Gradualism, Bifurcation, and Fading Qualia

Miguel Ángel Sebastián and Manolo Martínez*

May 20, 2023

Abstract

When reasoning about dependence relations, philosophers often rely on gradualist assumptions, according to which abrupt changes in a phenomenon of interest can only result from abrupt changes in the low-level phenomena on which it depends. These assumptions, while strictly correct if the dependence relation in question can be expressed by continuous dynamical equations, should be handled with care: very often the descriptively relevant property of a dynamical system connecting high- and low-level phenomena is not its instantaneous behavior, but its stable fixed points (those in the vicinity of which it spends most of the time, after comparatively short transitory periods), and stable fixed points can change abruptly as a result of infinitesimal changes of the low-level phenomenon. We illustrate this potential gradualist trap by showing that Chalmers's *fading qualia argument* falls in it.

*Authors listed in random order.

1 Introduction

The following *gradualist assumption* is very common. Suppose there is a high-level variable of interest, H, which depends on a low-level variable L, in the sense that there is a lawlike relation between the value taken by H and the value taken by L. The gradualist assumption is *the expectation that abrupt changes in H require abrupt changes in L.* This assumption in turn relies on the notion that the laws that govern such dependence relations are continuous and that, as a consequence, we should not expect discontinuities in the relation between L and H.

While gradualist assumptions are strictly valid, reasoning based on them can lead to grossly inadequate conclusions: very often what actually matters to the behavior of a system is not how instantaneous changes in the value of L are translated into instantaneous changes in the value of H, but how *stable* values of H —the values in the vicinity of which H spends most of the time, after comparatively short transitory periods— change as we change L. And, even in systems governed by continuous functions, stable values of a high-level variable *can* change discontinuously and abruptly in response to gradual changes in a low-level variable: for example, if the relation between H and L is described by nonlinear dynamical equations exhibiting *bifurcations*, it is entirely possible to have sudden, abrupt, irreversible changes in H's stabilities in response to infinitesimal changes in L.[1]

In this paper we bring bifurcations to the attention of philosophers, as

[1] The argumentative flaw that we target in this piece relies on overlooking the possibility of bifurcations in nonlinear systems, and not merely on overlooking the kind of sensitive dependence of the magnitude of some variables on others derived from nonlinearity—cf. Chalmers 1996, pp. 237–239.

an important corner case to keep in mind when reasoning about real-world dependence relations. In particular, we focus on Chalmers' *fading qualia argument* (Chalmers 1996), to this day an extremely popular and widely discussed argument on the metaphysics of consciousness, which, we will show, illegitimately relies on a gradualist assumption. The fading-qualia argument aims at establishing that functional duplicates have qualitatively identical experiences (in the actual world). To achieve this, the argument invites us to imagine a scenario where neurons in a certain cognitive system are gradually replaced with artificial units that perform the same function, and then appeals to our intuitions as to what would happen in that situation with the resulting phenomenal states of the system. Even if the conclusion of the argument is true, the argument is flawed: we should not assume that the relationship between neuronal goings-on and phenomenal consciousness can be adequately described by dynamical equations lacking bifurcations; and if it is governed by equations exhibiting (subcritical) bifurcations, then, *pace* Chalmers, *suddenly disappearing qualia* are perfectly possible—a cautionary tale for gradualists.

In §2 we summarize Chalmers' fading qualia argument, stressing its reliance on a gradualist assumption. In §3 we introduce the relevant theory of nonlinear dynamical systems, and make the connection between bifurcations and the possible shortcomings of gradualist reasoning. We evaluate the argument in the light of this theory in §4. Section §5 offers some concluding remarks.

2 Fading Qualia

Chalmers' *principle of organizational invariance* [OI henceforth] is the claim that "given any system that has conscious experiences, then any system that has the same fine-grained functional organization will have qualitatively identical experiences" (Chalmers 1996, p. 249). That is to say, there is a sufficiently fine-grained specification of the functional organization (according to Chalmers, fine-grained enough to fix behavioral dispositions is fine-grained enough for OI) such that all systems meeting that specification, regardless of their particular realization, will enjoy qualitatively identical experiences.[2]

Chalmers' *fading-qualia* thought experiment (1995; 1996; 2010) is one of the most prominent argument in favor of OI.[3] Chalmers invites us to consider the brain of some human, let's call them Geppetto, currently enjoying a perceptual experience as of a red patch. Assume that we have a sufficiently fine-grained functional specification of this brain's activity (say, its full connectome plus a measure of activation for each unit—or whatever else is needed.) Consider also a robot, GPTto, whose sensory processing happens in a silicon-based computer meeting the exact same specification. We also assume that GPTto is not enjoying an experience as of red. Finally, a sorites series is launched, in which at each step Geppetto's brain is rewired so that one of its neurons is replaced by its silicon analog in GPTto.

