This is an accepted manuscript of a paper forthcoming in ***Synthese***. Please cite the published version once available.

# How to Explain Degenerate Mechanisms

**Abstract**

Degeneracy, the ability of structurally different elements to perform the same function and give rise to the same phenomenon, is believed to be ubiquitous at all levels of mechanisms in neurobiology. Given its biological salience, degeneracy has become an emerging topic in recent scientific literature. In this paper, I will present a new strategy for researchers to offer mechanistic explanations for degenerate mechanisms in the nervous system as complementary to the received new mechanist account. Specifically, I argue that, due to degeneracy, exemplar mechanistic models built using averaging techniques are sometimes beset by the 'failure of averaging' problem. To avoid this problem, many researchers opt for an alternative strategy – which is to offer what I call the *population-based mechanistic explanations*. According to this strategy, a large population of models instead of single exemplar models are generated to capture the real-life variability of biological mechanisms. By examining an example from cellular neuroscience, I offer an account of population-based mechanistic explanations and show that they differ from the 'ordinary,' exemplar-based mechanistic explanations by providing different explanatory information. Finally, I argue that the use of population-based mechanistic explanations has implications for the new mechanist philosophy. Specifically, I suggest that the notion 'how-actually mechanistic models' needs to be reassessed when degeneracy is taken into account.

## 1. Introduction

No two leaves are alike – the spirit of this saying is exemplified all around the biological world. The nervous system, in particular, exhibits enormous variability. From molecule structures to system-wide properties, structural differences are found at all levels of the nervous system among animals with the same functional behaviours. The property of structurally different elements performing the same function and giving rise to the same phenomenon is termed *degeneracy* (Edelman&Gally, 2001). Given its biological significance, degeneracy has become a heated topic in fields including genetics, evolutionary biology, molecular and cellular neuroscience, and cognitive neuroscience.[1] In recent philosophical literature, degeneracy has been mentioned for its implications for the modularity of causal explanations (Mitchell, 2008; Boone, 2024) as well as its role in distributed cognition (Gillet, 2022). However, attention to degeneracy is lacking from the new mechanist literature. This paper will place degeneracy in the framework of new mechanist philosophy, present a new strategy to explain degenerate mechanisms in the nervous system, and address its implications for our understanding of mechanistic explanations in neuroscience. Note that the following proposal is complementary to the received account of mechanistic explanations – it aims to offer a strategy, as observed from scientific practice, for investigators to develop mechanistic explanatory insight when the default method is impeded by the prevalence of degeneracy.

[1] For genetics, see Gonzalez et al.(2019); for evolutionary biology, see Whitacre (2010); for cellular and systems neuroscience, see Calabrese&Marder (2024); for cognitive neuroscience, see Sajid et al. (2020).

Specifically, in section 2, I will clarify the notion of degeneracy in more detail and argue for its significant role in many explanatory projects in neuroscience. In section 3, I will show how degeneracy poses a problem for the generalisability of mechanistic explanations, and how the problem is usually glossed over by the default strategy, namely building *exemplar-based* mechanistic models using the mean values of experimental data. However, I suggest that this default strategy stumbles when the so-called 'failure of averaging' problem occurs, i.e. when the averaged techniques overlook hidden correlations among the recorded parameters (cf. Golowasch et al., 2002; Marder&Taylor, 2011). In section 4, I introduce what I call the *population-based* mechanistic explanation as an alternative solution that avoids this difficulty. Specifically, many researchers are now using a large population of models to capture the real-life variability of biological mechanisms. By examining a case study from cellular neuroscience (i.e. Lamb&Calabrese, 2013), I will suggest that population-based mechanistic explanations differ from the exemplar-based ones by providing information about all the 'mechanism variants' for the phenomenon and answering different kinds of what-if-things-had-been-different questions. However, in section 5, I argue that this alternative strategy has implications for the new mechanist philosophy. In particular, I argue that the notion 'how-actually models' needs to be re-examined in light of degeneracy.

## 2. The case for degeneracy

Degeneracy, the one-to-many mapping between structures and functions, is recognised as ubiquitous in the nervous system. Just to name a few examples, different ion channel combinations can give rise to the same neurophysiological properties,

different presynaptic and postsynaptic signalling cascades can result in the same synaptic plasticity, and different neuronal populations can produce the same cognitive phenomena (Rathour&Narayanan, 2019). Mechanisms that exhibit degeneracy also have many adaptive advantages. On the top of the list is its contribution to the robustness of biological functions – having a number of different structures capable of producing the same function means that minor perturbations are not likely to have lethal consequences for the organism or the population (Edelman&Gally, 2001). Moreover, degeneracy is argued to be a driving force for biological complexity and adaptive innovation (Whitacre, 2010). Therefore, more and more researchers believe that we cannot fully understand neural mechanisms without taking into account the notion of degeneracy.

How should we understand degeneracy in the context of new mechanist philosophy? Familiarly, Bechtel and Abrahamsen in their seminal work propose that '[a] mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel&Abrahamsen, 2005, p.423; see also, Machamer et al., 2000; Glennan, 2002). The organised components of the mechanism are regarded as at a lower (mechanistic) level than the overall phenomenon (cf. Craver, 2007, chap.5). Given that degeneracy is defined as a one-many mapping between function and structure, I believe that under the new mechanist framework, naturally, degeneracy should be understood as a one-many mapping between the higher-level phenomenon and lower-level mechanistic details. I will use 'degenerate mechanisms' to refer to the mechanisms responsible for the same phenomenon in virtue of different component parts, component operations, or

organisations.[2] Degenerate mechanisms can be found across different individuals or within one individual across time (e.g. in the case of neural plasticity). Note that this paper will *only* focus on the former cases, i.e. degeneracy that occurs among a population of systems (e.g. organisms, cells).[3]

Degeneracy is conceptually closely related to the notion of *multiple realisation (MR)*, which deserves some clarification. Specifically, MR gained attention among philosophers along with the debate on the relationship between mental properties and neural properties, and it is usually associated with philosophical issues such as the autonomy of psychology, anti-reductionism, etc. On the other hand, degeneracy is a notion favoured by biologists and neuroscientists and has a root in quantum physics.[4] Regardless of the different intellectual backgrounds, both terms refer to lower-level variabilities underlying the same higher-level function, hence one might wonder if they are essentially the same concept. However, Figdor (2010) suggests that 'there are cases of degeneracy that do not or may not count as MRs,' and that 'MRs are special, perhaps paradigmatic, cases of degeneracy' (Figdor, 2010, p.428). Specifically, after closely examining the notion of degeneracy in the context of cognitive neuroscience, Figdor points out that degeneracy is broader than MR for it includes cases such as coactivated redundant systems, dual-route cognitive functions, and a more inclusive list of

[2] Note that 'degenerate mechanism' in scientific literature sometimes refers to the mechanism of *how* a system recovers its function from structural change (i.e. the mechanism *of* degeneracy). However, here I use this term to solely refer to mechanisms underlying the same function despite structural heterogeneity (or change) in relevant aspects.

