NEURONAL CAUSES AND BEHAVIOURAL EFFECTS:

a Review on Logical, Methodological, and Technical Issues With Respect to Causal Explanations of Behaviour in Neuroscience

A PREPRINT

Kayson Fakhar*[1] and Lisa Schmors,[2]Dominic Gonschorek,[2]Natalia Z. Bielczyk[3]

[1]*Institute of Computational Neuroscience, University Medical Centre Eppendorf, Hamburg University, Hamburg, Germany | kayson.fakhar@gmail.com*

[2]*Department of Neuroscience, University of Oldenburg, Oldenburg, Germany*

[3]*Stichting Solaris Onderzoek en Ontwikkeling, Nijmegen, the Netherlands*

August 19, 2019

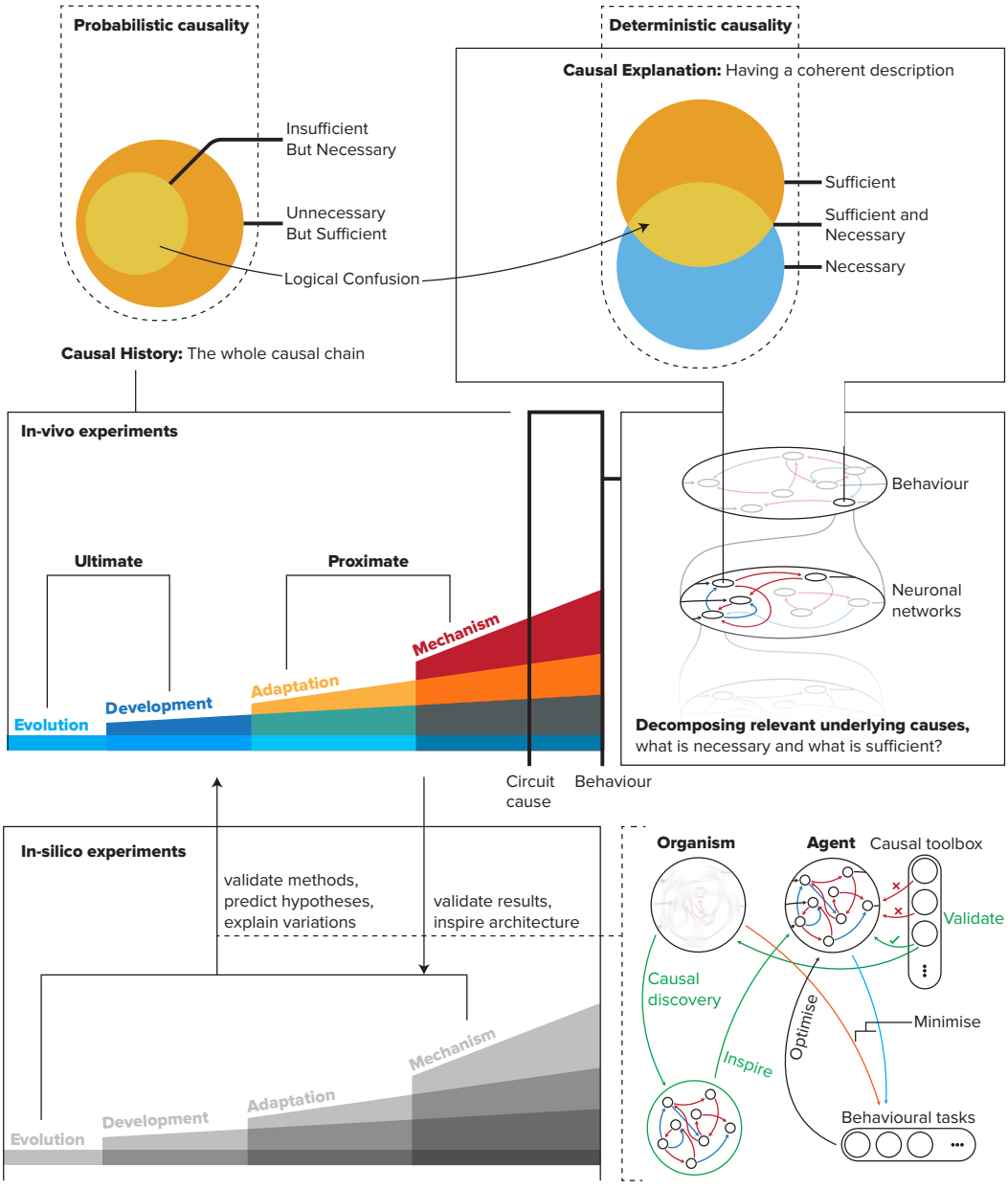
Abstract

Elucidating causal, neurobiological underpinnings of behaviour is an ultimate goal of every neuroscientific study. However, due to the complexity of the brain as well as the complexity of the human environment, finding a causal architecture that underlies behaviour remains a formidable challenge. In this manuscript, we review the logical and conceptual issues with respect to causal research in neuroscience.

First, we review the state of the art interventional and computational approaches to infer causal brain-behaviour relationships. We provide an overview of potential issues, flaws, and confounds in these studies. We conclude that studies on the causal structure underlying behaviour should be performed by accumulating evidence coming from several lines of experimental and modelling studies. Lastly, we also propose computational models including artificial neuronal networks and simulated animats as a potential breakthrough to causal brain-behaviour investigations.

Keywords: causality, cause and effect, Necessity and Sufficiency, causal inference, brain interventions, computational models, brain-behaviour relations

Visual abstract



1 Introduction

The notion of "*cause and effect*" plays a fundamental role in the understanding of our environment. This specifically applies to scientific endeavour and discoveries related to the mechanics of the physical world. However, understanding the realm of complex systems, especially biological ones, is vast enough that it can easily confuse us. Many of the scientific fields, such as cognitive science, psychology, ethology, and neuroscience, employ the question of "*what causes a specific behaviour in a living organism?*" For example, searching for causes of a relatively well-investigated phenomenon such as storing information in a biological system can be found in multiple spatial scales from molecules to neuronal circuits and multiple temporal scales from few milliseconds to millions of years (Josselyn, Köhler, & Frankland, 2015; Bielczyk, Buitelaar, Glennon, & Tiesinga, 2015).

To capture such a complex causal structure in one framework, Ernst Mayr (Mayr, 1961) advocated the separation of causes into two major categories of *ultimate* complex causal and *proximate* causes. Proximate causes can explain the immediate causes, such as neuronal events while ultimate causes aim to explain more distant causes such as the evolution of the behaviour. Inspired by Aristotle's four causes, this account was further elaborated by Niko Tinbergen, dividing each of the categories into a static/dynamic dichotomy (Tinbergen, 1963). In other words, one should consider four causes with respect to behaviour, as an example, aggression:

- **Proximate - dynamic cause (ontogeny):** that explains the *individual's* developmental history of aggressive tendencies, such as what was the role of the mother's hormonal state during the pregnancy?

- **Proximate - static cause (mechanism):** that elucidates the *individual's* underlying neurobiological processes like which brain circuits are responsible for aggressive behaviour?

- **Ultimate - dynamic cause (phylogeny):** that reveals the *species'* evolutionary history, for example, of how aggressive traits led to an increased rate of survival.

- **Ultimate - static cause (adaptation):** that describes the function of the observed behaviour with respect to the current environment like how aggression can be an adaptive strategy to cope with the environmental forces.

Even though the mentioned framework facilitates the organisation of the scientific findings in a way that builds a more comprehensive causal explanation for behaviour, there are several weaknesses of this approach. We will review these issues in the following section.

2 Logical reasoning and conceptual foundations of causal inference

Besides the overemphasis of the proximate factors that lead to a so-called *reductionistic bias* (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017), one major confusion arises from how we define and subsequently establish a proximate causal relationship in laboratories. As interventionists, we mainly infer such a link by manipulating the cause and measuring its influence over the produced effect (J. Woodward, 2009). Again, this manipulation might happen in different scales from gene knockout experiments (Gavériaux-Ruff & Kieffer, 2007) to pharmacological interventions (Afraz, Boyden, & DiCarlo, 2015) and the brain's regional deactivation using non-invasive tools (Parkin, Ekhtiari, & Walsh, 2015). What is often concluded afterwards is that the target of intervention has a causal role in the occurrence and/or quality of the observed behavioural changes.

2.1 Levels of organisation: Necessary and sufficient causes

So far, an intuitive but powerful approach known as "*Necessity and Sufficiency*" conditions (NS) was used to draw causal conclusions in behavioural studies (Gomez-Marin, 2017): if deactivating the target leads to the deactivation or malfunctioning of the effect, the target is assumed to be a *necessary cause*. And, if we can observe the effect by merely activating the target then the target is assumed to be a *sufficient cause*. Consequently, a sufficient and necessary cause is a cause that has both properties.

Although this line of logical reasoning has been extensively used, helping to construct logically sound arguments, Yoshihara and Yoshihara criticised it as misleading when it comes to causal conclusions in biological systems (Yoshihara & Yoshihara, 2018). They used gene knockout

experiments as an example in which a gene is considered to be a factor necessary and sufficient to produce a developmental phenomenon. As they argue, the problem with this line of reasoning is that the gene alone (e.g., the *eyeless gene*) is incapable to produce such an effect (e.g., retinal development) in isolation, therefore, the conclusion of the gene is sufficient for the retinal development and consequently "***if** gene **then** develop*" is misleading in a way that a small proportion of the process was considered to be all that is needed to explain the phenomenon.

