this later, noting for now that it is possible for an organism to make a mistake without there

function is what a trait contributes right now to the working of a system (Walsh 1996, 558).

¹² On which see further below.

¹³ Aristotle, *Nicomachean Ethics* I.7, 1097b22-30 (Crisp ed. 2004, 11).

¹⁴ We can set aside various complexities for present purposes – for example, hybrid sterility, which appears to be rare, usually artificial, and often an effect of chromosomal abnormality. We also omit other details concerning, say, eusocial colonies such as bees and ants, where not every member has the function of reproducing.

being any *compromise* or *damage* to any of its trait functions, however they are defined or identified by the lights of a given theory of function in the narrower sense of the functions debate.

Much of the current functions debate has swirled around the idea that a correct definition of trait function must be fit for evolutionary theory. Historical accounts of function wear this requirement on their face, but it is also a feature of fitness-contribution theories (Garson 2016, ch. 4) and of the view that functions must be adaptations (Ruse 1971). On our view, although mistake theory must of course cohere with prior biological knowledge, the reality of mistake-making in biology requires us to think of function not merely in the specific ‘trait function’ sense that is the focus of the functions debate, but *also* more generally in a certain, normatively laden way that we will elaborate in terms of effective action.

3. What Mistake Theory Rules out: Mistakes vs Malfunctions vs Mere Failures

The idea of biological mistakes is, we submit, distinctive in virtue of its being both normatively laden and agential: an organism gets something wrong in a way that threatens its effective action in its environment. As such, whatever position one may take in the functions debate, room must be made for the possibility of mistakes as a distinct category of ‘going wrong’ – distinct both from malfunctions and also what we call ‘mere failures’. Mistake, malfunction, and mere failure are distinct but overlapping phenomena, as depicted in Fig. 1:

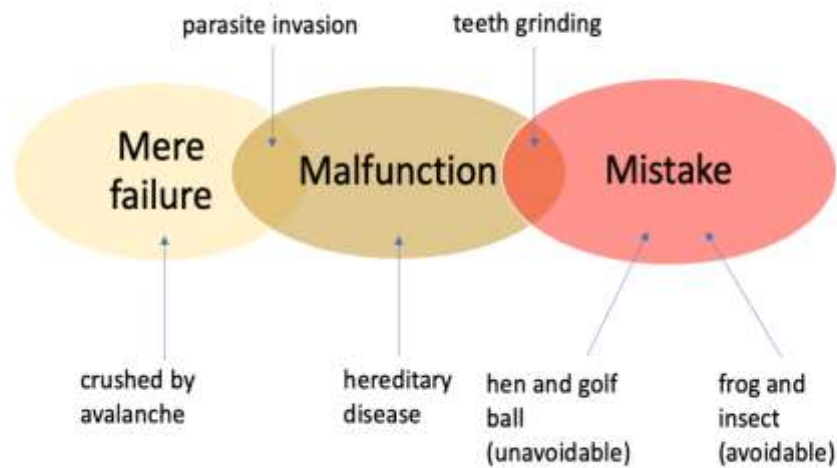


Fig. 1. Examples of mere failures, malfunctions, mistakes and their overlaps.

With the caveats in place that context is all-important and that the devil is in the details, moving from left to right we see that a mere failure is something that merely *happens* to an organism, such as being crushed by an avalanche. A malfunction is exemplified by a hereditary disease: it is any adverse state of trait or systemic breakdown or disadvantageous operation – things do not work the way they *should* – that undermines the organism’s health, integrity, survival, reproduction, etc. Mere failure and malfunction overlap, for example, in a parasite invasion that degrades the body, causing multiple system malfunctions. A *mistake* would be, for example, the case of a broody domestic hen trying to hatch a golf ball or other egg-resembling object. It gets it wrong when it comes to reproduction. Another would be the case of a frog mistiming its leap at a nearby insect. At the intersection of mistake and malfunction might be, for example, teeth grinding: one should not grind one’s teeth as it damages tooth structure and integrity, leading to other complications. But the act of grinding itself is a kind of malfunction of the operation of mastication. Moving back to the mistake category, we make the crucial point that the domestic hen that tries to hatch a golf ball is *not* malfunctioning and is not subject to a mere failure. There is no trait or systemic malfunction, no damage or breakdown. The hen *gets it wrong* with respect to reproductive function but it does not *suffer a malfunction*.

The hen fails to act effectively in its environment, to use our terminology, and so – in the broad Aristotelian sense – is not ‘living up to its *ergon*’. But this does not require us to identify any specific malfunction of the kind constituted by damage, disease, or breakdown.

Hens simply do not have the discriminative capacity to distinguish real eggs from vague simulacra such as golf balls and ceramic eggs. We call such a mistake *unavoidable*. As far as we know, hens cannot be trained out of it. The behaviour is part of their nature and not the result of pathology or of not having been sufficiently conditioned.¹⁵ An avoidable mistake is one that an organism can learn from or be trained out of; the frog that mistimes its strike at prey can do better next time. Another example of an avoidable mistake is poison-shyness by rats.¹⁶ All we need assume is that the organism has enough flexibility in its capacities that margins for error can be closed.