[3] This restriction is made because most of the current scientific works that apply the methodology discussed in this paper deal with cases of degeneracy across individuals. In my view, the same or similar methodologies have potential application for the intra-individual cases (that is, it is possible generate a population of models to represent the mechanistic configurations of the same individual at different time points). However, I believe that such modelling strategies might be less suitable for representing the physiological *changes* that are crucial for maintaining the functional robustness within individuals, hence crucial for intra-individual degeneracy. Future work is needed to address the difference in the mechanistic explanations of inter- and intra-individual degeneracy.

[4] See Mason et al. (2015) for a historical review of the use of the term 'degeneracy.'

structural variations which are not usually considered as cases of MR.[5] Thus, degeneracy and MR seem to be distinct concepts, with the latter being a subset of the former.[6] Nevertheless, although a detailed conceptual comparison is out of the scope of this paper, I believe that discussions on degeneracy could be of interest to the MR debate – specifically, since MR can be understood at least as a special case of degeneracy, my account of how to explain degenerate mechanisms might point to a novel take on how to acknowledge MR while preserving a causal/mechanistic approach to explanation.[7]

Another thing to be clear up front: degeneracy is relative to the grain of description. According to cognitive scientist Uta Noppeney and collaborators, '[s]tructure-function relationships depend implicitly on the descriptive level at which the structural and functional elements are specified' (Noppeney et al., 2004, p.434). To borrow their example, suppose that the cognitive function of wording reading can be performed by multiple degenerate neuronal populations. The attribution of degeneracy

[5] More specifically, according to Figdor (2010), when two degenerate systems are coactivated to realise a single instance of the function, it is still degeneracy but might not count as MR; regarding dual-route cognitive functions (e.g. reading by the semantic route vs. reading by the phonological route), it is not MR at least if cognitive functions are individuated by routes; finally, some variations in neuroanatomical details may appear in the degenerate mapping, but do not constitute MR.

[6] However, note that different philosophical accounts of MR can have more or less overlap with scientists' use of degeneracy. For example, Polger and Shapiro (2016) require that the lower-level variations qualified for MR should be *relevant* to the higher-level function and involve taxonomical difference, which could exclude many cases of degeneracy. They also endorse Figdor's (2010) conceptual distinction and conclude that 'some neuroscientists who hypothesize widespread degeneracy do not distinguish relevant from irrelevant differences' (Polger&Shapiro, 2016, p.134).
On the other hand, Aizawa's and Gillett's account of MR (Gillett, 2003; Aizawa&Gillett, 2009a; 2009b) allows a wider range of lower-level variations to be qualified as cases of MR, which seems to align more closely at least with the cases of degeneracy introduced in this paper. Similarly, Boone offers an analysis of MR in light of causal explanatory frameworks, and argues that '[i]n systems neuroscience, the study of robustness very much is a science of multiple realization' (Boone, 2018, pp.82). The relationship between degeneracy and MR hence in a sense depends on one's favoured characterisation of MR.

[7] I thank an anonymous reviewer for bringing up this connection. See also Boone (2018) for a discussion on MR and causal explanations with cases studies similar to the ones cited in the present paper.

will not hold either if the structure is specified in a more coarse-grained way (e.g. in terms of cortical regions), or if the function specified in a more fine-grained way (e.g. reading via spelling-sound relationships). Hence, we need to talk about degeneracy in a specific and contrastive manner (i.e. whether function P is degenerately sustained by structure $X_1$-$X_n$).[8] However, this description-dependency does not make degeneracy trivial. Although the full criteria for the attribution of degeneracy would be out of the scope of this paper, I argue that not any structural variation can make a case of degeneracy – for one thing, the difference in the color of the neurons does not make the activity of neurons degenerate.

However, given this description-dependency, it seems that researchers can always avoid talking about degeneracy by shifting the granularity of descriptions and splitting the function into fine-grained variants. To be sure, such practice is commonplace in science. To borrow Kaplan's (2017) example, the research project on the phenomenon of animal sound localisation was split into one for sound localisation *in birds* and one for sound localisation *in mammals* when structural difference in the neural circuits of birds and mammals was discovered. Craver (2004) points out that some researchers in cognitive neuroscience even take a step further and assume that it is not only commonplace but also necessary to split the description of the function into more fine-grained ones when structural variations are observed. Nevertheless, as a matter of fact, researchers do not always dismiss degeneracy by shifting descriptions.[9] I believe that there are many reasons for researchers to adopt descriptions that support the attribution of degeneracy and refrain from splitting higher-level phenomena

[8] Polger and Shapiro (2016, Chapter 4) make similar remarks for the attribution of MR.

[9] Examples of researchers embracing degeneracy in cognitive neuroscience include Noppeney et al. (2004); Sajid et al., (2020).

according to lower-level variations. Aizawa and Gillett (2011) argue that whether or not researchers choose to split the higher-level property is determined by both higher- and lower-level theories – considerations from higher-level theories can prevent the splitting in the face of lower-level variability. Moreover, I believe that in many cases, explanatory interest also plays a role in choosing whether to dismiss or embrace degeneracy.[10] Given these considerations, this paper goes under the assumption that phenomena produced by degenerate mechanisms are genuine explanatory targets for scientific endeavours, and I will focus on the cases where researchers do not dismiss degeneracy by shifting descriptive granularity. I believe that such a premise can lead to descriptively adequate philosophical accounts of many explanatory projects in neuroscience.