Chalmers considers two possible predictions for those who reject OI to make in this scenario: either the qualitative feel of the experience enjoyed

[2]For a recent discussion of functionalism about qualia, see (Van Gulick 2017). We would like to thank an anonymous reviewer for prompting us to be more precise here.

[3]Chalmers' argument relies on keeping the actual laws of nature fixed. For the sake of brevity we omit "... in the actual world" and related qualifications in what follows.

by Geppetto fades as we progress through the sorites series (as we substitute each of its neurons by a silicon unit), until no experience is left by the time we have no neurons left, or *qualia suddenly dissappear*; that is, there is some point in the series at which "the replacement of a single neuron ... could be responsible for the vanishing of an entire field of conscious experience." (1996, p. 238). Chalmers quickly dismisses this second option, for two reasons. First, he claims that if suddenly disappearing qualia were possible, "we could switch back and forth between a neuron and its silicon replacement, with a field of experience blinking in and out of existence on demand" which seems "entirely bizarre" (p. 238). Let us call this possibility "flickering qualia". Second, he claims that suddenly disappearing qualia require "brute discontinuities in the laws of nature unlike those we find anywhere else." In what follows, we show that suddenly disappearing qualia are, in fact, entirely possible if the dynamical system governing the relationship between neurons and phenomenal states exhibits subcritical bifurcations. Furthermore, as it happens, the possibility of suddenly disappearing qualia does not necessitate the possibility of flickering qualia.

Before we are in a position to show this, we need to introduce the relevant dynamical-systems theory.[4] We do so in the next section.

[4]Our introduction draws heavily from Strogatz 2001, which is where we learned about these ideas.

3 Nonlinear Dynamic Systems, Fixed Points and (Subcritical) Bifurcations

Dynamical systems are collections of evolving variables. We describe them using differential equations, which represent how each of these variables change with time as a function of its own and other variables' values.

Consider, to begin with, a simple system consisting of a particle moving in one dimension. We can describe this system completely by giving the position of the particle, x, and its velocity (the instantaneous rate of position over time), $\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}$. Assume that the evolution of the system is fully described by the following differential equation:

$$\dot{x} = x^2 - 1 \tag{1}$$

One common way to make sense of the behavior of dynamical systems such as this is to find their fixed points—those points at which the velocity of the system is zero. This is shown in the so-called *phase portrait* of this system, in fig. 1.

We can see that there are two fixed points of the system, (i.e., points at which the velocity is zero, and therefore the curve intersects with the x-axis) at $x = -1$ and $x = 1$. Examining how the velocity changes as we move slightly away from these points, we see that the leftmost fixed point, marked with a black circle, is *stable* (in the sense that trajectories that are nudged away from it quickly return to it), and the rightmost one, marked with a white circle, *unstable* (because trajectories nudged slightly away from

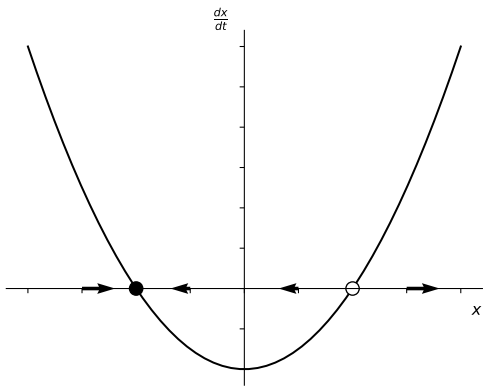

Figure 1: A simple phase portrait. A magnitude of interest, x, is represented on the horizontal axis, and its rate of change, $\frac{dx}{dt}$, on the vertical axis.

it and to its left will leave it and approach the stable fixed point; while trajectories nudged slightly away from it and to its right will leave it and approach infinity.)