The concept of effective action used by mistake theory is irreducibly and objectively normative.¹⁷ Organisms act effectively insofar as they maintain their health, integrity, survival, reproduction, and overall flourishing as the kinds of thing they are. The *functions* of their traits, parts, subsystems, processes, and mechanisms – whether the answer is given by selected effects theory, or the biostatistical theory (Boorse 1977), or the causal role theory (Cummins 1975), and so on – should not, we submit, be defined in such a way as to preclude the very possibility of effective action thus construed. Again, although there are technical notions of malfunction correlative with the specific definitions of function found on various sides of the functions debate, there is also an essentially normative sense of malfunction only occasionally picked up by protagonists in the

¹⁵ <https://flockjourney.com/golf-balls-dont-hatch-but-they-help-manage-a-broody-hen/> [last accessed 14/10/24]. Farmers use this mistaken behaviour to manage their hens. Another example is the herring gull, which Niko Tinbergen notes is far better at distinguishing its own young than it is at distinguishing its own eggs (Tinbergen 1969, 149–50).

¹⁶ (Naheed and Kahn 1989). Note that a mistake by a rat in taking poison to which it has never before been exposed might well be unavoidable *at the time*, but by unavoidability we mean that by nature the organism is incapable of learning from or being trained out of the mistake.

¹⁷ Again, on irreducibility see (Hill et al. forthcoming).

debate.¹⁸ A part, sub-system, mechanism, process, and so on, malfunctions just in case it suffers intrinsic damage, disease, or breakdown rendering it incapable of performing its function, however its function is specified (e.g. as a selected effect, biostatistical norm, and so on). An organism whose part, sub-system, etc., suffers a malfunction in this sense will be (probably if not certainly) incapable of acting effectively in its environment and so can be said to malfunction *itself* – where organismic malfunction is simply derivative from the malfunction of parts, sub-systems, and the like. Mistakes and malfunctions, as we saw, are overlapping but distinct phenomena. A mistake can occur without any malfunction in the sense given.

Here we agree with selected effects theorists of function such as Ruth Millikan and Justin Garson that malfunction concerns what an organism¹⁹ ‘constitutionally’ (or ‘intrinsically’) cannot do (Garson 2019, 127; Millikan 2013, 40). However, whereas they fix intrinsicness by reference to what an organism can or cannot do in its selective (or ‘Normal’) environment (Garson 2019, 128–9) – more precisely, the environment in which the trait in question was selected for – we fix intrinsicness purely by reference to internal constitution, without indexing this to a privileged environment or range of privileged environments – whether selective, statistically typical for the organism, and so on. However else malfunction might be defined relative to a position in the functions debate, there is also clearly a sense of malfunction that requires some kind of damage, disease, or breakdown of a trait, system, process, and so on, such that it is incapable of promoting, enhancing, or protecting an organism’s capacity to act effectively in its environment. We will discuss this further, noting for now that malfunction – intrinsic incapacity to function – may of course be *caused* by external circumstances. An environmental pathogen can cause disease – a kind of intrinsic malfunctioning of some part or sub-system. A lack of oxygen in the environment can cause the malfunctioning of red blood cells. There is, then,

¹⁸ For example (Millikan 1989, 295).

¹⁹ Again, using ‘organism’ as a shorthand term for both the organism and a whole and its parts, sub-systems, traits, processes, mechanisms, and the like.

no incompatibility between a malfunction's being an intrinsic feature of an organism yet dependent, causally, on environmental factors.

Returning to the contrast between mistake and malfunction, it might seem plausible to suppose that mistakes are just a certain kind of malfunction: having a brain disease, or defective eyesight, or too few teeth, are so closely connected to mistakes that it looks like two terms are being used for one phenomenon. If a person gets indigestion through failure to chew their food properly, and this because they have too few teeth, they seem to have malfunctioned: the mistake appears to be *in* the malfunctioning. But this is not quite right. The malfunction, on our view, is the possession of too few teeth and the incomplete digestion. The *mistake* is not masticating the food sufficiently. In other words, malfunctions can *cause* mistakes and *be caused* by them but that is different from saying that mistakes are *constituted* by malfunctions. Mistake theory sees malfunctions and mistakes as part of a complex causal network, without equating the phenomena.

That said, *some* mistaken behaviour can be constituted by a malfunction: limping due to malformed limbs is a case where the movement itself is a malfunction – an engineering breakdown. Incorrect colour discrimination due to colour-blindness is mistaken behaviour constituted by malfunctioning eyes. Yet mistake is not just a special kind of malfunction because it can happen with *no malfunction at all*. Broody domestic hens, as we have noted, do not malfunction when they try to hatch golf balls. Nor do fish malfunction when they take the bait. In such cases, absent a disease, an injury, or some kind of damage, there is no basis for identifying any malfunction. These organisms are doing what they do according to their own natures and accompanying suites of powers and liabilities. The hen that tries to hatch a golf ball is not deficient *as a hen* – and there is no other standard to which it should be compared. Hens just do not have the ability visually or tactilely to discriminate between a real egg and what to us is a vague simulacrum. So they get it wrong – they are not supposed to hatch anything other than their own eggs – but they do not malfunction because it is not part of their correct function, given what they are, to have the kind of perceptual ability that would make the needed distinction. The same applies to every species, at some fine enough grain of discrimination.