## 3. How to explain degenerate mechanisms? The default strategy and its limitations

### *3.1 Degeneracy and the generalisability of mechanistic explanations*

In a sense, degeneracy poses a problem for the generalisability of our mechanistic explanations. According to the mainstream new mechanist accounts, a mechanistic explanation (in the epistemic sense) for an explanandum phenomenon explains by representing the mechanistic details (i.e. component parts, component operations, and organisations) of the mechanism underlying the phenomenon (see Bechtel&Abrahamsen, 2005, p.425; Craver, 2007, Chap.4). Such an explanation usually takes the form of a mechanistic model. Importantly, explanandum phenomena

[10] Such explanatory interest can be affected by many factors – for example, when structural variability is discovered across groups with no clear taxonomical distinctions, researchers can be reluctant to describe their function as different ones.

in neuroscience are usually regularities observed in a population of organisms (e.g. a species or a class of animals). It follows that mechanistic explanations should offer *generalisable* claims that apply to all the individuals in the population in question. Degeneracy, on the other hand, threatens the ability of mechanistic explanations to meet this criterion. This is because, due to degeneracy, individuals in a population are heterogeneous in terms of their mechanistic details despite exhibiting the same phenomenon. If a mechanistic model is faithful to one individual, it will diverge from other individuals in the population, hence the challenges the applicability of the model to all instances of the explanandum phenomenon.

Relatedly, degeneracy threatens the generalisability of the explanatory information provided by the mechanistic models. As many new mechanists would agree, explanations provide explanatory information by offering answers to the interventionist what-if-things-had-been-different questions (the 'w-questions' for short). According to Woodward (2003), w-questions are counterfactual questions about how the explananda would change given interventions on the conditions cited in the explanans. The more systematic w-questions it answers, the deeper the explanation. Many new mechanists believe that mechanistic explanations explain by answering w-questions regarding how the phenomenon will change if interventions are conducted on the components of the mechanism (e.g. Craver, 2007, chap.4). Consider the example of explaining how neurons can achieve a low firing rate. It is believed that a low firing frequency is a result of a higher density of ion channels with an inactivating (A-type) potassium conductance, $I_A$. In other words, a mechanistic model of a neuron with low firing rate is supposed to answer the question 'what would happen to the firing rate of the neuron if $I_A$ was changed?' However, Drion and colleagues showed that the same effect could also be produced by unrelated ion channels with a non-inactivating (L-type-like) calcium

conductance, $I_{Ca}$ (Drion et al. 2015). That is, the mechanism for achieving a low firing rate is degenerate, where two different kinds on components can give rise to the same neuronal activity. This means that the effect of intervention on $I_A$ in the membrane will differ from neuron to neuron depending on its specific composition of $I_A$ and $I_{Ca}$ – if a neuron maintains its low firing rate mainly in virtue of $I_{Ca}$, intervention on $I_A$ would have little effect on the neuronal activity. Thus, no general, systematic answer can be provided to the question 'what would happen to the firing rate if $I_A$ was changed?'; the generality of the explanatory information offered by a single mechanistic model is compromised.

In sum, degeneracy challenges the generalisability of our mechanistic explanations. Indeed, one can argue that, if a mechanistic account fails to apply to different instances of the explanandum, it is always an option to propose multiple mechanistic explanations for one phenomenon. However, it can be impractical to offer a tailored mechanistic explanation for each degenerate individual in the population – some degree of generalisability is needed for most, if not all, scientific explanations. Thus, more effective methods are required for explaining degenerate mechanisms.

### *3.2 The default strategy: Exemplar-based mechanistic explanations*

As a common strategy practiced by many investigators and endorsed by most new mechanist philosophers, exemplar mechanistic models are constructed to account for the general explanandum phenomena. According to Bechtel and Abrahamsen (2005), mechanistic models can offer generalised explanations by serving as exemplars or prototypes. Craver and Kaiser echo this account by arguing that 'mechanists deny that an explanatory model must be formulated in terms of generalizations,' and that

'exemplar models … by their very nature describe representative instances rather than general types' (Craver&Kaiser, 2013, p.140-141). In other words, an exemplar model can offer a general mechanistic explanation by representing the mechanistic details of a single prototypical individual (or an idealised individual with prototypical features) that can be extrapolated to similar but non-identical individuals in the population. This is what I call the *exemplar-based* mechanistic explanations.

Such a default treatment can sometimes be applied to degenerate cases. By using exemplar models, the lower-level variability among the population does not prevent (approximate) generalisation as long as the extrapolation is justified.[11] For example, much early research on neural transmission use the squid giant axons to build exemplar models for all neurons, assuming that the mechanisms for neural transmission across different neurons are similar enough. However, rather than embracing degeneracy, this default strategy avoids the talk of degeneracy by focusing on a single exemplary case and downplaying the non-exemplary ones. Thus, exemplar-based mechanistic explanations can at best serve as an expedient method for dealing with degenerate mechanisms.

More importantly for my purpose, there are practical problems when building the exemplar models for degenerate mechanisms. Specifically, what features should be regarded as exemplary? Such a choice can be hard when it comes to quantitative models, where the quantitative values of mechanistic details are relevant to the phenomenon of interest. For example, the quantitative features of neuronal activities (e.g. spike rates) are determined by the maximal conductances of ion channels on the membrane. An

[11] As for the justification of such extrapolation, Craver and Kaiser suggest that the comparative studies described by Steel (2008) and phylogenetic proximity between organisms can serve as potential grounds for the generalisation.

exemplar model for spiking neurons should hence specify the values of parameters representing the maximal conductances. But what parameter values are exemplary for a degenerate population? According to Marder and Taylor (2011), in the past, some researchers chose their 'best' recording – that is, the fastest and largest currents – to fit the exemplar model. However, due to the unreliability of using single measurement, more researchers prefer to record from multiple preparations, *average* the measurement results, and calculate the mean value to fit the models. As we all know, averaging is perhaps one of the most commonplace and standard practices for model building in neuroscience.[12] However, the use of averaging techniques conceals the underlying variability of the individuals. As Prinz and colleagues argues, 'successful as these strategies [i.e. averaging] may be, they implicitly assume that variability between preparations or animals is "experimental noise" rather than an essential characteristic of the nervous system' (Prinz et al., 2004, p.1345).

In short, exemplar-based mechanistic explanations usually employ averaging techniques to idealise away degeneracy by blotting out the variability across individuals. However, as we know, models are bound to have idealisations. How much of a problem can the method of averaging bring about?