It often makes analytical sense to consider a whole family of dynamical systems which differ from one another in the value of certain *control parameters*, to be fixed independently. The presence or absence, and the stability, of fixed points often depends on the value taken by these parameters. For another very simple example, take

$$\dot{x} = x^2 + r \tag{2}$$

Here the behavior of the system depends on the value of r—our previous example was just the particular case in which $r = -1$. Consider what happens as we change r, from some positive value all the way to some negative

value. While $r > 0$, the system has no fixed points, and the speed of the particle at x always increases quadratically as the absolute value of its position increases. However, when $r = 0$ a fixed point appears at the origin and "splits into two, one stable, one unstable" (Strogatz 2001, p. 48), when $r < 0$. The appearance and disappearance of fixed points, and in general the change of behavior as a control parameter changes value, is called *bifurcation*.

An example might help build intuition: consider a bead that can slide freely along a vertical hoop. We know that the bead will end up, after a certain period of time, at the bottom of the hoop (the stable fixed point) unless it is located in a precarious equilibrium at the very top of the hoop (the unstable fixed point). Now we make the hoop spin on its vertical axis at angular velocity r, starting from zero and increasing it little by little. At the beginning, nothing happens: very low angular velocities are unable to overcome friction, and the bead remains at its stable bottom position. But, at some point, the bottom position will cease to be a fixed point, and two new symmetrical fixed points will appear, at an angular distance to the bottom position that depends on the angular velocity of the hoop. That is a bifurcation.

We can now start to see what the potential problem with gradualist assumptions is: if the dynamic behavior of a system is described by an equation like eq. 2, as r changes its value from just above 0 to just below 0, the system changes its behavior from there being no fixed point to there being two, one of them stable. This is qualitatively very different: when there are no fixed points the only possible evolution to the value of x is to increase or decrease forever. When there is a stable fixed point, that's

where the system will land, after a short transitory period. Arbitrarily small variations in r have substantial consequences for the behavior of the system.

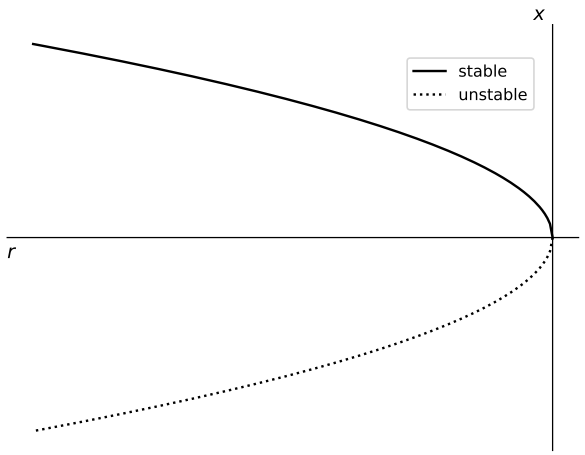

Figure 2: Bifurcation diagram for eq. 2.

Fig. 2 is the bifurcation diagram of eq. 2: it displays the stabilities of the quantity of interest, x, as a function of the control parameter, r. In particular, the fixed points of the system are represented as dotted (for unstable fixed points) and continuous lines (for stable ones).

The diagram illustrates how new fixed points are created as soon as $r \leq 0$. It also displays how, in the dynamics represented by eq. 2, they do not abruptly change location as we modify r: small changes in r correspond to small changes to the location of the stable points. But also this need not necessarily be the case. In so-called *subcritical pitchfork bifurcations* (Strogatz 2001, ch. 3), the position of fixed points changes abruptly, too. Consider eq. 3, where H is the quantity of interest and L a control parameter

(for reasons that will become obvious in the next section, we change here to the variables L and H that we used in the introduction):

$$\dot{H} = L \cdot H + H^3 - H^5 \tag{3}$$

The resulting dynamical system presents a subcritical pitchfork bifurcation. Fig. 3 shows the bifurcation diagram for eq. 3. The parameter L is represented on the horizontal axis, and the quantity of interest H on the vertical axis. Fixed points are represented as dotted and continuous lines following the same convention as before.

Suppose that, in the beginning, the parameter L is somewhere between the point marked L_s and 0. Also suppose that $H = 0$. In that situation, H will remain 0, as this is a stable fixed point. There are also two high amplitude stable branches above and below, but, assuming that H is only subject to small variations, those are currently unreachable.