Moreover, by considering the spatial scale or "*levels*" (Craver, 2015; Sejnowski, Churchland, & Movshon, 2014), we can see that for an effect to emerge at a higher level of organisation, e.g., a specific behaviour, there are many underlying factors from lower levels of organisation that can be experimentally manipulated and counted as a *necessary and sufficient cause*. For example, a specific set of genes on the molecular level, a specific neurotransmitter on the cellular level, and a specific neuronal organisation on the systems level might, in fact, pass the NS condition independently. This abundance of "*sufficient*" causes is, by itself, in contradiction with the definition of "*sufficiency*". Interestingly, the philosopher John Leslie Mackie already pointed to this issue and proposed a more complicated set of conditions to address it (Mackie, 1993). In his work, Mackie assumes a short-circuit as causal for a burnt house since short-circuit ignited a flame while the presence of flammable material and absence of a fire alarm are other parts of the causal chain. However, he argues that these factors are **I**nsufficient but **N**ecessary factors (since none of them by themselves can burn a house down while they were necessary for the effect to take place) in a bigger set of **U**nnecessary but **S**ufficient causes (since besides them other factors can potentially burn a house down). This **INUS** conditioning might be a potential alternative to the NS approach. To put it into perspective, declaring a manipulated circuit being sufficient and necessary, therefore, causal with respect the behavioural change might be misleading since:

1. it is not sufficient *per se* (one circuit alone is not able to completely generate a behaviour without, e.g., muscles or vascular system) and
2. it is not necessary by nature due to the high redundancy in the brain.

However, the circuit can be called a cause that is an insufficient but necessary factor in an unnecessary but sufficient set of factors (Figure 1). This line of reasoning acknowledges the

latent influence of other potential factors and consequently, other causal explanations of the same end product by not considering only one factor of interest as sufficient or necessary. Yoshihara and Yoshihara also pointed towards this issue and argued that if a particular circuit or gene was already found to be necessary and sufficient for the observed behavioural effect, it does not mean that no other causes contribute to the causal explanation (Yoshihara & Yoshihara, 2018). Also, if a target region did not pass the conditions, it should not imply that they do not play any causal role in the occurrence of the event. Interestingly, by moving towards a more pluralistic set of conditions to declare a variable causal, we will be able to surpass the limitations of the deterministic approach towards causality (see Potentials of connecting biological and artificial systems) and explore the interactions among multiple variables, attributing a "*causal contribution score*" to each (Keinan, Sandbank, Hilgetag, Meilijson, & Ruppin, 2006).

2.2 Levels of investigation: type and token causes

One other logical issue when investigating causes underlying behaviour, is the "*type-token distinction*" of causes. A *token* cause can be observed in an event, e.g., person 'A' threw a stone and the stone broke the window. In this scenario, person 'A' then can be pinpointed as the cause of the broken window, however, in domain of type causes neither person 'A' nor the stone are necessary for a window to be broken since there might have been another event in which person 'B' throws a hammer (J. F. Woodward, 2010). Therefore, concluding that either person 'A' or the 'stone' or even a combination of both is "*causally crucial*" for the effect might skew or even hinder the exploratory aspect of science because we generalised a token (event-related) cause to a type (general) cause and therefore ignored, or at best, underrated other potential causes. Translating this to neuroscience, a manipulated circuit might reveal that it has *a causal role* in the behaviour of interest. This experimental finding is comparable to the person 'A' being *a cause* of the broken window in an event. However, there are redundancies and compensations from other brain regions that might have taken over the computation if needed, much like another person could have had broken the window using a potato. In experimental setups, this leads us to control the role of other regions as much as possible, however, even though a complete control over the entire brain might be possible in future, yet that brain itself can be considered as a token cause. The reason is that this particular configuration is one of many

implementations that is evolved to produce under-investigated behaviour, e.g., signalling the presence of a predator. It is quite possible that the same behaviour can be produced by a plant with a completely different biological mechanism that makes the underlying causes being realised differently, e.g., through chemical signalling instead of vocalising. Therefore, behaviour as a system's product is not necessarily bound to one particular implementation and might be fruitful to consider a higher level of analysis such as **the computational level** (what problem the system tries to solve?) or **algorithmic level** (how can it be done?) instead of the **implementation level**, which translates to how it is done by this system (Marr & Poggio, 1976).

2.3 Levels of implementation: The methodological obstacles

Focusing on the implementation level, one of the most puzzling questions in neuroscience is "*What can we call causal in the brain with respect to the produced behaviour? Neuronal elements themselves? The information they transmit? Or both and generally, how?*" (Figure 3). Using the broken window thought experiment, there is an unknown number of potential factors that can break the window and the question of "*how the window can be broken*" has numerous potential and true answers as "*how the behaviour is manifested in this particular way*". Therefore, to better understand the causal structure underlying behavioural effects, we might need to adjust our view, focusing more on the properties of behaviours as phenomena and not all the different combinations of events that can produce them since it is argued to be an "algorithmic barrier" ahead of this approach (Ramaswamy, 2019). The described issue was discussed extensively in an influential paper by Krakauer and others (Krakauer et al., 2017).

Jazayeri and Afraz also argued that biological systems are optimised to perform in their natural environments which might not be possible to reconstruct in laboratory settings (Jazayeri & Afraz, 2017a). In other words, the stimulation parameters used in an optogenetic experiment might be unnatural for the targeted neuronal circuit. By taking the circuit to such obscure space of being constantly activated for seconds, one might observe unnatural responses so that their ecological validity is not guaranteed. This might render the mechanistic conclusion irrelevant and obsolete since it will not give us reliable information about the causal structure of the system as is in its "*umwelt*" and may be hard to generalise.

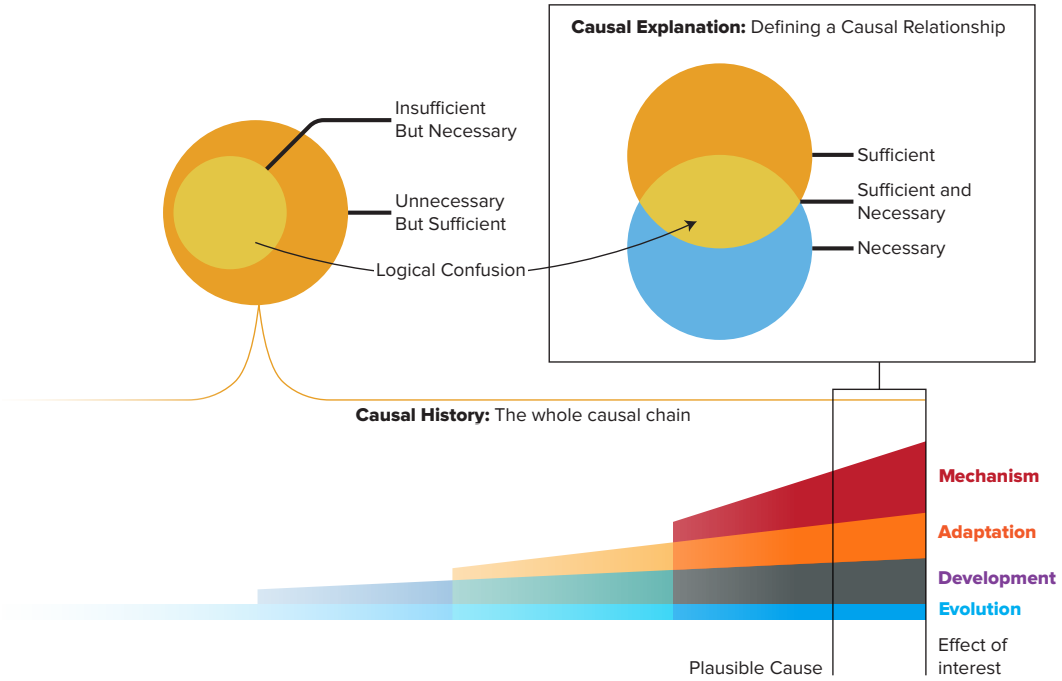


Figure 1: ***Logical confusion raised by NS conditions:*** *In principle, every behaviour has a long, unknown causal history behind. However, a plausible cause will be chosen for a causal explanation. NS (top right) only captures one or a few of these causal links and by claiming a cause as "sufficient" or "necessary" we at least underestimated other potential causes in the causal history of the event. Alternatively, the INUS (top left) concept might give a more accurate description of causal relations by 1.acknowledging that the constructed causal explanation does not contain all the necessary elements and 2.is not sufficient for a comprehensive explanation.*

2.4 Levels of description: Causal history and causal explanation

Overall, the difference between the "*causal history*" of an event and the "*causal explanation*" of it is a crucial point to bear in mind (Lipton, 2010). The causal history of behaviour is not limited to the underlying biological mechanism as we can follow the causal chain to the genetic configuration of the organism and even further to the evolutionary background of it (Tinbergen, 1963). Behaviour has a multiscale causal structure behind - and this structure cannot be simplified to any of the levels alone. Causal explanation of the behaviour, however, is a part of this history that is being used to propose a coherent story that suffices her explanation.