The assertion that not all mistakes are mere failures, as we have defined both, might seem to go without saying. After all, it might be thought, every failure is a kind of malfunction and, since we have shown that not all mistakes are malfunctions, it follows that not all mistakes are mere failures. Yet some failures are arguably *not* malfunctions in the sense of involving internal system breakdown. Perhaps death through ageing and disease is the ultimate malfunction – eventual total breakdown. Sudden death, however, does not look like a malfunction except in a superficial sense: being smashed to bits by an avalanche, vaporised by a laser, crushed by deep sea pressure, or electrocuted are all fortunately uncommon, but it is a stretch to call them malfunctions. They are, nevertheless, a kind of *mere* failure, by which we mean something that merely *happens* to an organism, not something it *does*. The same goes for more common conditions that clearly fall into the malfunction category – getting sick and being injured or damaged. Being thrown off course by the wind, or placed in an inhospitable environment, are not in themselves malfunctions, perhaps, but they can well lead to malfunction, and of course to mistake.

The concept of a biological mistake is of an organism's *getting* things wrong, not merely being subject to conditions in which things *go* wrong for it. Mistakes only 'happen' in the jargon of evasive bureaucrats and politicians; the truth is that mistakes are *made*, and making requires *agency*. On mistake theory,²⁰ the Minimal Biological Agent is the organism inasmuch as it interacts causally with its environment such that some other thing ceases to exist, or comes into existence, or undergoes a change of intrinsic feature, without the organism itself ceasing to exist or changing intrinsically. Alternatively, the organism might not act on its environment but may act only on itself, so parts of it will cause other parts to cease to exist, come into existence, or change intrinsically. A bird eating a fish is an MBA on the fish. A wolf watching out for prey is not doing anything *to* its environment but it is doing things *to itself* and so is an MBA on itself – getting itself into a certain predatory posture, adjusting its perceptual organs, moving its limbs in various ways.²¹

²⁰ See (Oderberg et al. 2023) for the details.

²¹ It is also an MBA *with respect to* its environment: it is not acting *on* it but acting with some environmental goal in mind, such as prey.

4. Going Wrong and Getting it Wrong

It is illuminating to see the extent to which mistake theory aligns with a recent account of how organisms can ‘go wrong’ (Matthewson and Griffiths 2017). John Matthewson and Paul Griffiths identify four ways of going wrong: (i) Mechanism failure – the breakdown of some mechanism within the system. (ii) There is no failure of mechanism but the organism is forced to operate in an environment in which it or the relevant trait was not selected for. (iii) The organism operates in an ‘inhospitable’ environment – one in which it or its traits were selected for, but in which conditions have deteriorated, thereby affecting fitness. (iv) The mechanisms work correctly, the environment is one in which original selection applied, the conditions are not inhospitable, but the outcome of a developmental trajectory is no longer ideal – their example²² being a water flea (*Daphnia cucullata*) born with protective parts that use developmental resources but have become unnecessary due to a decline in the predator population.

Mistake theory makes finer distinctions than this fourfold classification allows. Note that the authors do not but distinguish general malfunction (‘dysfunction’) from specific mechanism failure (Matthewson and Griffiths 2017, 453). One can infer that for them malfunction *probably* means simply ‘going wrong’, as per the title of their paper. However, a mechanism failure (their first way) may be a *mere failure* – something caused by exogenous damage. Or it may be a kind of mistaken behaviour. Their example is instructive – a genetic mutation in mice leading to a deficiency of the hormone leptin, causing the mice to eat to the point of obesity (Matthewson and Griffiths 2017, 453). The mutation may be a mistake, for example by RNA polymerase in transcription or aminoacyl tRNA synthetase in translation – enzyme behaviour being about as clear an example of biochemical agency as one can hope for.²³ But if it is caused by environmental toxicity this would look more like a mere failure. The hormone deficiency *itself* is not a mistake, just a

²² See also (Garson 2015, 70, 144).

²³ See further (Kuykendall 2024/2021).

failure in the sense of a mere lack of something needed. However, the misregulation of appetite caused by it is a kind of mistaken behaviour, in particular eating beyond satiety.

Operating in a non-selective environment (their second way) might lead to mere failure – starvation, freezing, or just death. Matthewson and Griffiths' example, however, is a male glow-worm in an urban environment, where the abundance of light makes it impossible to locate a mate by its light signature. Here, we have a possible mistake due to the worm being confused or deceived by an incorrect light source. It is not clear, absent further detail, that there is either malfunction or mere failure. Rather, the worm simply *mislocates* mates, which could be a mistake of *omission*, or a mistake of *commission* by moving towards light-emitting objects that are not mates.

Concerning their third way – an inhospitable environment – for us a mistake is made when an organism does something that threatens its effective action in the environment; and we mean whatever environment the organism finds itself in – selected for, hospitable, or not. Similarly, malfunctions and mere failures are evaluated relative to whatever environment the system is in. Matthewson and Griffiths' example is of a plant growing in a poor location yet within its selective environment. It may grow less and flower less, yet it is doing all the things it was selected for, having traits that evolved to deal with variable conditions.

Whether plants make mistakes is an intriguing question.²⁴ Suppose a plant sends out roots that are too shallow to get enough nutrition; this is a mistake, not a mere failure. It may not involve malfunction – a constitutional inability to achieve proper root growth. In a harsh environment, say where the soil is too compacted to allow growth, the plant will have acted in a mistaken way, albeit unavoidably. A plant flooded by excess water is subject to mere failure. A pest-ridden environment can cause disease and so affect the plant's intrinsic ability to function. Much more can be said about such cases; the inevitable

²⁴ For interesting lines of enquiry, see (Goldshtein et al. 2020); (Gruntman et al. 2017). We hypothesise that mistake-proneness, as a feature of biological systems, is also true of plants.

vagueness²⁵ will present conceptual problems. That overlooked distinctions can be made in this context should, however, be clear.