Suppose now that L starts increasing gradually. While $L < 0$, the situation is qualitatively as described above: the only low-amplitude stable point for H is $H = 0$, and nothing changes. But, when L hits zero (the intersection of L and H axes,) this changes dramatically: suddenly, $H = 0$ becomes unstable, and the only stable points are the high-amplitude branches above and below. This means that the system will abruptly jump to one of those high-amplitude branches, and then move along it. This is what the upward arrow along the vertical axis, and the rightward arrows along the higher branch, are marking. Again here, very significant changes in the landscape of stabilities for H have resulted from infinitesimal changes in L. Although

eq. 3 is perfectly continuous, the change in the stabilities and locations of the fixed points of the system is not, and there are abrupt *jumps* from one "landscape" to the next.

The surprises do not end here. By making L positive, we have moved from having one stable fixed point at $H = 0$ to having two stable branches of high absolute value for H, above and below. Now, intuition might suggest that by reverting L to a value lower than or equal to zero we would regain our $H = 0$ stable point. But this is not how things work in subcritical bifurcations: as fig. 3 shows, once we are in a high-amplitude branch, *it remains stable well below* $L = 0$. This is what the leftward arrows along the higher branch are marking. This lack of responsiveness to parameter changes is called *hysteresis*: some changes to the stable states of a system are easy to make and hard to unmake. Once L falls below L_s, then, yes, the high-amplitude branches become unstable again, and H "jumps off a cliff" to the only remaining stable point at $H = 0$: we are back where we started.

The combination of jumps and hysteresis is far from a mathematical curiosity, and has important engineering consequences (e.g., Chen, Moiola, and Wang 2000). For the sake of simplicity of exposition we are considering only one-dimensional systems. More complex dynamics appear as we increase the number of dimensions. The analogous of pitchfork subcritical bifurcations in two-dimensional systems are called *subcritical Hopf bifurcations*. These occur, for example, in aeroelastic flutter and other vibrations (Dowell and Ilgamova 1988; Thompson and Stewart 1986), instabilities of fluid flows (Drazin and Reid 1981), and, closer to home, in the dynamics of nerve cells (Rinzel and Ermentrout 1989). Subcritical bifurcation offers a

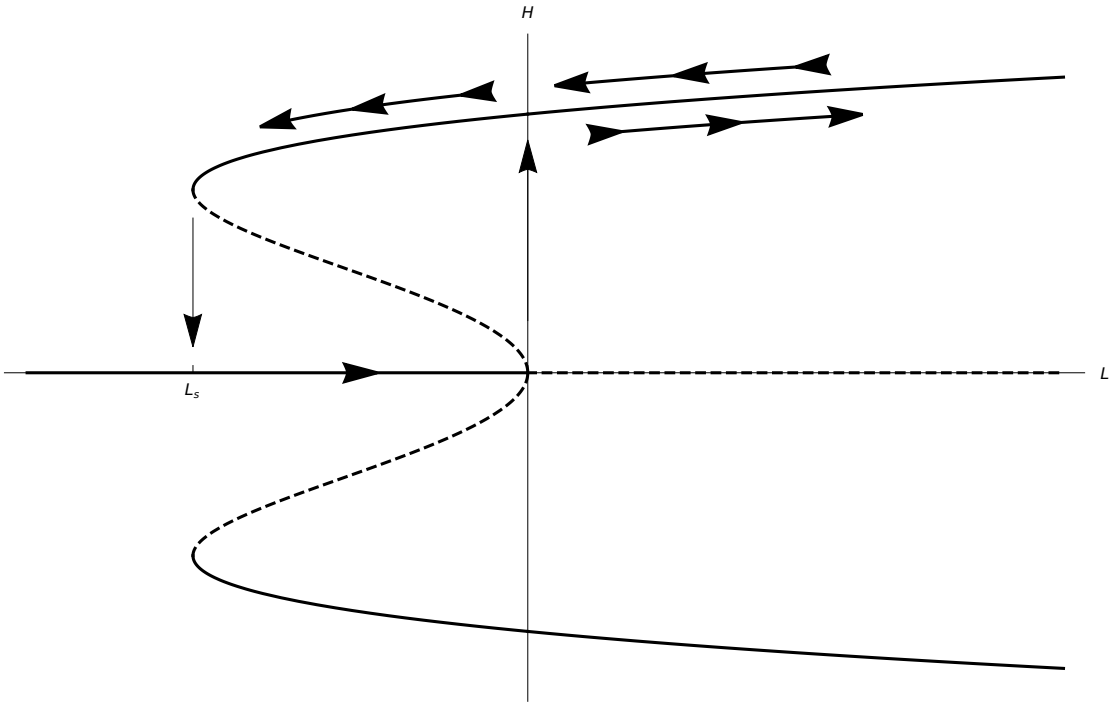

Figure 3: Bifurcation diagram for eq. 3 (a subcritical pitchfork bifurcation). Redrawn from Strogatz 2001, Fig. 3.4.8.

clear, naturalistically unobjectionable model of sharp cut-offs in the dependence relation between variables linked by continuous functions. In the next section we exploit it in the context of Chalmers's fading qualia argument.