### *3.3 Failure of averaging: the limitations of exemplar-based mechanistic explanations*

I suggest that the use of averaging for exemplar-based mechanistic explanations is bugged by what is called the *'failure of averaging'* problem. Specifically, as more and more researchers in cellular and systems neuroscience recognise, averaging is blind

[12] For example, Hodgkin and Huxley derived their classic model of action potential by averaging over the measurement results from multiple axon preparations (Hodgkin&Huxley, 1952, p.435).

to hidden correlations between measured quantities, which could result in misrepresentation of the mechanism that produces the phenomenon.

In particular, Golowasch, Goldman, and Marder (2002) conducted a modelling study to illustrate this possibility. These researchers constructed a conductance-based model neuron with five voltage-dependent conductance parameters. To mimic the biological variability, they randomly varied the maximal conductances and generated a number of bursting neurons with the same behaviour: firing a single action potential at the peak of a slow membrane potential depolarization (i.e. one-spike bursters). As a surprising result, the model constructed using the mean of the parameter values of this population of neurons is *not* itself a one-spike burster – that is, the 'typical' exemplar model constructed using the mean values does not itself produce the target phenomenon. Golowasch and collaborators suggest that this result is due to a hidden correlation between two maximal conductances, which is not reflected by the averaged model. In a multi-dimensional parameter space, this correlation gives rise to a L-shaped distribution of this population of neurons, and the mean value of the parameters hence does *not* fall within the region that generates the target phenomenon (see Fig. 2 for illustration).

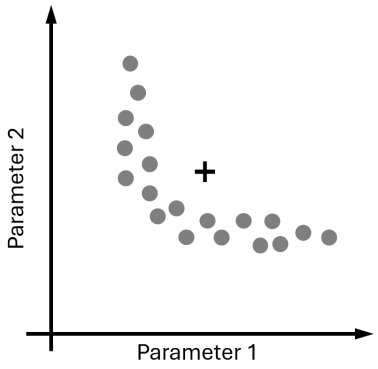


Fig.2 Schematic illustration of the failure of averaging. Dots represent individual neurons that exhibit the target behaviour. The cross represents the mean of the distribution. The neurons lie in an L-shaped region that does not include the model constructed with the mean values. After Fig.3a from Golowasch et al., (2002, p.1130) and Fig.2b from Marder&Taylor (2011, p,135).

The researchers hence conclude that '[a]veraging will fail whenever the mean of the distribution of relevant data points lies outside the region they occupy. Standard statistical measures do not indicate when this occurs because they do a poor job of characterizing the boundary of a region. However, scatter plots of the relevant parameters, as used here, should be sufficient to reveal a failure of averaging by showing regional boundaries' (Golowasch et al., 2002, p.1131). In other words, averaging techniques sometimes overlook important correlations among parameters and hence misrepresent salient features of mechanisms. On the other hand, drawing a scatter plot representing all individuals in the population can reveal hidden correlations, but it is at odds with the idea of exemplar-based mechanistic explanations, which by definition only represents single individuals. Of course, to avoid the failure of averaging, an exemplar model can always abstract away from the precise quantitative properties that are specific to individuals and only represent general qualitative feature (e.g. to model the neuron as having *some* sodium conductance). However, this abstraction would prohibit the model from offering further mechanistic details and answering more w-questions, which can lead to a loss in explanatory power of the model (see Section 5).

Therefore, my contention is that, in some cases – specifically, in cases where quantitative parameters are crucial to the mechanism in question, the mean value of the parameters are regarded as the exemplary features, and certain correlations exist among these parameters across individuals – building exemplar models cannot offer mechanistic explanations for degenerate mechanisms. How pervasive are these cases? In current scientific practice, the majority of models are featured with quantitative parameters; there is usually no obvious choice for the exemplary parameter values other

than the mean values. However, indeed, those problematic correlations are only found in some systems. Therefore, not all cases of averaging will fail, and exemplar-based mechanistic explanations still work for many research projects. However, importantly, without checking this possibility – without drawing a scatter plot showing all individuals in the population – it is unknown whether the system in question hold such correlations among the parameters or not. Hence, the failure of averaging impairs the reliability of all unchecked exemplar models. Therefore, successful as they are, the exemplar-based strategies leave room for alternatives. In fact, many researchers are aware of this limitation and now opt for an alternative solution to deal with degeneracy – that is, to offer what I call *population-based mechanistic explanations*.

## 4. Population-based mechanistic explanations

Recently, many researchers working on small neuronal circuits are starting to construct populations of models instead of single exemplar models to capture the real-life variability of biological mechanisms (cf. Prinz, 2010; Marder&Taylor, 2011). Specifically, a large number (usually thousands or even millions) of model instances are generated to represent individual neurons or individual neuronal circuits with the same behaviour or activity pattern. In other words, rather than idealising away the variability, this population-based modelling approach accommodates numerous distinct instances and represents a degenerate population, hence it avoids the failure of averaging. Moreover, as will be clearer later, such a modelling approach can offer mechanistic explanations – what I call population-based mechanistic explanations – for the behaviour of a population of neurons or neuronal circuits despite their lower-level variability, hence it preserves the generalisability of the explanations.

Note that according to this population-based modelling approach, researchers do not usually use actual experimental data for each model instance in the population due to the impracticality of experimentally recording millions of preparations. Instead, to mimic experimental data, they generate the population by assigning randomised parameter values to model instances and select those instances that exhibit the target behaviour. In such a model population, each model instance is defined by a distinct set of parameter values, occupying a point in the multi-dimensional parameter space, while representing a unique individual. The region in the parameter space where instances show the target behaviour is called the *'solution space.'* Thus, the solution space at least approximately represents the population of all the individuals exhibiting the explanandum phenomenon.[13] A general goal for the researchers who adopt this population-based approach is to figure out the shape and the contour of the solution space. There are many different ways to explore the solution space. For example, researchers can vary the parameters with a certain step size and analyse all the resulting models (that is, solution space exploration on a grid, see Fig. 3, left). More recently, multi-objective evolutionary algorithms are proven efficient in generating good model instances (see Fig. 3, right). Inspired by natural selection, these evolutionary algorithms 'breed' model instances exhibiting good performance to derive new generations of model instances that have better performance for the target phenomenon (Smolinski&Prinz, 2009).