The fourth way of going wrong, which Matthewson and Griffiths call a ‘heuristic failure’, occurs because ‘developmental trajectories must be initiated in the setting of imperfect information’ (Matthewson and Griffiths 2017, 456). Their example of morphogenesis in water fleas involves no malfunction and looks like a mere failure – something that happens to the organism. But if the unnecessary resource expenditure involved a kind of *misregulation*, or perhaps *mistiming* of development, we should call it a mistake. ‘Imperfect information’ also may be attributable to the organism’s actions, since information needs to be *registered* and *processed*. If the present case involved an incorrect quantification of predator levels,²⁶ it would be an informational mistake. As always, the devil will be in the details: mistake theory should encourage experimentalists to attend to the kinds of mistake that could potentially be made in any given case of things ‘going wrong’.²⁷

5. Looking Forward and Looking Back

As we have outlined, the theory of biological mistakes largely prescind from the particular technical debate about how to define trait and related functions. That said,

²⁵ A problem for most boundary-drawing in most areas.

²⁶ This case differs interestingly from the hen and the golf ball. The hen trying to hatch a golf ball is taking a reading in real time of whether eggs are in its environment. The water flea that grows defensive parts in a low-predator environment is not responding to anything it represents in real time. Rather, its production of such parts is an epigenetic response to *historical* representation by its ancestors of a high-predator environment. Does this mean that its growing these parts is not a mistake? It depends on whether correct function is threatened. If having those extra parts made swimming less efficient, thus increasing energy expenditure and consequently, for example, the risk of starvation, then it would be a mistake.

²⁷ Note that Matthewson and Griffiths say of this case (456): ‘It is always better to be in a situation of low predation than not, but nevertheless, this failure of the heuristic decision method will mean the offspring have paid a developmental cost they did not need to pay. These water fleas would have been better off if they had not prepared for predation.’ This seems to imply a decision mistake by *both* mother and offspring! In any case, the details of this particular case are not relevant to our general theoretical point.

however, an instructive comparison and contrast can be made with the popular – even perhaps orthodox – selected effects (SE) theory of function, in order to highlight where mistake theory fills gaps or makes contributions that are largely absent from SE theory, whether it be the classic version focusing on natural selection (Millikan 1984; Neander 1991) or the generalised form (Garson 2019). First, note that our emphasis on interrogating organisms for the *potential* to make mistakes – not simply trying to explain mistakes already made – gives it a forward-looking character that is absent from an historically oriented theory of function such as SE theory. The latter aims to pin down the function of a trait by identifying what it was naturally selected for, or what gave it an advantage over competing traits, thereby allowing it to persist. This backwards-looking feature makes it hard, in our view, for SE theory to handle the essential *normativity* of functions, even if it is conceded to be a correct account of function according to its purely descriptive aspects.

Although a forward-looking approach is not incompatible with the framework of SE theory, the conceptual orientation of mistake theory does not privilege history over contemporaneous fact and future prospects for organisms. By its very nature, SE theory asks us to look backwards, into a history we often grasp only dimly, to understand what an organism's powers and capacities are. We seek to reorient the investigative perspective to consider equally (if not more so) the potential of an organism to make mistakes or deviate from standards of correctness. Mistake theory is designed in large part for scientists looking to develop novel research hypotheses concerning current and potential future behaviour. It is less clear that SE theory is suited to drive lab-bench biological testing in a similar fashion, especially when normativity is in play.

For a selected effects theorist such as Garson, the expression 'normativity of function' involves 'nothing having to do with values or goals, oughts and shoulds, prescriptions or commands, the good or the just' (Garson 2019, 15). According to him, 'dysfunction' or 'malfunction' is no more than a constitutional inability to perform that for which the trait in question was selected (that is, within its normal environment) (Garson 2019, ch.8). One could consider this to be a stipulation of the theory: if functions are defined as selected effects, then malfunctions must be defined accordingly. Still, there are normative aspects of malfunction about which SE theory has less to say, and to which

mistake theory – not as a theory of function, which it is not, but as a theory about biological behaviour generally – makes a useful contribution.

So, for instance, present-day humans do not, constitutionally, have the skills for performing various functions in the savanna, especially surrounded by wild animals. We cannot run for long periods in rough terrain with bare feet; we are far less resistant to various physical challenges – certain microbes, unpalatable food sources, extreme temperatures – than in the savanna. That does not imply we are now malfunctioning in a normative sense,²⁸ even if in some respects we have ‘gone soft’.²⁹ Constitutional degeneration over time does not entail normative malfunction – malfunction that undermines organismal welfare – even though it might involve a constitutional inability, as per SE theory, to do things for which certain traits were historically selected.