4 Flickering and Suddenly Disappearing Qualia

Suppose, with Chalmers, that we are interested in ascertaining the dependence between Geppetto's qualitative experience when looking at a red patch and the neural activity happening in their brain. Suppose in particular that the dynamic dependence between Geppetto's phenomenal experience and

their neural activity is governed by eq. 3, where H is a measure of the vividness of their experience—high absolute value $|H|$ corresponds to a vivid experience; $H = 0$ to no experience at all—,[5] and L is the amount of active neurons in their brain, in units such that $L = 0$ corresponds, say, to a few million neurons. We assume that, throughout the exercise, Geppetto has their eyes open, in front of a red patch. At the outset, we make sure that the system is located in a stable state with high $|H|$—that is to say, Geppetto's experiencing a red patch corresponds to a fixed point somewhere in the high-amplitude branches of fig. 3.[6]

Now we try to launch Chalmers's sorites: we replace one of Geppetto's neurons by a silicon chip, reducing L as a result. The value of H will change in a way governed by eq. 3, ending up, after a transitory time, in a different fixed point. The trajectory taken to reach this point and the transitory time spent in reaching it are irrelevant: what matters is that it *will* quickly end up on a point along the same branch in fig. 3, to the left of where it was before the replacement. We keep substituting neurons and the value of $|H|$ keeps decreasing. So far, so gradualist. However, at some point the number of neurons in Geppetto's brain (L) falls below L_s. At this point, the upper branch is no longer an attractor. The system "falls off a cliff", and quickly evolves to the only remaining attractor, $H = 0$: Geppetto's experience is suddenly extinguished. An infinitesimal change in L (say, a single neuron replacement) was responsible for this change! That is to say,

[5] The sign of H has no designated interpretation in this toy model.

[6] This is, of course, an entirely fanciful model of consciousness. We, like Chalmers, are focusing on *how-possibly* aspects of the dynamical dependence between neural activity and consciousness. Chalmers' discussion of the fading-qualia argument makes no assumption whatsoever about the relevant dynamics.

in this idealized model suddenly disappearing qualia are, *pace* Chalmers, not "entirely bizarre" but the fully foreseeable consequence of minute parameter changes in the vicinity of a subcritical bifurcation.

How about the even more bizarre flickering qualia that supposedly would result as we replace back and forth a silicon chip with a neuron? The answer is that there is no flickering: as we can see in the bifurcation diagram, the origin is still an attractor and H is still zero when the neuron is replaced back. We are in the subcritical part of the hysteresis cycle, and nothing flickers. Chalmers claims that if "we could switch back and forth between a neuron and its silicon replacement, we would see a field of experience blinking in and out of existence on demand" (1995, p. 315). But this is not merely mandated by the fact that such an abrupt change happened in one direction (the disappearing qualia discussed above): hysteresis prevents it from happening in the other direction. Neurons will not start being stably active until L is again greater than zero. At this point there is a bifurcation and we get high $|H|$ again. Only then does Geppetto (suddenly) recover their red quale.

To be clear, we do not dispute the implausibility of qualia depending only on neuronal activity, all the while behavior is sensitive to functional organization implemented in both neuronal and silicon-based activity. Our point is that this implausibility cannot be spelled out as the claim that *suddenly disappearing qualia are naturalistically unacceptable, because there is no nomological mechanism that could account for them.* Sure there is: jumps "off the cliff" in a subcritical bifurcation is a perfectly suitable candidate, with impeccable naturalistic credentials. It also cannot be spelled

out as the claim that *switching one neuron off and on again will result in even more implausible flickering qualia*, because hysteresis cycles may prevent such flickering from happening. We stress again that our rejection of Chalmers' plausibility argument for OI does not depend on the existence of discontinuous laws governing the relation between the relevant magnitudes. For all we know, these laws might not exist—and, indeed, eq. 3 is perfectly continuous. Finally, we are, of course, not assuming that brain dynamics somehow conform to eq. 3. Brain dynamics are massively more complex than that. What we do claim is that Chalmers, and gradualists in general, cannot assume that they are even simpler.