[13] I acknowledge that the solution space can only be an *approximate* representation of the population in questions. For one thing, the population of models, however densely sampled, can hardly exhaust all the parameter combinations for the target behaviour. Moreover, the multi-dimensional parameter space does not usually reflect the possibility of, say, two individuals having exactly the same set of parameters, which is another idealisation.

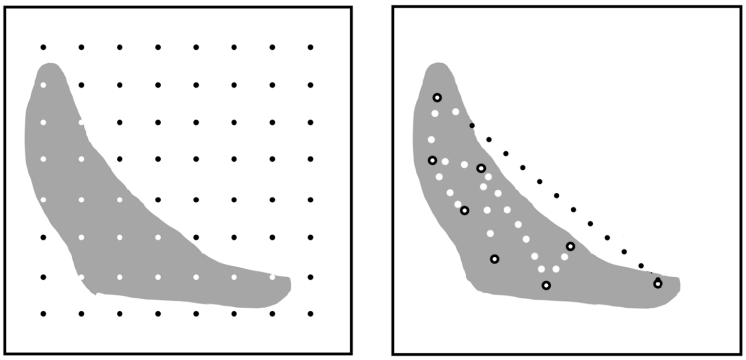

Fig.3 Schematic illustrations showing solution space exploration. White boxes represent the parameter space. Grey areas represent the solution space. Dots represent model instances, including solutions (white) or non-solutions (black). *Left*: solution space mapping on a grid. *Right*: solution space exploration using evolutionary algorithms, where solution space can be further explored by hyperplanes spanned by subsets of results (dots with white centre and black outline). After Fig.4d from Prinz (2010, p.2402).

An example can better illustrate the idea. Lamb and Calabrese (2013) employed the population-based modelling approach to account for the activity of heart (HE) motor neurons in the leech. The HE motor neurons in the leech exhibit signature rhythmic activities that are largely the same across individuals, whereas the intrinsic membrane properties of these neurons – mainly the maximal conductances determined by the density of voltage-gated ion channels – show significant animal-to-animal variability. To study the contribution of different maximal conductances ($\bar{g}$) to the activity pattern, Lamb and Calabrese developed a multi-compartmental Hodgkin-Huxley style model for the HE motor neurons, parameterised by eight different maximal conductances. The target neuronal activity was defined by a complete input-output data set recorded from one preparation, with simultaneous recording of the activities of premotor interneurons and HE motor neurons as well as their synaptic weight. To find the parameter values that produce the target activity, Lamb and Calabrese used multi-objective evolutionary algorithms to generate model neuron instances and evaluate them against a list of output activity metrics (including measures of spike frequency, duty cycle, phase, etc.). As a result, a population of more than 40,000 model instances of the HE motor neurons were

constructed. All of these model instances have good performance of the target activity and differ in their maximal conductance parameters, hence constitute the solution space. The researchers believed that this population of model neurons captures the key functional characteristics and structural variability of actual HE motor neurons, hence can serve to provide further explanatory insights.

We can now turn to the important questions: What kind of explanatory insights does this population-based modelling approach offer? First, recall that degeneracy hampers the explanatory power of mechanistic models by threatening the generalisability of the answers to w-questions regarding interventions on single components (section 3.1). Specifically, in the example above, if one asks, 'how would the spike frequency of the HE neurons change if the maximal conductance of sodium channels ($\bar{g}_{Na}$) is increased by 20%?', no generalisable answer can be offered for the whole population of HE motor neurons. This is because, given their variability, some neurons in the population might have abundant sodium channels, whereas some have very few. Hence, the same intervention on the sodium channels would result in different quantitative effects on the phenomenon in different neurons. Granted, some coarse-grained, qualitative generalisations can be proposed (e.g. 'the spike frequency will *more or less* increase if $\bar{g}_{Na}$ is increased'). However, such answers to the w-questions are less systematic and hence 'shallow' in terms of the explanatory insight they provide.

A population of models, on the other hand, can yield the distributions of parameter values in numerous individuals constituting a solution space, which allows us to gain different explanatory insight. In particular, after successfully explored the solution space with the population of models**,** Lamb and Calabrese turn their attention to 'how the model instances were distributed in parameter space, or how the parameter

values for successful model instances were related to each other' (Lamb&Calabrese, 2013, p.10) In other words, they aimed to find *patterns or regularities* among all the instances in the population. Specifically, they evaluated the partial correlations [14] between each pair of parameters, and showed that, roughly, many outward conductance parameters (e.g. $\bar{g}_{K2}$, $\bar{g}_{KA}$, $\bar{g}_{K1}$) were positively correlated with inward conductance parameters (e.g. $\bar{g}_{P}$, $\bar{g}_{Na}$), whereas outward conductance parameters were negatively correlated with each other, and so were the inward conductances. Moreover, manipulationist experiments were conducted on the whole population to determine (qualitatively) *the functional role of each conductance*. Lamb and Calabrese perturbed each parameter by ±25% and ±50% in more than 1000 model instances, and observed that, roughly, increases in outward conductances reduced spike frequency and duty cycle while inward conductances had the opposite effects. They hence concluded that '[c]onductances that had positive correlations opposed one another and had the opposite effects on activity metrics when perturbed, whereas conductances that had negative correlations could compensate for one another and had similar effects on activity metrics.' Thus, they further inferred that *co-regulation* (i.e. regulatory factors that modify multiply conductances at the same time) might be in place to sustain the robustness of neuronal activities in HE motor neurons and further explain why there are such correlations between conductances.

In short, in their population-based modelling study, Lamb and Calabrese (a) presupposed and represented the variability in maximal conductances underlying the activity of HE motor neurons, and (b) provided information about all the parameter values that can produce the activity by exploring the solution space. Such information

[14] The partial correlation between two parameters is the correlation after compensating for the remaining parameters with a linear model.

involves (i) regularities among all the model instances, (ii) a functional interpretation of the parameters, and (iii) further insights on why there are such regularities in the solution space (i.e. the co-regulation). More generally, I propose that, ideally, a population-based mechanistic explanation of a phenomenon

> (a) presupposes and represents the variability in the mechanisms underlying the phenomenon, and (b) provides information on all the *mechanism variants* of the phenomenon, which usually involves (i) information about the regularities among these mechanism variants, (ii) information on the functional role of component type variants, and (iii) further explanations for the regularities.