Conversely, the constitutional inability to do what a trait was selected for – even if this is stipulatively a malfunction on SE theory – is not necessary for what would be a clear case of normative malfunction. Take the familiar scenario of a polar bear unhappily being dumped in the Sahara Desert. It will not take long for the polar bear seriously to malfunction on any reasonable normative interpretation of that term. Its system will break down catastrophically, leading to perhaps the ultimate malfunction – death. Yet, on the SE stipulation of what counts as malfunction, the polar bear is not malfunctioning³⁰ because it is not *constitutionally* incapable of performing, in its ‘normal’ (i.e., selective) environment, the functions for which it and/or its traits were historically selected. It is simply that its current environment is ‘abnormal’ and hence ‘uncooperative’. Again, what might not count as a technical malfunction on historical, purely descriptive grounds, could still count

²⁸ Or, perhaps more precisely in the present context, it does not mean that our locomotive or digestive traits, for example, are malfunctioning.

²⁹ Many of us but not all of us, of course. And we would prefer current life expectancy to the life expectancy in the savanna. There is plenty of room for disagreement over whether a given lack of ability in contemporary humans could be recovered through intensive training or conditioning, but see (Chirchir et al. 2014) on the apparent constitutional degeneration of bone density relative to past *homo sapiens*.

³⁰ Or as Garson often says, ‘dysfunctional’.

as one on contemporaneous, normative grounds. Function pluralism would seem to many to be an attractive option in this context.

The selected effects theorist might reject the pluralist suggestion at this point, replying that the above characterisation of SE theory is too quick. They might insist that a trait malfunctions just in case it cannot, constitutionally, do what it was selected for; and there is nothing else that need be added. In a hostile environment such as the desert, a polar bear will be constitutionally – intrinsically – incapable of thermoregulation; its constitutional abilities were conferred by natural selection, and these either can or cannot be performed in any given environment. This will always be a matter of intrinsic capacity except where there is no system breakdown, disease or malformation, and so on, but there is some external interference with or blocking of the exercise of a perfectly healthy trait. This might be the view of some selected effects theorists such as Karen Neander,³¹ but it is not Garson's view. As he explicitly puts it, 'Something dysfunctions just when it can't perform its most proximal function in its normal environment' – this being the environment in which the trait was selected for (Garson 2019, 125). Hence, by implication, the polar bear in the desert does not malfunction on this definition, even though there is a normative sense in which it clearly does malfunction.

We submit that relativisation to selective environment is the view that *ought* to recommend itself to SE theorists. For if the function of a trait is determined by selection in a given environment (better, range of suitably similar environments), then so by parity of reasoning should *malfunction* be determined relative to that environment. Perhaps more importantly, if the SE theorist does not index malfunction to selective environments, then selection itself no longer does any work in the determination of function, with the result that the theory collapses into a type of non-historical theory. Suppose we took malfunction to be no more than the constitutional ability to perform a function. In other words, even though the trait was selected for in a given environment, malfunction would be judged relative to *whatever* environment the trait-bearer found itself in. Then it would be hard to see why we should appeal to selection as a determiner of function in the first place. In

³¹ It is implicitly the view of (Neander 2017), at least in broad terms.

order, then, to maintain a pluralist account whereby a broad, normative sense of function – such as undergirds mistake theory – can co-exist with a more technical, or as we’ve suggested stipulative, account of function in terms of selected effects theory, the SE theorist needs to index malfunction to a selective environment, in order for SE theory to maintain its distinctive status alongside a more normative approach to function. On the latter approach, which has intuitive support in our view, the polar bear in the desert *does* malfunction even if on a more narrow SE reading it is not malfunctioning but *merely* in a hostile or uncooperative environment. Things *go wrong* with it in a comprehensive and catastrophic way, with its homeostatic traits constitutionally incapable of doing what they are supposed to do, leading to many other traits – locomotive, nutritive, vegetative – also unable to perform their functions.

Leaving selected effects to one side, the attempt to strip evaluation from all conceptions of function (even while seeking to retain the term ‘normativity’)³² leads, we suggest, to an overly narrow conception of the latter and a lack of connection between what organisms *do* and their *welfare*, without which important distinctions collapse – such as health/disease, integrity/breakdown, flourishing/deterioration, and generic ones such as correctness/incorrectness and success/failure. In this we fully agree with Georges Canguilhem: ‘life is what is capable of error’ (Canguilhem 1991, 22). Suppose the concept of ‘constitutional (in)ability’ carried nothing essentially normative in its meaning. Then abilities and inabilities would, of themselves, be mere physical variations in an organism’s behaviour or processes: physically speaking, ability and inability, and so function and malfunction, would be on a par. But if they are not, because malfunction is a departure from behaviour or processes that support, maintain or promote an organism’s *welfare* – as we say, its *effective* or *successful* interaction with its environment – then there must be real normative features of constitutional ability or inability. By contrast, we rightly do not say that gold is dysfunctional because it cannot dissolve in water, or that if salt does not

³² Albeit for Millikan this term has to go, along with anything ‘mysterious’: (Millikan 1984, 17).

dissolve in water – say, because the water is already supersaturated – the salt must be *failing* or departing from what it is *supposed* to do.³³

On a broad, normative sense of function understood as effective action or whatever subserves effective action – the focus of mistake theory – powers and abilities are at the heart of our evaluation of whether an organism is able to act effectively in its environment. This understanding can complement, without replacing, more technical senses of ‘function’ such as we find in the functions debate. Mistake theory requires us to ask what mistakes an organism not only has made but does, will or even *could*: the answer to such questions will involve interrogation of the organism’s actual powers and abilities, even as their ancestry gives us a particular answer to the question of what the function of a given trait is in a more narrow sense.