In other words, causal explanations of behaviour are not the full story and must not be confused to serve such purpose (Figure 1). In the following section we review a few of the multiple tools that are used to manipulate neuronal circuits in search for causal explanations of behaviour. We also briefly compare a few of the main methods that are being used for the observational studies. Lastly, we discuss the potential benefits of *in silico* models in search for causes of behaviour.

3 Interventional methods for causal inference

Experimental approaches towards mapping brain regions to behaviours can be traced back to classic case studies on patients in which natural lesions in specific brain circuits caused drastic changes in the behavioural domain (Damasio, Grabowski, Frank, Galaburda, & Damasio, 1994). Nowadays, advancements in technology enabled us to induce artificial and reversible lesions as well as acutely activate specific brain structures. Two major types of artificial lesions were *cooling* and the usage of *neurotoxins* before the presence of even more advanced methods such as optogenetics. Briefly, the cooling approach reduces the activity of the targeted neuronal population by reducing the temperature to at least below 10°C in that area (Coomber et al., 2011). The other approach works by injection of various types of toxins such as Muscimol, to inactivate targeted brain region (Levy, 2001). Furthermore, pharmacological drugs are also used to mainly manipulate brain networks by changing the interaction of certain neurotransmitters (Cools, Nakamura, & Daw, 2011; Tsou, 2012). However, pharmacological manipulations are spatially not as precise as cooling nor neurotoxin-induced lesions (Figure 2) since they can influence receptors across the brain and even body (Ogawa & Watabe-Uchida, 2017; PollakDorocic et al., 2014). In addition to the aforementioned methods, it is now possible to activate or inhibit different cortical areas by using Transcranial Magnetic Stimulation (TMS) that can be implemented in a form of a self-standing therapeutic device, playing a potent role in neurorehabilitation purposes (Polanía, Nitsche, & Ruff, 2018; Parkin et al., 2015). However, there are still debates over the efficacy and precision of non-invasive interventional methods (George et al., 2009). In this section, we aim to introduce and compare a few established and spatially precise techniques developed and widely used to manipulate neuronal circuits.

3.1 Electrical Stimulation

Arguably, electrical stimulation is among the oldest methods of manipulating neuronal elements. The conceptual foundations root in the field of "*medical electricity*" that is older than neuroscience itself and can be traced back to the classic experiments done by Luigi Galvani during the 18^{th} century. Galvani induced electrical currents to the sciatic nerve of a dead frog's leg and observed its twitches (Galvani, 1791). However, counting human brains as targets can be attributed to seminal investigations of Wilder Penfield and his colleague Herbert Jasper during 50^{th}s. In their experiment, they systematically induced weak electrical currents to various cortical regions of epileptic patients, identifying the source of epilepsy (Penfield & Jasper, 1954). Their results led to intricate sensory motor maps known as homunculus and the crucial role of the temporal lobe in recalling memory pieces. This functional specialisation of the brain regions later inspired Jose Delgado to locate candidate brain structures for his famous bull experiment in 1964. During his show, he was able to stop a raging bull from attacking him simply via a radio controller, connected to an electrode inside the bull's Caudate Nucleus and Thalamus (Delgado, 1969). Eventually, these early experiments led to further applications of electrical stimulation to psychiatric and neurological difficulties. For example, Deep Brain Stimulation (DBS) is used to treat a range of disorders highly resistant to pharmacological medication, e.g., Parkinson's Disease, Major Depression, Obsessive Compulsive Disorder, and Alcoholism (for an extensive review see (Lozano & Lipsman, 2013)). Furthermore, electrical stimulation is practised in neural prosthetic devices (Urdaneta, Koivuniemi, & Otto, 2017) since more invasive solutions such as optical implants are not feasible yet. Hence, the need for having precise and natural stimulation parameters has given rise to great challenges and outstanding results.

Technically, one free parameter is the stimulation frequency, i.e., the number of electrical pulses per second, which can be adjusted during the procedure. Another crucial parameter is the intensity of the applied electrical stimulation. A strong stimulation has the ability to directly induce spikes while microstimulations are subthreshold currents, driving neurons closer to their thresholds (Rattay, 1999). Therefore, increasing the intensity can drastically extend the spatiotemporal scale of neuronal response (Figure 2) since higher intensities start a cascade of activity through distant structures and induce long-term synaptic changes in the network,

causing sustained activation or inhibition (Histed, Ni, & Maunsell, 2013). However, Histed and colleagues (2009) found that the spatial scale can be reduced to only a few hundred microns away from the electrode using microstimulations. They also showed that most of the neurons were activated via the electrode itself and could only activate a small fraction of other connected neurons (Histed, Bonin, & Reid, 2009). This so-called "*indirect effect*" is responsible for the activation of distant neuronal circuits even located few centimetres away. To illustrate this effect, Tolias and colleagues (2005) used functional Magnetic Resonance Imaging (fMRI) to measure Blood Oxygenation Level Dependent (BOLD), while stimulating V1 using a high-intensity current of 1 μA for 4 seconds. The results indicated that non-local neuronal populations from extrastriate cortex were also affected by the stimulation (Tolias et al., 2005).

As described, electrical stimulation is one of the established tools in systems-, behavioural- and cognitive neuroscience, however, it cannot target specific cell types as genetic tools do. They rather affect all the neurons of the targeted region, leading to artificial patterns of activity for the brain to process. In addition, few studies showed that axons are more sensitive to the electrical stimulation comparing to cell bodies (Nowak & Bullier, 1998; McIntyre & Grill, 2000). This sensitivity can drastically distort spatial specificity if the electrode is not precisely implanted since the responsible circuit for the observed behaviour can be located at a different region, while having its axons passing near the stimulation site (Histed et al., 2013). Moreover, silicon electrodes are susceptible to Foreign Body Response (FBR) in which glial cells cover the electrode and impair its performance in chronic implantations (Turner et al., 1999).

3.2 Chemo and Optogenetics

A considerable wealth of the neuronal dynamics happens on milliseconds scale (Deco, Cruzat, & Kringelbach, 2019). Thus, in order to create a manipulation paradigm that mimics natural neural events in the network, there is a need for a temporal precision, which optogenetics provides. Optogenetics allows us to either activate or deactivate the targeted circuit using a combination of optics and genetics (Deisseroth et al., 2006). The first steps towards the fully functional optogenetics framework were undertaken in the 70^{th}s when bacteriorhodopsin has been found to be light-sensitive (Oesterhelt & Stoeckenius, 1971). However, it took roughly 40 years for

this idea to become practical, leading the first successful implementation of optogenetics in neuroscience being carried out in 2005 (Boyden, Zhang, Bamberg, Nagel, & Deisseroth, 2005). Despite its promising results, this method is not easy to employ yet, demanding many steps of preparation. Briefly, the initial step of the procedure is selecting the proper *opsin* following the selection of the proper viral vector and dose of injection. In the next step, the neuronal circuit can be targeted by choosing an appropriate set of parameters such as optic parameters, e.g., wavelength and the duration of the laser (Yizhar, Fenno, Davidson, Mogri, & Deisseroth, 2011).

To elaborate, there are two major types of opsins. The light sensing system of type I opsins is based on a single-component since light sensing and ion conductance are carried out by one protein (Yizhar et al., 2011). Therefore, type I opsins are able to directly activate ion channels or pumps due to their ionotropic receptors that enable fast and reliable responses to the light. On the contrary, type II opsins are based on metabotropic receptors. Here, the light indirectly opens ion channels as the signal transduction is carried out by a G-protein-coupled receptor (GPCR). This receptor type elicits a signal cascade comprising various components, thus the light response is relatively slow, reaching response times in the range of seconds (Zhang et al., 2018). Overall, different types of opsins have different properties that allow the cell to be inhibited or excited in a fast or a slow manner. For example, while lightened with a beam of blue light, Channelrhodopsin - a type I opsin - responds in roughly 15 ms leading to depolarisation (Kleinlogel et al., 2011), whereas inhibitory chloride pump opsins need yellow light in order to be activated within 4 ms (Gradinaru et al., 2010).