Condition (a), I believe, is what fundamentally distinguishes population-based mechanistic explanations from 'ordinary' exemplar-based mechanistic explanations: the former explicitly embraces variability and represent a plurality of mechanisms underlying the same phenomenon, whereas the latter only offers singular exemplars that sweep degeneracy under the rug. In condition (b), what I call a 'mechanism variant' is one that underlies at least one instance of the explanandum phenomenon. As mentioned, all the mechanism variants are approximately represented by the solution space. A population-based mechanistic explanation counts as a mechanistic explanation because it sheds light on how the organised mechanistic component parts and operations give rise to the phenomenon by specifying the mechanism variants and providing information (i)-(iii) listed in condition (b).[15]

How does this information differ from the explanation provided by the exemplar-based alternative? I argue that population-based mechanistic explanations can deepen explanatory insight by offering answers to different kinds of w-questions.

[15] Note that a population-based mechanistic explanation might only provide some but not all of this information; the information provided might also be incomplete.

First, they can answer individual-level w-questions (i.e. w-questions specific to individuals) for all individuals in the population. For example, one can answer whether a certain HE motor neuron can maintain its activity given perturbations on its conductances based on whether its parameter values stay in the solution space. But what about the more generalised 'population-level' w-questions? I suggest that population mechanistic explanations at least have the potential to offer generalised insights regarding the relation between the phenomenon and *multiple* types of components. For example, recall that Lamb's and Calabrese's results suggest that several outward conductances can compensate for one another in neurons with the same activity. Hence, although single outward conductances contribute differently in different neurons, it might be the case that multiple outward conductances as a group have a definitive contribution to the neuronal activity that is universal across neuron instances. In other words, a stable, linear dependency relation might hold between the phenomenon and the sum of the contribution of several component types.[16] Lamb and Calabrese's model population can hence (at least potentially) offer generalisable answers to questions such as 'how would the neuronal activity change if a group of outward conductances is changed'. More generally, population-based mechanistic explanations can (at least potentially) offer generalisable answers to w-questions with the form 'how would the phenomenon change if a group of components is changed?' By providing systematic answers to these w-questions, population-based mechanistic explanations deepen the explanatory insight into degenerate mechanisms.

However, I acknowledge that the modelling strategy presented above might only apply to some but not all cases of degenerate mechanisms. Specifically, due to the

[16] This echoes Aizawa's (2013) idea of multiple realization by compensatory differences.

limitation of computation power, the members in the model population are usually inherently similar to each other; they might have to share a common 'skeleton model' with a limited degree of freedom. Thus, if the degenerate mechanisms have more 'eccentric' or more complex variations, they might not be captured by such a model population. Moreover, in order to represent all the variations in terms of the model parameters, we must have enough prior knowledge about the mechanisms. This might also require the mechanisms in questions to be relatively simple. Thus, the population-based mechanistic explanations might only work for a class of degenerate mechanisms, namely degenerate mechanisms that are relatively simple and well-understood with more tractable variations. Nevertheless, population-based mechanistic explanation as a burgeoning explanatory strategy has been employed in many research projects in cellular and molecular neuroscience (e.g. Prinz et al., 2004; Caplan et al., 2014; Alonso&Marder, 2020). With the rapid development in modelling techniques, I believe that its application can only become wider.

In sum, in this section I introduced population-based mechanistic explanations as a novel strategy for dealing with degeneracy in neural systems. Instead of dismissing degeneracy or idealising it away, such an explanatory strategy embraces the lower-level variability and offers explanatory insights on all the mechanism variants underlying the phenomenon. By doing so, researchers can deepen the explanation while preserving the generalisability of the explanation. However, I argue that such an explanatory scheme has consequences for our understanding of mechanistic explanations. In the next section, I will explore some of such implications.

## 5. Implications for the 'how-actually' mechanistic model

Current discussions in the new mechanist literature are mostly based on the default, exemplar-based view of mechanistic explanations. For instance, Craver (2006, 2007, see also Machamer et al. 2000) famously argues that mechanistic models should be '*how-actually models*,' which represent the real mechanisms underlying the explanandum phenomenon. Such a normative claim is endorsed or debated by many new mechanist authors (e.g. Craver, 2014; Povich, 2015; 2018; Brainard, 2020). However, examples of how-actually models in this literature are mostly, if not all, exemplar models. I believe that the notion 'how-actually model' needs to be reconsidered in light of degeneracy and population-based mechanistic explanations. Rather than posing objections to the how-actually-vs.-how-possibly distinction, I want to raise some caveat about the use of the term 'how-actually models.' Specifically, I argue that the model instances representing the unique individuals in a population are more entitled to be termed 'how-actually' than the exemplar models with the same grain of description derived from averaging techniques. However, the target of a model population should not be confined with the actual mechanism instances currently existing in the world, but encompass non-actual, past or potentially future instances. Thus, there is more nuance to what it means when we argue that good mechanistic models should be how-actually models.

According to Craver, a how-actually model represents (or approximately represents) 'real components, activities, and organizational features of the mechanism that in fact produces the phenomenon' (Craver, 2007, p.112). In contrast, how-possibly models might 'have explanatory purport, but they are only loosely constrained conjectures about the sort of mechanism that might suffice to produce the explanandum phenomenon' (Craver, 2007, p.112). According to the proponents of this how-actually vs. how-possibly distinction, the former has clear advantage for mechanistic

explanations because they offer reliable information about the mechanism in question. Based on this distinction, I believe that exemplar models with parameter values fixed by averaging techniques are hardly ever qualified as how-actually models. In particular, as introduced in section 3, averaging techniques are usually applied under the assumption that the mean value is the representative parameter value of the target mechanisms in the population in question. However, such an assumption is usually unsupported by evidence – that is, it is a loosely constrained conjecture. In other words, this assumption is a how-possibly element featuring in such exemplar models. And as a result, the failure of averaging occurs when this conjecture misfires – that is, the problem occurs when what was believed to be possible (or even plausible) fails to capture how the mechanisms underlying the phenomenon actually are. The model instances in a model population, on the other hand, avoid this pitfall by implementing parameter values that can actually produce the target phenomenon and hence providing reliable mechanistic information. Thus, I believe that the model instances score better than the said exemplar models on the how-actually-vs.-how-possibly axis.[17]