5 Conclusion

If the law governing the relation between two variables is continuous, the gradualist assumption that there cannot be abrupt changes in one without abrupt changes in the other is well motivated. However, caution is required when reasoning on the basis of such a gradualist assumption: in most cases, what is relevant to the description of a phenomenon of interest are the fixed points of the dynamical system that implements this phenomenon. As we have seen, systems whose dynamic behavior is described by perfectly continuous, non-linear equations with subcritical bifurcations present sharp cut-offs in the evolution of the fixed points as a function of a parameter. In those cases, an abrupt change in the location and the nature of fixed points can be due to an infinitesimal change in the parameter at the bifurcation.

In this paper, we have illustrated the perils of gradualist reasoning, using

as our main example Chalmers's fading qualia argument. We have shown that, *pace* Chalmers, it is perfectly coherent and consistent with the laws of nature being perfectly continuous, for a single neuron to be responsible of the extinction of consciousness (through a jump in the vicinity of a subcritical bifurcation), without this committing us to the possibility of flickering qualia as the neuron is restored (because a hysteresis cycle might prevent it).

Many philosophical arguments concerning dependence relations rely on gradualist assumptions. These arguments include classical metaphysical arguments as well as contemporary ones in the philosophy of mind. For example, Lewis 1986's argument for unrestricted compositionality is gradualist, as it relies on the idea that there cannot be cut-off points in whatever relation composition depends upon. Sider 2001's modification of Lewis' argument in favor of perdurantism, and Tye 2021's recent argument for panpsychism on the basis of physical properties having no cut-off points are similarly gradualist. It is not impossible that vulnerabilities in these arguments, and other analogous ones, might be uncovered by paying attention to the dynamical underpinnings of gradualist reasoning, as we have done here for Chalmers' fading qualia. In general, caution in the application of gradualist assumptions is advisable.[7]

[7]We are grateful to Elias Okon, Moises Macías-Bustos, Oliver Marshall, Angélica Pena-Martínez, Alessandro Torza, and three anonymous reviewers for this journal for their comments and discussion. Financial support was provided by the PAPIIT project IN402423; the Spanish Ministry of Science and Innovation, through grants PID2021-127046NA-I00 and CEX2021-001169-M (MCIN/AEI/10.13039/501100011033); and by the Generalitat de Catalunya, through grant 2017-SGR-63.

References

Chalmers, David J (1995). "Absent qualia, fading qualia, dancing qualia". In: *Conscious Experience*. Ed. by Thomas Metzinger. Ferdinand Schoningh.

Chalmers, David J. (Nov. 1996). *The Conscious Mind: In Search of a Fundamental Theory*. 1st ed. Oxford University Press, USA. ISBN: 0195117891.

Chalmers, David (2010). *The Character of Consciousness*. Oxford University Press.

Chen, Guanrong, Jorge L. Moiola, and Hua O. Wang (Mar. 2000). "Bifurcation Control: Theories, Methods, and Applications". In: *International Journal of Bifurcation and Chaos* 10.03, pp. 511–548. ISSN: 0218-1274, 1793-6551. DOI: 10.1142/S0218127400000360.

Dowell, E. H. and M. Ilgamova (1988). *Studies in Nonlinear Aeroelasticity*. Springer, New York.

Drazin, P. G. and W. H. Reid (1981). *Hydrodynamic Stability*. CambridgeUniversityPress, Cambridge, England.

Lewis, David (Jan. 1986). *On the Plurality of Worlds*. First Edition. Blackwell Publishers. ISBN: 0631139931.

Rinzel, J. and G.B. Ermentrout (1989). "Analysis of neural excitability and oscillations". In: *Methods in Neuronal Modeling: From Synapsesto Networks*. Ed. by C. Koch and I. Segev. MIT Press, Cambridge, MA.

Sider, Theodore (2001). *Four-dimensionalism: an ontology of persistence and time*. Oxford university Press.

Strogatz, Steven (2001). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press.

Thompson, J. M. T. and H. B. Stewart (1986). *Nonlinear Dynamics and-Chaos*. Wiley, Chichester, England.

Tye, Michael (2021). *Vagueness and the Evolution of Consciousness: Through the Looking Glass*. Oxford:Oxford University Press.

Van Gulick, Robert (2017). "Functionalism and Qualia". In: *The Blackwell-Companion to Consciousness, Second Edition.* Ed. by Susan Schneider and Max Velmans. Blackwell.