After selecting a proper opsin, we need to use a *viral vector* to transport the opsin to our targeted neuronal circuit. A dose of the viral vector with proper dilution level has a direct influence on how far the opsins will diffuse. For example, 1μl of the transporter virus AAV5 can spread through the whole hippocampus in a mouse brain while the same amount of the virus AAV2 will be diffused in a smaller volume (Burger et al., 2004). Thus, it is important to select the proper transport virus to aim the targeted spatial diffusion of opsin. Another factor is the light penetration. As shown by Aravanis et al., 2007, light intensity varies across different tissue types (Aravanis et al., 2007). This issue can restrict light penetration and is important to control for. However, cell-specific targeting is a great feature of optogenetics since we are able

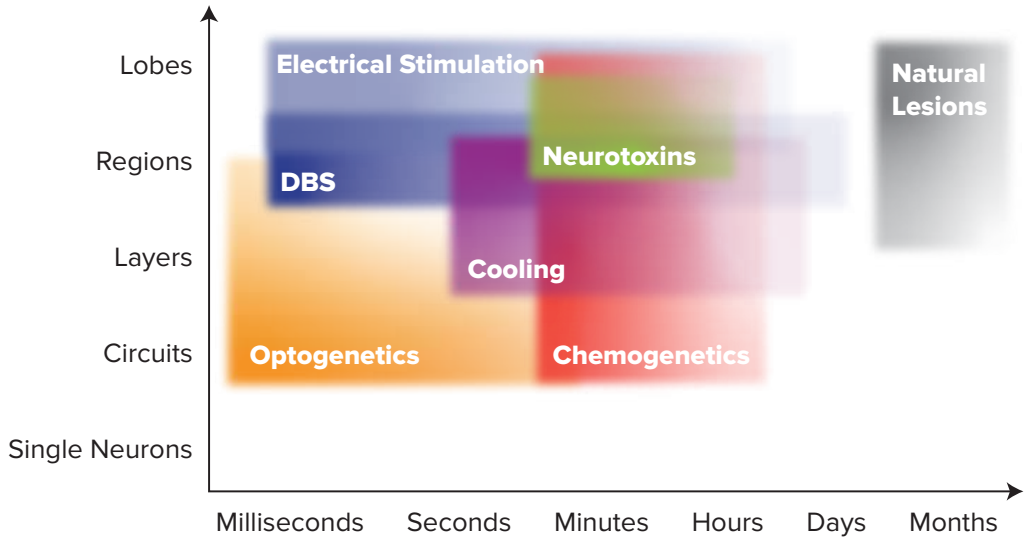


Figure 2: ***The spatiotemporal resolution of the interventional methods:*** *This is a schematic representation of the currently established interventional approaches towards manipulation of the neuronal circuits with respect to spatial and temporal resolution. The rough areas with high colour-density are those that the method is used conventionally or is most effective. DBS: Deep Brain Stimulation. Inspired from (Sejnowski et al., 2014)*

to investigate the role of two different cell types in the same circuit at the same time by shining light with different wavelengths to the tissue.

In summary, optogenetics is a powerful method to decompose neuronal substrate of a particular behaviour because of its remarkable temporal resolution (Figure 2) and the possibility to target multiple cell types in one region with different types of stimulation parameters simultaneously. However, like any other methods, optogenetics also has limitations and can be susceptible to confounds. As an example, tissue heating is a well-known confound in which employing sustained or prolonged usage of light can leads to tissue heating, leading to significant behavioural changes (J. H. Lee et al., 2010; Tsai et al., 2009). Inhibitory opsins are particularly vulnerable to this problem since they need steady light to operate properly. Therefore, the intensity of the light must be considered to avoid such problems (Gradinaru et al., 2010). Another possible limitation is not only restricted to optogenetics but refers to the expression of any other exogenous proteins. Namely, there is an issue of toxicity caused by injecting

a large volume of viral vectors or long-term expression of those proteins (Zimmermann et al., 2008). Although the toxicity can be avoided, the need for advanced technology in order to stimulate parameters itself is critical and still remains to be solved. As mentioned above, Jazayeri and Afraz warned against the issue of ecological validity of our interventions (Jazayeri & Afraz, 2017b). Most of the laser stimulation patterns are not yet close to the normal pattern of activity in the targeted circuit and these so-called "*off-manifold perturbations*" may bring the cell to obscure behaviour, which would not be exhibited in the animal's natural environment. This argument is also valid with respect to many other interventional methods, if not all (Jazayeri & Afraz, 2017b). In other words, while we are using such manipulations, we are the cause of the behaviour and not the circuit since this is not the normal input that it always processes (Gomez-Marin, 2017).

In addition, optogenetics has the issue of light penetration, reducing its efficiency for further cells, while electrical stimulation surpasses the problem via its indirect effects. However, a significant problem of both methods is that the animal, although behaving freely, has optical fibres or electrical wires implanted that need to be connected to the implantation site during the experiment. Recent advances provide us two other options to address this issue; one is the wireless optical and electrical devices (Park et al., 2015; de Haas et al., 2012) and the other is chemogenetic tools also known as **D**esigner **R**eceptors **E**xclusively **A**ctivated by **D**esigner **D**rugs (DREADD) (Armbruster, Li, Pausch, Herlitze, & Roth, 2007). The former still demands invasive surgery to implant external material, which makes it susceptible to FBR and infections in long-term investigations. The latter is a promising tool developed by Roth and colleagues (Roth, 2016) to tackle the aforementioned difficulties. By definition, DREADD is "*a method by which proteins are engineered to interact with previously unrecognized small molecule chemical actuators*" (Roth, 2016). In other words, this technique uses engineered GPCRs to activate or inhibit the targeted cell (Armbruster et al., 2007).

As optogenetics, these receptors can be introduced to the circuit via viral vectors and they can be modulated using synthetic compounds such as clozapine N-oxide (CNO) (Armbruster et al., 2007). Another way of having GPCRs, or opsins for optogenetics, in specific cells is to use transgenic animals specifically manufactured for such purposes. This will make DREADD a

precise tool to target specific cell types in a partially non-invasive manner since the vectors can be injected not directly to the brain but anywhere in the body. After expressing locks or the designed receptors, we are now able to inject the key - ,e.g., CNO -, again, anywhere in the body (Urban et al., 2016) or using eye drops (Keenan, Fernandez, Shumway, Zhao, & Hattar, 2017) to activate the receptor. This allows to perform long-term behavioural tasks with no optical restrains.

In general, optogenetics and chemogenetics should be considered as complementary tools since DREADD acts as a long-term manipulator with lower temporal resolution as compared to optogenetics (Whissell, Tohyama, & Martin, 2016). Interestingly, in terms of spatial resolution (Figure 2), since chemogenetics does not suffer from the problem of light penetration in the larger scale of interest, it can be used as a tool to manipulate large-scale brain networks (Grayson et al., 2016). The non-invasive nature of activating GPCRs also make DREADD a promising tool for translational studies and probably clinical usage (H.-M. Lee, Giguere, & Roth, 2014). Nevertheless, DREADD is also the subject to toxicity caused by inappropriate viral injection and off-manifold perturbations of neuronal activity since there is no control over the temporal parameters.

Box1: Methods for causal inference, applied to observational studies

According to the classic view on the causal inference by Holland (Holland, 1986), *no causation without manipulation*. Indeed, whenever the experimental conditions allow, interventional studies are preferred over observational (a.k.a. quasi-experimental (Marinescu, Lawlor, & Kording, 2018)) studies as they provide with a *direct* evidence for causal links between brain and behaviour. However, in human studies, experimental manipulation is often too invasive and cannot be implemented. Therefore, besides some exceptions (such as TMS), causal inference is performed by means of computational techniques in which brain states cannot be manipulated in the experiment. Here, we list the main computational approaches commonly practised in human studies.

Regression analysis:

Multivariate regression analysis is a study of a relationship between a set of dependent variables (in this case, representing components of behaviour) and one or more independent variables, also known as *predictors* or *explanatory variables* (in this case, brain states), using either a linear or a nonlinear model. In epidemiology and clinical psychiatry, regression models are widely used both for prediction and for causal discovery.

In multivariate *linear* regression, the set of dependent variables is modelled as a *linear* combination of predictor variables. In this model, regression coefficients fitted to the model are the parameters which are associated with causal links between independent and dependent variables. The solution of the inference procedure is a set of regression coefficients which allows for fitting the dependent variables as a linear combination of predictors, with the minimum inaccuracy (expressed in the model as a noise term). This solution is most often obtained with the use of the Ordinary Least Squares approximation (OLS, (Edwards, 1976; Hayashi, 2000)).

In linear regression, we assume that the dependent variable is continuous. Another type of regression often used in epidemiological studies, is logistic regression. Logistic regression is used when the dependent variable is *categorical* (Cox, 1958), especially when it is binary (Prentice & Pyke, 1979). In logistic regression, the relation between the independent and the dependent variables is modelled with the use of a nonlinear, logistic (or, sigmoidal) function.

There are two things to keep in mind while interpreting the results in regression studies. Firstly, the Ordinary Least Squares approximation for finding regression coefficients does not provide a measure of *confidence* for the resulting coefficient parameters. This issue can be overcome in at least two different ways: through permutation testing and through model comparison. Secondly, p-values per each causal link should be corrected for multiple comparisons.

Bayesian nets:

Bayesian networks belong to the group of graphical models and represent a probabilistic approach to causality (Pearl, 1995; Bielczyk et al., 2019). They describe complex systems with interacting entities displayed in a graphical network structure. Each node in the network represents a random variable while their interactions are defined by connections called edges or links. They are displayed as directed arrows, indicating probabilistic causal relations between those variables. Each link stands for one conditional probability. The observance of one variable is more probable if the connected variable also is observed. The joint probability of two or more events occurring together can be calculated as the product of all individual conditional probabilities in the graph. If a node is independent of all other variables in the system it occurs with the marginal probability. That is the probability of that variable taking a specific value.

Usually, prior knowledge is required to define the structure of the graph. However, there are algorithmic approaches to learn the structure of the network directly from the data (Koller, Friedman, Getoor, & Taskar, 2007). If the structure of the graph is eventually defined, the connection strength between the nodes needs to be evaluated. Two approaches might be suitable to do so: finding the set of parameters which best describes the underlying data, or using the Bayesian approach that defines a prior distribution over all parameters and calculates a posterior distribution using Bayesian conditioning (Koller et al., 2007).