However, a new mechanist might turn her attention to models with a coarser grain of description. In particular, as indicated in section 3, if a model refrains from specifying the quantitative details in the mechanism – that is, if it refrains from specifying parameter values and only contains qualitative descriptions – no averaging technique will be needed, hence the loosely constrained conjecture mentioned above can be avoided. Moreover, such a coarse-grained, abstract model can be true of all

[17] This result should also be expected from the perspective of Darwinian theories. Specifically, according to Mayr (2004), Darwinian theories of evolution introduced a new mode of thinking – namely, the Darwinian 'population thinking'. This mode of thinking emphasises the uniqueness and variability of every individual in a population. According to the Darwinian population thinking, 'it is this variation among the uniquely different individuals that has *reality*, while the calculated statistical mean value of this variation is an abstraction' (Mayr, 2004, p.88, original emphasis). In other words, only the particular individuals, not essentialist archetypes, are real or actual.

individuals in the target population, hence qualify as a how-actually model without compromising its scope. However, recall that Craver (2007; also Machamer et al. 2000) makes another normative distinction, namely the distinction between mechanism sketch and complete mechanistic models. In particular, a mechanism sketch is an incomplete mechanistic model that leaves important details unspecified, whereas a complete mechanistic model fills in all such details. I believe that a model that sacrifices quantitative details are likely to score lower on the axis of mechanism sketch vs. complete mechanistic models. This is because, as shown in my case studies, quantitative details are often relevant to the explanandum phenomena that modern neuroscience is interested in. In other words, although a model can stay how-actually by abstracting away from certain details, it might do so at the cost of its completeness. Population-based modelling strategy on the other hand, is capable of better achieving both actuality and completeness by representing the population with a large number of models.

However, the sense of actuality in the context of model populations deserves further examination. Consider a population of models of the activity of leech HE motor neurons. Suppose that a model instance *M* captures the parameter values of an actual, living HE motor neuron *N*, hence represents an actual mechanism instance underlying an instance of explanandum phenomenon. When neuron *N* dies and ceases to exist, the parameter values are no longer instantiated by any actual living neuron, hence *M* no longer represents an actual mechanism instance underlying the phenomenon. However, I believe it would be implausible to say that *M* suddenly becomes a misrepresentation and should be excluded from the solution space of the model population. This is because, although neuron *N* no longer exists, the mechanism variant represented by model *M* still counts as a 'solution' for the phenomenon, which might be instantiated by other

individuals in the population again. The same argument applies to potential future neurons in the population: suppose that a model instance represents the features of a neuron that will exist and produce and phenomenon in the same population some time in the future; it seems implausible to say that such a model instance is only valid when the potential future neuron embodying the features of the model is brought to life.

Therefore, if a how-actually model is understood as one that represents the actual mechanistic details of a system that exists *at the present time*, then members of a model population in a population-based mechanistic explanation are not necessarily how-actually models. Rather, the mechanistic configurations that used to be instantiated in the past and the ones that are potentially instantiated in the future are all valid target for the model population. In my view, this extension stems from the scope of scientific explanations. In particular, when researchers want to explain, say, the heart rhythm in the leech, it seems that they are not only interested in the currently living leeches – since if this is the case, we should renew our explanatory accounts every several years for each new generation of leeches. Rather, the explanations they seek should also be applicable to past and future organisms of the same species. More generally speaking, to the extent that mechanistic explanations have scope (i.e. are generalizations), they apply to a wider range of instances, some of which are actual, some non-actual (i.e. past or future). Such a distinction is masked by the assumption of a homogenous population. However, thinking about degeneracy can reveal this gap, since the past and future mechanism instances might diverge from the present ones.

In sum, model instances in a population-based mechanistic explanation are promising candidates for how-actually models, if actuality is understood in terms of how much loosely constrained conjecture about the phenomenon is involved. However,

if actuality is understood in terms of existence at the present time, then it seems problematic to require all the model instances in a model population to be actual. Of course, a new mechanist might broaden the talks of 'how-actually models' to accommodate both past and potential future mechanisms – perhaps by endorsing a growing block theory of time and some determinist assumptions. However, this option would still call for careful examination of what counts as actual for mechanistic models and how we can distinguish them from how-possibly models. While I do not intend to raise objections against the how-actually-vs.-how-possibly distinction, I do invite new mechanist philosophers to reassess their framework in light of the ubiquity and importance of degeneracy among biological populations.

## 6. Conclusion

In this paper, I discussed the notion of degeneracy in the framework of new mechanist philosophy, presented a new strategy to explain degenerate mechanisms in the nervous system, and addressed its implications for our understanding of mechanistic models in neuroscience. Specifically, I argued that the standard, exemplar-based mechanistic explanations are sometimes beset by the failure of averaging problem. I then proposed an alternative strategy for researchers to explain degenerate mechanisms, namely, to offer what I call population-based mechanistic explanations which provide information of all the mechanism variants underlying the explanandum phenomena. Finally, I suggested that the use of population-based mechanistic explanation has implications for the notion 'how-actually models' in the new mechanist literature. To be sure, with the increasing scientific attention paid to degeneracy in the nervous system and the rapid development of modelling technologies, the application of population-

based mechanistic explanations and similar strategies can only be on the rise. I believe that the above account serves as an adequate complement to the new mechanist philosophy that reflects this rising trend in scientific practice.

**Acknowledgements**

For helpful comments on this article, many thanks to Kenneth Aizawa, Mazviita Chirimuuta, Mark Sprevak, the audience at the 2024 meeting for the International Society for the Philosophy of the Sciences of the Mind (ISPSM), and my anonymous reviewers.

**Reference**

Aizawa, K., & Gillett, C. (2009a). *Levels, individual variation, and massive multiple realization in neurobiology*.

Aizawa, K., & Gillett, C. (2009b). The (multiple) realization of psychological and other properties in the sciences. *Mind & Language*, *24*(2), 181–208.

Aizawa, K., & Gillett, C. (2011). The autonomy of psychology in the age of neuroscience. *Causality in the Sciences*, *6*, 202–223.

Alonso, L. M., & Marder, E. (2020). Temperature compensation in a small rhythmic circuit. *Elife*, *9*, e55470.

Baumgartner, M., & Gebharter, A. (2016). Constitutive relevance, mutual manipulability, and fat-handedness. *The British Journal for the Philosophy of Science*.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press.