Similar remarks, in the spirit of ‘function pluralism’,³⁴ can also be made concerning the so-called ‘Cummins function’ of a trait.³⁵ A Cummins function is a feature of an entity that adequately explains the capacity of a system containing that entity as a component to behave or operate in a certain way. It would take us too far afield to assess causal role functional analysis (Cummins functions, broadly construed) in its own right.³⁶ That said, we suggest that causal role functions can co-exist with, and be informed by, mistake theory.³⁷ Mistake theory can help itself to descriptive analyses of how complex systems are subserved by their components, or how whole-system functions are decomposable (at least in some cases) into sub-systemic constitutive functions. However,

³³ It is arguable that the concept of *selection* itself is itself normative, because otherwise selection and non-selection would be mere physical variations – ‘just more physics’, as it were. But biology sees a big difference, since selection is precisely for *adaptiveness*, for *successful* negotiation of the environment (even if indirectly via a trait’s contribution to organismic performance).

³⁴ We cannot explore function pluralism here; see (Garson 2016, ch.5.3) for an overview.

³⁵ (Cummins 1975), with a useful precis in (Cummins 2010, 290–92).

³⁶ See (Garson 2016, ch.5) for a survey of some problems with this theory.

³⁷ Cummins occasionally indulges in quasi-normative talk such as ‘breakdown or abnormal functioning’ (1983, 181), but there is no explicit indication of any normative attribution.

reference to the well-being of the organism (perhaps species or other collective), and of the components themselves insofar as they subserve the well-being of the organism, would enable us to understand better how causal role functions can contribute to getting things *wrong*. To take one example, the circadian system - one of the principal time-keeping processes within virtually all living things (Sehgal ed. 2015; Albrecht ed. 2010) – is a complex network of systems and sub-systems. Molecular circadian clocks are even found in individual body cells, such as mammalian peripheral body cells, all ‘entrained’ – to use the circadian terminology – by a master circadian clock in the brain’s hypothalamus, itself entrained by the twenty-four-hour day-night cycle (Kornmann et al. 2007). From the perspective of mistake theory as an organising framework for hypothesis generation, probing diverse systems for what might be underlying, highly general timekeeping mechanisms could yield important results. Moreover, interrogating systems for general time-keeping properties should give insight into whether there are mechanisms for *protection* against mistakes, or *correcting* them once they have occurred. An unanswered question is: why does a circadian system not undergo a phase shift just by exposure to a flash of lightning? To what extent is measurement of the *duration* of the exposure a factor? Understanding the role of the *normative* in organismal structure and behaviour enables questions to be asked about what departure from a standard amounts to. If, say, a time-keeping mistake is made by an organism and it synchronises to a lightning flash, this will be deleterious; which is why, if it happens, it must be exceedingly rare. Interrogating a system for its circadian causal role functions is a pre-requisite to understanding how the system works; but interrogating it for its mistake potential, based on standards of correctness and incorrectness, takes us a step further both theoretically and experimentally.

This approach is generalisable inasmuch as, once we have a grasp of the function of a trait in a more narrow sense – answering the question of, say, what it was selected for, or what its contribution is to the operation of some larger system – we can then investigate how the trait contributes to organismal (and hence indirectly species) well-being in environments quite removed from those in which the trait was selected, or became an exaptation, or was stabilised epigenetically, among other processes. More particularly, the mistake theorist’s understanding of normativity is essentially *evaluative*, as per the

standard usage of the term. But to say that it is evaluative does *not* mean that it involves anyone *valuing* anything, or ethics, or preferences, or judgments.³⁸ Rather, to say that normativity is evaluative – specifically, in the biological realm – is to say that living systems rely essentially on such phenomena as *timeliness* and *effectiveness* of action, *accuracy* of behaviour or operation, and so on. There are *correct* and *incorrect* ways for them to act so as to maintain *good* or *bad* states of being – health or disease, integrity or disintegration, ultimately life or death. Malfunctions can be investigated not merely as constitutional inabilities, but as diminutions of welfare. A broken limb prevents an organism from moving *well*. Blocked arteries compromise *health* – physical *wellness*. An impaired sense of taste or smell undermines the ability to distinguish food from that which is not food, or spoiled food, or even toxic material – all of which are *bad* for the organism.³⁹

The normative conception of function in a broad sense that we espouse is what underwrites the normativity built into the concept of mistake. A mistake need not itself be a malfunction, as shown earlier, but mistakes *threaten* effective action: if the causal link between action and outcome is not blocked, effective action is undermined. Note that a threat need not be significant or immediate. When a deer stands in the road dazzled by headlights, it must correct its behaviour rapidly, or its life is in danger. But in the case of a

³⁸ Compare (Garson 2016, 15). While we agree with Garson that normativity of function does not involve ethics or ‘values’ (except in the human case, at least – and we would argue that ethics is at least in part a specific case of more generic biological normativity), we do not draw the conclusion he draws that it is about *no more than* capacity for performance. The approval of a value-free conception of normativity of function is also clear in (Neander 1991), (Neander 2017), (Millikan 1989), and (Millikan 1984). By contrast, we claim that goals are essential to functions, no matter how narrowly specified, e.g. in terms of selected effects, and that function in our *broad* sense has evaluative features, as explained above; they cannot be understood independently of a whole raft of evaluative notions. In other words, it is not *mere* performance that defines function in the broad sense; rather, performance must always be adequate, effective, timely, proportional, and so on – in other words, infused by that raft of normative notions that together comprise what we call a ‘standard of correctness’.