Bayesian networks have the advantage of a possible switch between the factorised representation and the graphical representation. In the visual graph, conditional independence can be spotted without any additional analytical tools (Lauritzen, 1996; Pearl & Dechter, 1996). Their great flexibility also allows graphs to be extended, split into sub-graphs or get combined during the modelling process. Deficiencies in Bayesian networks lay in their incapability of accounting bidirectional relations and their susceptibility to hidden common causes in the graph.

Recent developments in the area of observational methods of causal inference shift towards decomposing one-dimensional causal interactions into sets of independent

requirements for causation (existence, composition, information, integration and exclusion, (Albantakis, Marshall, Hoel, & Tononi, 2019)). The applications of these new approaches to experimental datasets are still pending.

4 Potentials of connecting biological and artificial systems

So far, we aimed to provide an overview of the notion "*causality*" and its subsequent issues in a biological system. We first reviewed the logical and conceptual foundations, followed by a comparison of the prominent experimental and computational tools used to infer causal brain-behaviour relations. We mentioned a number of caveats to causal studies of behaviour that we believe should be considered. However, it is important to note that the issues we raised in this paper are still the subject of discussion and new tools and approaches are being proposed regularly to enrich the understanding of the causal links from the brain to behaviours.

As an example, we asked, "*Can a brain region be claimed to have a causal role or is it the neural code that plays such a role*?" Answering this question requires a solid definition of the concept "*neural code*" that by itself is argued to be inappropriate as a metaphor (Brette, 2017). Also, the conceptual discrepancy between probabilistic causality, i.e., the cause raises the probability of the event to occur, and deterministic causality, i.e., the event will reliably follow the cause, leads to an open discussion on which type of causal model is a better assumption when it comes to characterising the brain and its function (Mannino & Bressler, 2015)?

In terms of fundamental assumptions, one other crucial debate is the question of "*decomposability*" of the brain and behaviour into subsystems (Figure 3). It is not clear if we can precisely differentiate functions arising from the brain, and if yes, how much decomposable to discrete functional modules the brain is (Bechtel & Richardson, 2010). In other words, if specific modules in the brain (e.g., Fusiform face area) are causally involved just in a specific task (e.g., face processing) (Kanwisher, 2010) and if processing face can be counted as a phenomenon with clear boundaries. This issue is particularly important for causal brain-behaviour relations: in such conclusions, we implicitly assume that the brain and its products are decomposable to pieces and there are one-to-one or at least traceable relationships between neuronal elements and behaviours. Several lines of research violate such assumption by demonstrating that the

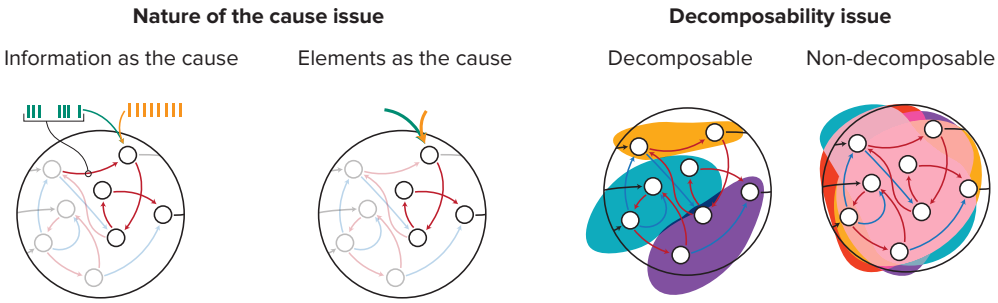


Figure 3: ***Two main issues with respect to neuronal causes of behaviour:*** *Left: Assuming information as the cause leads us towards using different interventional parameters such as stochastic lesioning (i.e, scrambling the activity of the element, which will lesion the "code" (Aharonov et al., 2003)) instead of biological lesioning, i.e., removing the element itself. Right: The assumption of decomposability of the network, although modularity in the brain networks is being under heavy investigation, several lines of research also points towards the so-called polymorphic networks in which each network can have multiple functions (Marder, 1989).*

different configurations of the neuronal network can give rise to the same behaviours (Prinz, Bucher, & Marder, 2004) and different behaviours also can be produced by the same networks (Marder, 1989). This itself leads to probably the most fundamental question that is "*how to define and distinguish a behaviour from its neighbours?*" See (Berman, Bialek, & Shaevitz, 2016; Cande et al., 2018) for an exquisite approach on this issue. For example, is motor output a necessary aspect of behaviour (e.g., acting aggressively versus tendency towards aggression). Altogether, these issues are fundamental yet not addressed properly to date rendering the mechanistic causal claims hard to validate.

As a possible proxy to bypass several of such issues, researchers often use computational models in which all the system's properties are artificially designed or known (Figure 4). Such models are potentially useful since the underlying causal mechanism of the observed behaviour is either transparent or at least accessible for in-depth analyses, at least in theory. Thus it is possible to validate tools and techniques in combination with computational models (Ruppin, 2002; Xu, Jha, & Nachev, 2018). Proposed by Daniel Dennett in an influential commentary (Dennett, 1978) one can "*make up a whole cognitive creature* [...] *and an environmental niche for it to*

cope with" in hope of understanding the "*deep principles of human cognitive psychology*" (Dennett, 1978). This idea of using so-called "*animats*" in their simulated environments as scientific models later developed further into hybrid systems in which a biological neural network was receiving information from and sending commands to a simulated rat in a virtual maze (DeMarse, Wagenaar, Blau, & Potter, 2001). Doing so enables us to investigate the circuit of interest as isolated as needed while still able to contribute to or even perform a behaviourally relevant task.

2.pdf

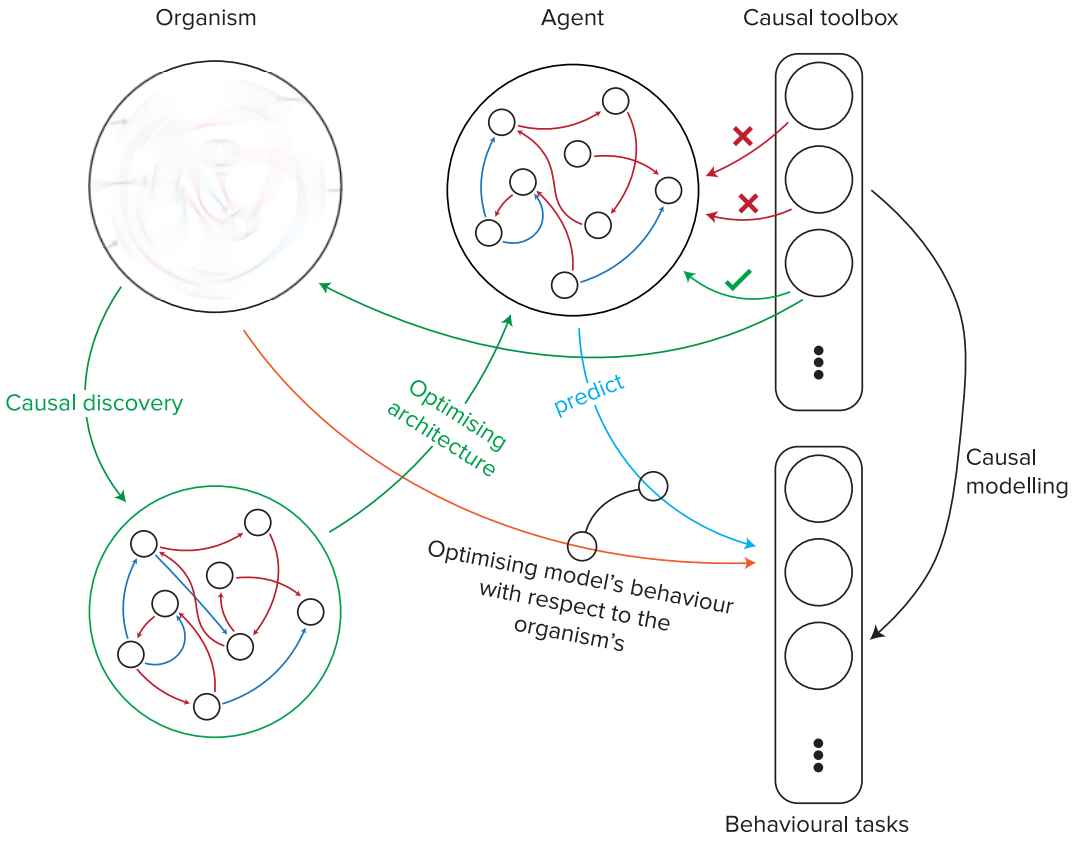


Figure 4: ***Using computational ground-truth models in neuroscience:*** *We can use synthetic systems such as artificial neural networks, simulated agents, and animats as our scientific models to firstly: evaluate our toolboxes and methods and secondly: gain more insight into algorithmic strategies that might be used by the biological systems.*

One interesting example is to explore the functional similarity of Deep Neural Networks (DNN) and the visual system to grasp a more comprehensive understanding of systems that solve the

same task. The potential benefits lay on the assumption that even though the brain and DNNs are different in nature, there might be similarity in the algorithmic level or even functional organisation (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016), therefore, using DNNs as scientific models might provide unprecedented benefits (Cichy & Kaiser, 2019; Barrett, Morcos, & Macke, 2019). It has been assessed by several studies that the hierarchical structure of DNNs, as well as the representations at layers of non-linear computations have correspondence to the real world visual system (Yamins et al., 2014; Kriegeskorte, 2015; Cichy et al., 2016). Here, mechanistic approaches could be useful to explore the causal role of network components, such as the effect of degeneracy by introducing sparseness to the network or exploring why some network configurations trigger a certain behaviour while others do not. However, a natural limitation of DNNs is the lack of interpretability although a few research projects have tackled this issue recently (see, e.g., Montavon, Samek, & Müller, 2018). Another interesting way to use systems with known structure is to investigate methodological assumptions. For example, Jonas et al. (Jonas & Kording, 2017) used a microprocessor as their model to investigate if pervasive analytical tools used in neuroscience can actually reveal how a microprocessor plays games. Crucially, this research showed that the results are structured and contain interpretable patterns, however, these patterns did not meaningfully reveal how the processor solved the given tasks and therefore causally irrelevant (Jonas & Kording, 2017). This points towards an epistemological pitfall that can be addressed using such *ground-truth* models. However, assuming artificial neural networks or simulated agents being identical to their biological counterparts has its own pitfalls and drawbacks that are beyond the scope of this review, nonetheless one of the key points against simulated agents is the problem of generalisability. Even if we gain a complete understanding of the simulated rat in its virtual environment, it might not necessarily generalise to the real rat in its real environment (for example see (Webb, 2009)).