Boone, T. (2018). Multiple Realization and Robustness. In M. Bertolaso, S. Caianiello, & E. Serrelli (Eds), *Biological Robustness: Emerging Perspectives from within the Life Sciences* (pp. 75–94). Springer International Publishing. https://doi.org/10.1007/978-3-030-01198-7_4

Boone, T. (2024). Robustness and modularity. *The British Journal for the Philosophy of Science*, *75*(2), 417–442.

Brainard, L. (2020). How to explain how-possibly. *Philosophers' Imprint*, 20 (13):1-23.
Calabrese, R. L., & Marder, E. (2024). Degenerate Neuronal and Circuit Mechanisms Important for Generating Rhythmic Motor Patterns. *Physiological Reviews*.
Caplan, J. S., Williams, A. H., & Marder, E. (2014). Many parameter sets in a multicompartment model oscillator are robust to temperature perturbations. *Journal of Neuroscience*, *34*(14), 4963–4975.
Couch, M. B. (2011). Mechanisms and constitutive relevance. *Synthese*, *183*, 375–388.
Craver, C. F. (2004). Dissociable realization and kind splitting. *Philosophy of Science*, *71*(5), 960–971.
Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376. https://doi.org/10.1007/s11229-006-9097-x
Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.
Craver, C. F. (2014). The ontic account of scientific explanation. *Explanation in the special sciences: The case of biology and history*, 27-52.
Craver, C. F., Glennan, S., & Povich, M. (2021). Constitutive relevance & mutual manipulability revisited. *Synthese*, *199*(3), 8807–8828.
Craver, C. F., & Kaiser, M. I. (2013). Mechanisms and laws: Clarifying the debate. *Mechanism and Causality in Biology and Economics*, 125–145.
Drion, G., O'Leary, T., & Marder, E. (2015). Ion channel degeneracy enables robust and tunable neuronal firing rates. *Proceedings of the National Academy of Sciences*, *112*(38), E5361–E5370.
Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763–13768.
Figdor, C. (2010). Neuroscience and the Multiple Realization of Cognitive Functions. *Philosophy of Science*, *77*(3), 419–456. Cambridge Core. https://doi.org/10.1086/652964
Gillett, A. J. (2022). Development, Resilience Engineering, Degeneracy, and Cognitive Practices. *Review of Philosophy and Psychology*, *13*(3), 645–664. https://doi.org/10.1007/s13164-021-00550-9
Gillett, C. (2003). The metaphysics of realization, multiple realizability, and the special sciences. *The Journal of Philosophy*, *100*(11), 591–603.
Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(S3), S342–S353.
Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.
Golowasch, J., Goldman, M. S., Abbott, L. F., & Marder, E. (2002). Failure of averaging in the construction of a conductance-based neuron model. *Journal of Neurophysiology*, *87*(2), 1129–1131.
Gonzalez, D. L., Giannerini, S., & Rosa, R. (2019). On the origin of degeneracy in the genetic code. *Interface Focus*, *9*(6), 20190038. https://doi.org/10.1098/rsfs.2019.0038
Hodgkin, A. L., Huxley, A. F., & Katz, B. (1952). Measurement of current-voltage relations in the membrane of the giant axon of Loligo. *The Journal of Physiology*, *116*(4), 424.
Kaplan, D. M. (2017). Neural computation, multiple realizability, and the prospects for mechanistic explanation. *Explanation and Integration in Mind and Brain Science*, 164–189.

Lamb, D. G., & Calabrese, R. L. (2013). Correlated conductance parameters in leech heart motor neurons contribute to motor pattern formation. *PLoS One*, *8*(11), e79267.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. JSTOR.
Marder, E., Goeritz, M. L., & Otopalik, A. G. (2015). Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Current Opinion in Neurobiology*, *31*, 156–163.
Marder, E., & Taylor, A. L. (2011). Multiple models to capture the variability in biological neurons and networks. *Nature Neuroscience*, *14*(2), 133–138.
Mason, P. H., Winter, B., & Grignolio, A. (2015). Hidden in plain view: Degeneracy in complex systems. *Biosystems*, *128*, 1–8.
Mayr, E. (2004). *What makes biology unique?: Considerations on the autonomy of a scientific discipline*. Cambridge University Press.
Mitchell, S. D. (2008). Exporting causal knowledge in evolutionary and developmental biology. *Philosophy of Science*, *75*(5), 697–706.
Noppeney, U., Friston, K. J., & Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy*, *205*(6), 433–442.
Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford University Press.
Povich, M. (2015). Mechanisms and model-based functional magnetic resonance imaging. *Philosophy of Science*, *82*(5), 1035–1046.
Povich, M. (2018). Minimal models and the generalized ontic conception of scientific explanation. *The British Journal for the Philosophy of Science*, 69:1, 117-137
Prinz, A. A. (2010). Computational approaches to neuronal network analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1551), 2397–2405.
Prinz, A. A., Bucher, D., & Marder, E. (2004). Similar network activity from disparate circuit parameters. *Nature Neuroscience*, *7*(12), 1345–1352.
Prychitko, E. (2021). The causal situationist account of constitutive relevance. *Synthese*, *198*(2), 1829–1843.
Rathour, R. K., & Narayanan, R. (2019). Degeneracy in hippocampal physiology and plasticity. *Hippocampus*, *29*(10), 980–1022.
Sajid, N., Parr, T., Hope, T. M., Price, C. J., & Friston, K. J. (2020). Degeneracy and redundancy in active inference. *Cerebral Cortex*, *30*(11), 5750–5766.
Smolinski, T. G., & Prinz, A. A. (2009). Multi-objective evolutionary algorithms for model neuron parameter value selection matching biological behavior under different simulation scenarios. *BMC Neuroscience*, *10*(Suppl 1), P260.
Steel, D. (2007). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.
Whitacre, J. M. (2010). Degeneracy: A link between evolvability, robustness and complexity in biological systems. *Theoretical Biology and Medical Modelling*, *7*(1), 6. https://doi.org/10.1186/1742-4682-7-6
Whitacre, J., & Bender, A. (2010). Degeneracy: A design principle for achieving robustness and evolvability. *Journal of Theoretical Biology*, 263(1), 143–153. https://doi.org/10.1016/j.jtbi.2009.11.008
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford university press.