³⁹ Although this concept of biological normativity might seem non-naturalistic, we think this is due to an impoverished conception of the ‘natural’, a diagnosis found in writers such as (Foot 2001), (Okrent 2018), and (De Caro and Macarthur 2010).

hen sitting on a golf ball, the threat is merely a lack of reproduction resulting from pseudo-hatching behaviour – if the hen is not moved away or given a real egg to hatch at some point. Still, absent such correction, effective action is undermined – in this case, reproductive function in the narrow trait function sense.

That mistakes can be made and are regularly made across all organisms and scales points to the existence of correct and incorrect ways for organisms to behave – correct and incorrect ways to act effectively in our broad normative sense. Mistaken behaviour by definition departs from correctness and is inconsistent with the health, well-being and overall flourishing of the organism – whether the inconsistency be immediate and significant or indirect and minor. There are some well-known experiments conducted on spiders to see what kinds of web they spin while influenced by various psychoactive substances (NASA 1995; Witt 1954). On the evaluative conception of function in the broad sense that we espouse, a spider that spins a web full of holes when intoxicated by chloral hydrate has made a clear mistake: it will not *act effectively in its environment* because it has compromised its ability to catch prey. A critic might wonder: how could the spider be making a mistake? What other kind of web would you expect it to make on that drug? Isn't it doing exactly what it is *supposed* to do in that circumstance? Our reply is that it is certainly doing what it is *expected* to do – producing a sub-optimal web. But it is decidedly not *supposed* to do that. It is supposed to make efficient webs for its *benefit*. This distinction reveals something metaphysically deep – that biological normativity is not identical to mere regularity or expectation. Correlatively, mistakes are not mere irregularities or kinds of atypical or unexpected behaviour.

6. Conclusion

The theory of biological mistakes offers a new way of looking at living systems. We do not pretend that biologists never have mistake-making in mind in their investigations; nor do we suppose the concept wholly alien to philosophers of biology. That said, we propose that the concept of mistake is an *organising* idea, one that can shape both philosophical and experimental research on living systems. It is not designed to replace or contradict what we already know from the existing fruitful frameworks in which biological research

takes place. Rather, it offers a fresh perspective on biology as a special science that has embedded within it normative phenomena not found in physics or chemistry.

In the present discussion we have argued that mistake theory need not stake out an official position in the ‘functions debate’ and can co-exist with a certain ‘function pluralism’. Still, it does require a broad sense of what we call ‘effective action’ – for which the term ‘function’ can be used in a wide sense compatible with specific positions in the debate – and that this is the focus of investigation when interrogating living systems for their mistake potential. On one hand, then, biological mistake theory can keep the functions debate at a comfortable arm’s length. On the other, some of the demands of the theory should, we suggest, encourage function theorists to think more about biological normativity as an organising framework and about its place in lab-bench research.⁴⁰

⁴⁰ The authors gratefully acknowledge the financial support of the John Templeton Foundation (#62220). The opinions expressed in this paper are those of the authors and not those of the John Templeton Foundation.

References

- Albrecht, Urs., ed. 2010. *The Circadian Clock*. Dordrecht: Springer.
- Boorse, Christopher. 1977. "Health as a Theoretical Concept." *Philosophy of Science* 44:542–73. doi:10.1086/288768.
- Canguilhem, Georges. 1991. *The Normal and the Pathological*. New York: Zone Books.
- Chirchir, Habiba, Tracy L. Kivell, Christopher B. Ruff, and Brian G. Richmond. 2014. "Recent Origin of Low Trabecular Bone Density in Modern Humans." *Proceedings of the National Academy of Sciences* 112:366–71. doi:10.1073/pnas.1411696112.
- Crisp, Roger, ed. 2004. *Aristotle: Nicomachean Ethics*. Cambridge: Cambridge University Press.
- Cummins, Robert. 1975. "Functional Analysis." *The Journal of Philosophy* 72:741–65. doi:10.2307/2024640.
- Cummins, Robert. 1983. *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press/Bradford Books.
- Cummins, Robert. 2010. *The World in the Head*. Oxford: Oxford University Press.
- De Caro, Mario, and David Macarthur. 2010. "Science, Naturalism, and the Problem of Normativity." In *Naturalism and Normativity*, ed. Mario De Caro and David Macarthur, 1–19. New York: Columbia University Press.
- Dobzhansky, Theodosius. 1973. "Nothing in Biology Makes Sense Except in the Light of Evolution." *The American Biology Teacher* 35:125–29. doi:10.2307/4444260.
- Foot, Philippa. 2001. *Natural Goodness*. Oxford: Clarendon Press.
- Garson, Justin. 2016. *A Critical Overview of Biological Functions*. Cham: Springer.
- Garson, Justin. 2019. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.
- Goldshstein, Aya, Marine Veits, Itzhak Khait, Kfir Saban, Yuval Sapir, Yossi Yovel, and Lilach Hadany. 2020. "Plants' Ability to Sense and Respond to Airborne Sound is Likely to be Adaptive: Reply to Comment by Pyke et al." *Ecology Letters* 23:1423–25. doi:10.1111/ele.13514.