Box2: A hypothetical in silico experiment

Let us take an example of how causal structure of behaviour can be investigated using artificial systems, acknowledging the natural environment and the evolutionary

background of the agent. Following Dennett's proposal (Dennett, 1978), one needs a model and an environment to start with. In case of investigating the causal underpinning of aggression, we can consider starting with an environment that aggression can be incorporated in, e.g., a strategic multiplayer game in which teams of agents have to compete for resources and eventually capture all territories (Stanley, Bryant, & Miikkulainen, 2005). With respect to the model itself, we can use Topology and Weight Evolving Artificial Neural Network (TWEANN) algorithms (Stanley, Clune, Lehman, & Miikkulainen, 2018) such as NeuroEvolution of Augmenting Topologies (NEAT) to evolve a desired number of embodied agents with ANNs as their brains (Stanley & Miikkulainen, 2002). Interestingly and due to the speciation feature, we can evolve agents with different survival strategies. In our case, for example, one evolved strategy might be the aggressive-shooter strategy in which some of the agents prefer to first invade the enemy units then capture their territories. The other strategy can, in contrast, be the sneaking/hiding strategy in which the agents tend to avoid direct costly conflicts, capturing the territories while the enemy is busy somewhere else on the map. Much like the classic case of the Hawk and Dove game, there can independently evolve many strategies that by themselves are interesting to investigate the dynamics of (see (Nitschke, Eiben, & Schut, 2012) for a more concrete case). The other feature is that, unlike more established approaches, there is no need to handcraft the network *a priori* defining the number of neuronal layers and initial patterns of connectivity. In other words, the network's topology can be evolved from unconnected input/output layers to more sophisticated structures with recurrent neurons that can be investigated to mechanistically explain the aggressive behaviour. In addition, this allows us to provide explanations with respect to ultimate causes such as how environmental restraints led to such topology and strategy?

Next, we can aim to answer the proximate causes by leveraging the evolved networks, feeding the topology to a different set of algorithms such as Reinforcement Learning algorithms that are inspired by the way organisms learn from the consequences of their actions. we now can aim to address the issues with respect to learning and development of the specific strategy. Clearly, the opportunity to explore the effects of different factors is

endless. Comparing the neuronal architecture of each strategy with one another, precisely manipulating the elements, having access to the evolutionary history of each agent, being able to examine the effects of all the environmental variables and so on. In fact, it is possible to produce valuable predictions to test in vivo, explain evolutionary forces on the observed circuit model, and eventually model the intervention-related confounds using such simulated ground truth models.

5 Conclusions

To conclude, in this paper we encourage researchers interested in understanding the underlying causes of the behaviours to be aware of the logical, methodological and conceptual foundations of the conclusions they draw from their experiments. As suggested by DiDomenico and Eaton in 1988 (DiDomenico & Eaton, 1988), one study might not be sufficient to draw causal links between variables and it might require validation from not only different tools but also lines of research which employ different definitions of causality since the purpose of the methodology is to serve theory. We then suggested ground-truth models as potential scientific models in the causal circuit explanations and noteworthy, we believe that not all the confusion arises from the logical or methodological issues and it is important to note the issue of "*language confusion*". Mathematical and computational approaches towards causality (see box1) can reduce the impact of such effects. Lastly, we point towards the role of causality in system identification, it might be helpful to keep in mind that even by quantifying a value for the causal contribution of each element in a system, derived from rigorous mathematical approaches (Keinan et al., 2006), one cannot necessarily answer "*what*" those elements are doing so a mechanistic explanation is still yet to reach. This mechanistic explanation itself if reached is just one of many causal explanations with regard to a specific behaviour. Therefore, a complete causal architecture of how the brain produces behaviour should not be considered to be the goal, but a milestone.

Author contributions

KF, LS, DG and NB discussed the topic in a group. KF wrote the first draft of the manuscript and designed the figures. LS, DG and NB took part in planning the manuscript and wrote minor parts of the text.

Funding

KF received a SMARTSTART scholarship, the joint training program in computational neuroscience of the Bernstein Network and the Volkswagen Foundation. LS, DG, and NB received no funding for conducting this research.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank to **Jens-Steffen Scherer, Go Ashida, Markus Lohmeyer** and **Jutta Kretzberg** that helped and encouraged the authors to write down the very first version of the manuscript. We also thank **Alex Gomez-Marin, Asif Ghazanfar, Alexandros Goulas,** and **Claus C. Hilgetag** for fruitful discussions and constructive comments that helped developing the current version. We then thank Brittni Devlin for editing and proofreading the text.

References

Afraz, A., Boyden, E. S., & DiCarlo, J. J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences*, *112*(21), 6730–5. doi: 10.1073/pnas.1423328112

Aharonov, R., Segev, L., Meilijson, I., & Ruppin, E. (2003). Localization of function via lesion analysis. *Neural Computation*, *15*(4), 885–913. doi: 10.1162/08997660360581949

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What caused what? a quantitative account of actual causation using dynamical causal networks. *Entropy*, *21*(5), 459. doi: 10.3390/e21050459

Aravanis, A. M., Wang, L. P., Zhang, F., Meltzer, L. A., Mogri, M. Z., Schneider, M. B., & Deisseroth, K. (2007). An optical neural interface: in vivo control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *Journal of neural engineering*, *4*(3), S143–56. doi: 10.1088/1741-2560/4/3/S02

Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S., & Roth, B. L. (2007). Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. *Proceedings of the National Academy of Sciences*, *104*(12), 5163–5168. doi: 10.1073/pnas.0700293104

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019, apr). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, *55*, 55–64. Retrieved from `http://arxiv.org/abs/1810.13373https://linkinghub.elsevier.com/retrieve/pii/S0959438818301569` doi: 10.1016/j.conb.2019.01.007

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

Berman, G. J., Bialek, W., & Shaevitz, J. W. (2016). Predictability and hierarchy in Drosophila behavior . *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1607601113

Bielczyk, N. Z., Buitelaar, J. K., Glennon, J. C., & Tiesinga, P. H. E. (2015). Circuit to construct mapping: a mathematical tool for assisting the diagnosis and treatment in

major depressive disorder. *Frontiers in Psychiatry*, *6*(29). doi: 10.3389/fpsyt.2015.00029

Bielczyk, N. Z., Uithol, S., van Mourik, T., Anderson, P., Glennon, J. C., & Buitelaar, J. (2019). Disentangling causal webs in the brain using functional magnetic resonance imaging: A review of current approaches. *Network Neuroscience*, *3*(2), 237–73. doi: 10.1162/netn_a_00062

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263–1268. doi: 10.1038/nn1525

Brette, R. (2017, jul). Is coding a relevant metaphor for the brain? *bioRxiv*. Retrieved from `http://biorxiv.org/content/early/2017/07/26/168237.abstract`

Burger, C., Gorbatyuk, O. S., Velardo, M. J., Peden, C. S., Williams, P., Zolotukhin, S., … Muzyczka, N. (2004). Recombinant AAV viral vectors pseudotyped with viral capsids from serotypes 1, 2, and 5 display differential efficiency and cell tropism after delivery to different regions of the central nervous system. *Molecular Therapy*, *10*(2), 302–317. doi: 10.1016/j.ymthe.2004.05.024

Cande, J., Namiki, S., Qiu, J., Korff, W., Card, G. M., Shaevitz, J. W., … Berman, G. J. (2018). Optogenetic dissection of descending behavioral control in Drosophila. *eLife*. doi: 10.7554/eLife.34275

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305-317. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S1364661319300348` doi: 10.1016/j.tics.2019.01.009

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. doi: 10.1038/srep27755

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: unifying affective, activational, and decision functions. *Neuropsychopharmacology*, *36*(1), 98–113. doi: 10.1038/npp.2010.121

Coomber, B., Edwards, D., Jones, S., Shackleton, T., Goldschmidt, J., Wallace, M., & Palmer, A. (2011). Cortical inactivation by cooling in small animals. *Frontiers in systems neuroscience*, *5*, 53. doi: 10.3389/fnsys.2011.00053

Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society. Series B*, *20*(2), 215–42. doi: 10.1111/j.2517-6161.1958.tb00292.x

Craver, C. F. (2015). Levels. In *Open MIND.* doi: 10.15502/9783958570498

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The return of phineas gage: clues about the brain from the skull of a famous patient. *Science*, *265*(5176), 1159. doi: 10.1126/science.265.5176.1159-c

Deco, G., Cruzat, J., & Kringelbach, M. L. (2019). Brain songs framework used for discovering the relevant timescale of the human brain. *Nature Communications*, *10*(1), 1–13. Retrieved from http://dx.doi.org/10.1038/s41467-018-08186-7 doi: 10.1038/s41467-018-08186-7

de Haas, R., Struikmans, R., van der Plasse, G., van Kerkhof, L., Brakkee, J. H., Kas, M. J. H., & Westenberg, H. G. M. (2012, jul). Wireless implantable micro-stimulation device for high frequency bilateral deep brain stimulation in freely moving mice. *Journal of Neuroscience Methods*, *209*(1), 113–119. Retrieved from https://www.sciencedirect.com/science/article/pii/S0165027012002051 doi: 10.1016/j.jneumeth.2012.05.028

Deisseroth, K., Feng, G., Majewska, A. K., Miesenbock, G., Ting, A., & Schnitzer, M. J. (2006). Next-Generation Optical Technologies for Illuminating Genetically Targeted Brain Circuits. *Journal of Neuroscience*, *26*(41), 10380–10386. Retrieved from http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3863-06.2006 doi: 10.1523/JNEUROSCI.3863-06.2006

Delgado, J. M. R. (1969). *Physical control of the mind: Toward a psychocivilized society* (Vol. 41). World Bank Publications.

DeMarse, T. B., Wagenaar, D. A., Blau, A. W., & Potter, S. M. (2001). The neurally controlled animat: Biological brains acting with simulated bodies. *Autonomous Robots*. doi: 10.1023/A:1012407611130

Dennett, D. C. (1978). Why not the whole iguana? *Behavioral and Brain Sciences*, *1*(1), 103–104.

DiDomenico, R., & Eaton, R. C. (1988). Seven Principles for Command and the Neural Causation of Behavior. *Brain, Behavior and Evolution*, *31*(3), 125–140. Retrieved from https://www.karger.com/Article/FullText/116580 doi: 10.1159/000116580

Edwards, A. L. (1976). An introduction to linear regression and correlation. In (pp. 20–32). W. H. Freeman.

Galvani, L. (1791). D viribus electricitatis in motu musculari: Commentarius. *Bologna: Tip. Istituto delle Scienze, 1791; 58 p.: 4 tavv. ft; in 4.; DCC. f. 70*.

Gavériaux-Ruff, C., & Kieffer, B. L. (2007, mar). Conditional gene targeting in the mouse nervous system: Insights into brain function and diseases. *Pharmacology & Therapeutics*, *113*(3), 619–634. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0163725807000022 doi: 10.1016/j.pharmthera.2006.12.003

George, M. S., Padberg, F., Schlaepfer, T. E., O'Reardon, J. P., Fitzgerald, P. B., Nahas, Z. H., & Marcolin, M. A. (2009, jan). Controversy: Repetitive transcranial magnetic stimulation or transcranial direct current stimulation shows efficacy in treating psychiatric diseases (depression, mania, schizophrenia, obsessive-complusive disorder, panic, posttraumatic stress disorder). *Brain Stimulation*, *2*(1), 14–21. Retrieved from https://www.sciencedirect.com/science/article/pii/S1935861X08000375http://linkinghub.elsevier.com/retrieve/pii/S1935861X08000375 doi: 10.1016/j.brs.2008.06.001

Gomez-Marin, A. (2017). Causal circuit explanations of behavior: Are necessity and sufficiency necessary and sufficient? In *Decoding neural circuit structure and function* (pp. 283–306). Springer.

Gradinaru, V., Zhang, F., Ramakrishnan, C., Mattis, J., Prakash, R., Diester, I., ... Deisseroth, K. (2010). Molecular and Cellular Approaches for Diversifying and Extending Optogenetics. *Cell*, *141*(1), 154–165. doi: 10.1016/j.cell.2010.02.037

Grayson, D. S., Bliss-Moreau, E., Machado, C. J., Bennett, J., Shen, K., Grant, K. A., ... Amaral, D. G. (2016). The Rhesus Monkey Connectome Predicts Disrupted Functional Networks Resulting from Pharmacogenetic Inactivation of the Amygdala. *Neuron*, *91*(2), 453–466. doi: 10.1016/j.neuron.2016.06.005

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Histed, M. H., Bonin, V., & Reid, R. C. (2009). Direct Activation of Sparse, Distributed Populations of Cortical Neurons by Electrical Microstimulation. *Neuron*. doi: 10.1016/j.neuron.2009.07.016

Histed, M. H., Ni, A. M., & Maunsell, J. H. R. (2013). Insights into cortical mechanisms of behavior from microstimulation experiments. *Progress in Neurobiology*, *103*, 115–130. Retrieved from http://dx.doi.org/10.1016/j.pneurobio.2012.01.006 doi: 10.1016/j.pneurobio.2012.01.006

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–60.

Jazayeri, M., & Afraz, A. (2017a). Navigating the Neural Space in Search of the Neural Code. *Neuron*, *93*(5), 1003–1014. Retrieved from http://dx.doi.org/10.1016/j.neuron.2017.02.019 doi: 10.1016/j.neuron.2017.02.019

Jazayeri, M., & Afraz, A. (2017b). *Navigating the Neural Space in Search of the Neural Code* (Vol. 93) (No. 5). doi: 10.1016/j.neuron.2017.02.019

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*, *13*(1). doi: 10.1371/journal.pcbi.1005268

Josselyn, S. A., Köhler, S., & Frankland, P. W. (2015). Finding the engram. *Nature Reviews Neuroscience*, *16*(9), 521–534. Retrieved from http://dx.doi.org/10.1038/nrn4000 doi: 10.1038/nrn4000

Kanwisher, N. (2010, jun). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.1005062107 doi: 10.1073/pnas.1005062107

Keenan, W. T., Fernandez, D. C., Shumway, L. J., Zhao, H., & Hattar, S. (2017, nov). Eye-Drops for Activation of DREADDs. *Frontiers in Neural Circuits*, *11*. Retrieved from http://journal.frontiersin.org/article/10.3389/fncir.2017.00093/full doi: 10.3389/fncir.2017.00093

Keinan, A., Sandbank, B., Hilgetag, C. C., Meilijson, I., & Ruppin, E. (2006). Axiomatic scalable neurocontroller analysis via the Shapley value. *Artificial Life*, *12*(3), 333–352. doi: 10.1162/artl.2006.12.3.333

Kleinlogel, S., Feldbauer, K., Dempski, R. E., Fotis, H., Wood, P. G., Bamann, C., & Bamberg, E. (2011). Ultra light-sensitive and fast neuronal activation with the Ca 2+-permeable channelrhodopsin CatCh. *Nature Neuroscience*. doi: 10.1038/nn.2776

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning* (p. 13-56). MIT Press.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017, feb). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. Retrieved from http://dx.doi.org/10.1016/j.neuron.2016.12.041http://linkinghub.elsevier.com/retrieve/pii/S0896627316310406 doi: 10.1016/j.neuron.2016.12.041

Kriegeskorte, N. (2015, November). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Lee, H.-M., Giguere, P. M., & Roth, B. L. (2014). DREADDs: novel tools for drug discovery and development. *Drug discovery today*, *19*(4), 469–73. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24184433{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4004703 doi: 10.1016/j.drudis.2013.10.018

Lee, J. H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D. S., ... Deisseroth, K. (2010). Global and local fMRI signals driven by neurons defined optogenetically by type and wiring. *Nature*, *465*(7299), 788–792. doi: 10.1038/nature09108

Levy, R. (2001). Lidocaine and muscimol microinjections in subthalamic nucleus reverse parkinsonian symptoms. *Brain*. doi: 10.1093/brain/124.10.2105

Lipton, P. (2010). Causation and Explanation. *The Oxford Handbook of Causation*(December 2018), 1–15. doi: 10.1093/oxfordhb/9780199279739.003.0030

Lozano, A. M., & Lipsman, N. (2013). Probing and regulating dysfunctional circuits using deep brain stimulation. *Neuron*, *77*(3), 406–424.

Mackie, J. L. (1993). Causes and Conditions (1965). In *Causation.*

Mannino, M., & Bressler, S. L. (2015). Foundational perspectives on causality in large-scale brain networks. *Physics of Life Reviews*, *15*, 107–123. Retrieved from http://dx.doi.org/10.1016/j.plrev.2015.11.005 doi: 10.1016/j.plrev.2015.09.002

Marder, E. (1989). Modulation of neural networks. *Neural Mechanisms of Behavior*, 55–60.

Marinescu, I. E., Lawlor, P. N., & Kording, K. P. (2018). Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour*, *2*, 891–8. doi: 10.1038/s41562-018-0466-5

Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry.

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*(3489), 1501–1506. doi: 10.1126/science.134.3489.1501

McIntyre, C. C., & Grill, W. M. (2000). Selective microstimulation of central nervous system neurons. *Annals of biomedical engineering*, *28*(3), 219–233.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. doi: https://doi.org/10.1016/j.dsp.2017.10.011

Nitschke, G. S., Eiben, A. E., & Schut, M. C. (2012). Evolving team behaviors with specialization. *Genetic Programming and Evolvable Machines*, *13*(4), 493–536. doi: 10.1007/s10710-012-9166-5

Nowak, L., & Bullier, J. (1998). Axons, but not cell bodies, are activated by electrical stimulation in cortical gray matter ii. evidence from selective inactivation of cell bodies and axon initial segments. *Experimental brain research*, *118*(4), 489–500.

Oesterhelt, D., & Stoeckenius, W. (1971). Rhodopsin-like protein from the purple membrane of Halobacterium halobium. *Nature New Biology*, *233*(39), 149–152. doi: 10.1038/newbio233149a0

Ogawa, S. K., & Watabe-Uchida, M. (2017, may). Organization of dopamine and serotonin system: Anatomical and functional mapping of monosynaptic inputs using rabies virus. *Pharmacology Biochemistry and Behavior*. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0091305717300060` doi: 10.1016/J.PBB.2017.05.001

Park, S. I., Brenner, D. S., Shin, G., Morgan, C. D., Copits, B. A., Chung, H. U., … Rogers, J. A. (2015). Soft, stretchable, fully implantable miniaturized optoelectronic systems for wireless optogenetics. *Nature Biotechnology*, *33*(12), 1280–1286. doi: 10.1038/nbt.3415

Parkin, B., Ekhtiari, H., & Walsh, V. (2015, sep). Non-invasive Human Brain Stimulation in Cognitive Neuroscience: A Primer. *Neuron*, *87*(5), 932–945. Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627315006741 doi: 10.1016/J.NEURON.2015.07.032

Pearl, J. (1995). From bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence* (pp. 157–182). Springer. doi: 10.1007/978-1-4899-1424-8_9

Pearl, J., & Dechter, R. (1996). Identifying independencies in causal graphs with feedback. In *Proceedings of the twelfth international conference on uncertainty in artificial intelligence* (pp. 420–426). Morgan Kaufmann Publishers Inc.

Penfield, W., & Jasper, H. (1954). Epilepsy and the functional anatomy of the human brain.

Polanía, R., Nitsche, M. A., & Ruff, C. C. (2018). Studying and modifying brain function with non-invasive brain stimulation. *Nature Neuroscience*, *21*(2), 174–187. Retrieved from http://dx.doi.org/10.1038/s41593-017-0054-4 doi: 10.1038/s41593-017-0054-4

PollakDorocic, I., Fürth, D., Xuan, Y., Johansson, Y., Pozzi, L., Silberberg, G., ... Meletis, K. (2014, aug). A Whole-Brain Atlas of Inputs to Serotonergic Neurons of the Dorsal and Median Raphe Nuclei. *Neuron*, *83*(3), 663–678. Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627314005832?via{%}3Dihub{#}fig1 doi: 10.1016/J.NEURON.2014.07.002

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403–11. doi: 10.1093/biomet/66.3.403

Prinz, A. A., Bucher, D., & Marder, E. (2004). Similar network activity from disparate circuit parameters. *Nature Neuroscience*, *7*(12), 1345–1352. doi: 10.1038/nn1352

Ramaswamy, V. (2019). An Algorithmic Barrier to Neural Circuit Understanding. *bioRxiv*, 639724. Retrieved from https://www.biorxiv.org/content/10.1101/639724v1 doi: 10.1101/639724

Rattay, F. (1999). The basic mechanism for the electrical stimulation of the nervous system. *Neuroscience*, *89*(2), 335–346.

Roth, B. L. (2016). *DREADDs for Neuroscientists* (Vol. 89) (No. 4). doi: 10.1016/j.neuron.2016.01.040

Ruppin, E. (2002). Evolutionary autonomous agents: A neuroscience perspective. *Nature Reviews Neuroscience*, *3*(2), 132–141. doi: 10.1038/nrn729

Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. *Nature neuroscience*, *17*(11), 1440–1. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25349909{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4224030 doi: 10.1038/nn.3839

Stanley, K. O., Bryant, B. D., & Miikkulainen, R. (2005). Evolving neural network agents in the nero video game. In *Proceedings of the ieee 2005 symposium on computational intelligence and games (cig'05).* Piscataway, NJ: IEEE. Retrieved from http://nn.cs.utexas.edu/?stanley:cig05

Stanley, K. O., Clune, J., Lehman, J., & Miikkulainen, R. (2018). Designing neural networks through neuroevolution. *Nature Machine Intelligence*, *1*(1), 24–35. Retrieved from http://dx.doi.org/10.1038/s42256-018-0006-z doi: 10.1038/s42256-018-0006-z

Stanley, K. O., & Miikkulainen, R. (2002, jun). Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, *10*(2), 99–127. Retrieved from http://www.solver.com/neural-networks-predictionhttp://www.mitpressjournals.org/doi/10.1162/106365602320169811 doi: 10.1162/106365602320169811

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift f??r Tierpsychologie*. doi: 10.1111/j.1439-0310.1963.tb01161.x

Tolias, A. S., Sultan, F., Augath, M., Oeltermann, A., Tehovnik, E. J., Schiller, P. H., & Logothetis, N. K. (2005). Mapping cortical activity elicited with electrical microstimulation using fmri in the macaque. *Neuron*, *48*(6), 901–911.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., De Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, *324*(5930), 1080–1084.

Tsou, J. Y. (2012). Intervention, causal reasoning, and the neurobiology of mental disorders:

Pharmacological drugs as experimental instruments. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(2), 542–551.

Turner, J., Shain, W., Szarowski, D., Andersen, M., Martins, S., Isaacson, M., & Craighead, H. (1999). Cerebral astrocyte response to micromachined silicon implants. *Experimental neurology*, *156*(1), 33–49.

Urban, D. J., Zhu, H., Marcinkiewcz, C. A., Michaelides, M., Oshibuchi, H., Rhea, D., ... Roth, B. L. (2016). Elucidation of the Behavioral Program and Neuronal Network Encoded by Dorsal Raphe Serotonergic Neurons. *Neuropsychopharmacology*, *41*(5), 1404–1415. doi: 10.1038/npp.2015.293

Urdaneta, M. E., Koivuniemi, A. S., & Otto, K. J. (2017). Central nervous system microstimulation: Towards selective micro-neuromodulation. *Current Opinion in Biomedical Engineering*, *4*, 65–77.

Webb, B. (2009). Animals versus animats: Or why not model the real iguana? *Adaptive Behavior*, *17*(4), 269–286.

Whissell, P. D., Tohyama, S., & Martin, L. J. (2016). *The use of DREADDs to deconstruct behavior* (Vol. 7) (No. MAY). doi: 10.3389/fgene.2016.00070

Woodward, J. (2009). *Causation and Manipulability* (E. N. Zalta, U. Nodelman, C. Allen, & R. Lanier Anderson, Eds.). Stanford Encyclopedia of Philosophy. Retrieved from `http://plato.stanford.edu/archives/win2016/entries/causation-mani/`

Woodward, J. F. (2010). *Agency and Interventionist Theories* (Vol. 1; H. Beebee, C. Hitchcock, & P. Menzies, Eds.) (No. February 2017). Oxford University Press. Retrieved from `http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199279739.001.0001/oxfordhb-9780199279739-e-0012` doi: 10.1093/oxfordhb/9780199279739.003.0012

Xu, T., Jha, A., & Nachev, P. (2018). The dimensionalities of lesion-deficit mapping. *Neuropsychologia*, *115*(August 2017), 134–141. doi: 10.1016/j.neuropsychologia.2017.09.007

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual

cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. doi: https://doi.org/10.1073/pnas.1403112111

Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M., & Deisseroth, K. (2011). *Optogenetics in Neural Systems* (Vol. 71) (No. 1). doi: 10.1016/j.neuron.2011.06.004

Yoshihara, M., & Yoshihara, M. (2018). 'Necessary and sufficient' in biology is not necessarily necessary - confusions and erroneous conclusions resulting from misapplied logic in the field of biology, especially neuroscience. *Journal of neurogenetics*, *32*(2), 1–12. Retrieved from `http://www.tandfonline.com/action/journalInformation?journalCode=ineg20{%}0Ahttp://www.ncbi.nlm.nih.gov/pubmed/29757057` doi: 10.1080/01677063.2018.1468443

Zhang, F., Vierock, J., Yizhar, O., Fenno, L. E., Tsunoda, S., Kianianmomeni, A., ... Deisseroth, K. (2018, jan). The Microbial Opsin Family of Optogenetic Tools. *Cell*, *147*(7), 1446–1457. Retrieved from `http://dx.doi.org/10.1016/j.cell.2011.12.004` doi: 10.1016/j.cell.2011.12.004

Zimmermann, D., Zhou, A., Kiesel, M., Feldbauer, K., Terpitz, U., Haase, W., ... Sukhorukov, V. L. (2008). Effects on capacitance by overexpression of membrane proteins. *Biochemical and Biophysical Research Communications*, *369*(4), 1022–1026. doi: 10.1016/j.bbrc.2008.02.153