- Gruntman, Michal, Dorothee Gross, Maria Májeková, and Katja Tielbörger. 2017. "Decision-Making in Plants under Competition." *Nature Communications* 8:2235. doi: 10.1038/s41467-017-02147-2.
- Hill, Jonathan, David S. Oderberg, Ingo Bojak, and Jonathan M. Gibbins. 2022. "Mistake-Making: A Theoretical Framework for Generating Research Questions in Biology, With Illustrative Application to Blood Clotting." *The Quarterly Review of Biology* 97:2–13. doi:10.1086/718736.
- Hill, Jonathan, David S. Oderberg, Ingo Bojak, Jonathan M. Gibbins, Christopher Austin, and François M. Cinotti. Forthcoming. "Mistakes and Correctness: A New Look at the Reductionism Debate in Biology."
- Kornmann, Benoît, Olivier Schaad, Hermann Bujard, Joseph S. Takahashi, and Ueli Schibler. 2007. "System-Driven and Oscillator-Dependent Circadian Transcription in Mice with a Conditionally Active Liver Clock." *PLoS Biology* 5, no. 2:e34. doi: 10.1371/journal.pbio.0050034.
- Kuykendall, Davis. 2024/2021. "In Defence of the Agent and Patient Distinction: The Case from Molecular Biology and Chemistry." *British Journal for the Philosophy of Science* 75: 443–63. Originally published online 2021. doi:10.1086/715470.
- Matthewson, John, and Paul E. Griffiths. 2017. "Biological Criteria of Disease: Four Ways of Going Wrong." *The Journal of Medicine and Philosophy* 42:447–66. doi:10.1093/jmp/jhx004.
- Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
- Millikan, Ruth Garrett. 1989. "In Defense of Proper Functions." *Philosophy of Science* 56:288–302. doi:10.1086/289488.
- Millikan, Ruth Garrett. 2013. "Reply to Neander." In *Millikan and Her Critics*, ed. Dan Ryder, Justine Kingsbury, and Kenneth Williford, 37–40. Malden, MA: Wiley-Blackwell.
- Naheed, Ghazala, and Jamil Ahmad Khan. 1989. "'Poison-Shyness' and 'Bait-Shyness' Developed by Wild Rats (*Rattus rattus* L.). I. Methods for Eliminating 'Shyness'

- Caused by Barium Carbonate Poisoning.” *Applied Animal Behaviour Science* 24:89–99. doi:10.1016/0168-1591(89)90037-3.
- NASA. 1995. “Using Spider-Web Patterns to Determine Toxicity.” *NASA Tech Briefs* 19, no. 4:82. <https://arachnidlady.files.wordpress.com/2013/08/nasa-tech-brief.pdf> [last accessed 14/10/24].
- Neander, Karen. 1991. “Functions as Selected Effects: The Conceptual Analyst’s Defense.” *Philosophy of Science* 58:168–84. doi:10.1086/289610.
- Neander, Karen. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.
- Oderberg, David S., Jonathan Hill, Ingo Bojak, Jonathan M. Gibbins, Christopher Austin, and François M. Cinotti. 2023. “Biological Mistakes: What They Are and What They Mean for the Experimental Biologist.” *British Journal for the Philosophy of Science*. <https://www.journals.uchicago.edu/doi/pdf/10.1086/724444>.
- Oderberg, David S., Jonathan Hill, Ingo Bojak, Jonathan M. Gibbins, Christopher Austin, and François M. Cinotti. Forthcoming. “Getting It Wrong: Biological Mistake Making as a Cross-System, Cross-Scale Phenomenon.”
- Okrent, Mark. 2018. *Nature and Normativity*. London: Routledge.
- Oldstone, Michael B.A. 1998. “Molecular Mimicry and Immune-Mediated Diseases.” *FASEB Journal* 12:1255–65. doi:10.1096/fasebj.12.13.1255.
- Ruse, Michael. 1971. “Functional Statements in Biology.” *Philosophy of Science* 38:87–95. doi:10.1086/288342.
- Shcherbakov, Dmitri., Youjin Teo, Heithem Boukari, Adrian Cortes-Sanchon, Matilde Mantovani, Ivan Osinnii, James Moore, Reda Juskeviciene, Margarita Brilkova, Stefan Duscha, Harshitha Santhosh Kumar, Endre Laczko, Hubert Rehrauer, Eric Westhof, Rashid Akbergenov, and Erik C. Böttger. 2019. “Ribosomal Mistranslation Leads to Silencing of the Unfolded Protein Response and Increased Mitochondrial Biogenesis.” *Nature Communications Biology* 2: 381. doi: 10.1038/s42003-019-0626-9.
- Sehgal, Amita, ed. *Circadian Rhythms, Parts A and B*. 2015. London: Academic Press/Elsevier.

Tinbergen, Nikolaas. 1969. *The Study of Instinct*. Originally published 1951. Oxford: Oxford University Press.

Wildner, Gerhild. 2023. “Antigenic Mimicry – The Key to Autoimmunity in Immune Privileged Organs.” *Journal of Autoimmunity* 137:102942.
doi:10.1016/j.jaut.2022.102942.

Witt, Peter. “Spider Webs and Drugs.” 1954. *Scientific American* 191, no. 6 (1954): 80–87. doi:10.1038/scientificamerican1254-80.