Consciousness Science Underdetermined

A short history of endless debates

Matthias Michel

Sciences, Normes et Démocratie, Sorbonne Université, CNRS.

Author's version. Accepted for publication in **Ergo** (forthcoming).

Abstract: Consciousness scientists have not reached consensus on two of the most central questions in their field: first, on whether consciousness overflows reportability; second, on the physical basis of consciousness. I review the scientific literature of the 19th century to provide evidence that disagreement on these questions has been a feature of the scientific study of consciousness for a long time. Based on this historical review, I hypothesize that a unifying explanation of disagreement on these questions, up to this day, is that scientific theories of consciousness are underdetermined by the evidence, namely, that they can be preserved "come what may" in front of (seemingly) disconfirming evidence. Consciousness scientists may have to find a way of solving the persistent underdetermination of theories of consciousness to make further progress.

Introduction

Scientists studying consciousness have been unable to settle two central debates in the field. The first is about whether subjects are conscious of more than they can report: some researchers believe that consciousness and reportability are equivalent (Dehaene & Changeux, 2011; Naccache, 2018), while others think that consciousness overflows reportability (Block, 1995, 2007; Lamme, 2010). The second debate is about the identification of the physical basis of consciousness (e.g., Boly et al., 2017; Odegaard et. al. 2017). In this debate, theories according to which consciousness pervades the universe co-exist with theories suggesting that some specific parts of the cortex are responsible for consciousness (Tononi & Koch, 2015; Dehaene et al., 2014). My goal is to understand why consciousness scientists do not reach consensus on these questions.

One could argue that there is not much to explain here, for disagreement in the study of consciousness is not very surprising: consciousness is difficult to study, and we shouldn't expect to solve these challenging problems in a snap. After all, one might say, the scientific study of consciousness appeared relatively recently in the history of science. A popular opinion among consciousness scientists, indeed, is that the first attempts to scientifically answer all these questions began in the 1990s, with a series of landmarks articles, such as Crick & Koch's "Towards a neurobiological theory of consciousness" (1990).

This article has two parts: in the first part, I show that the view according to which lack of consensus on the overflow debate and on the physical bases of consciousness has been going on for a short period of time is wrong. To defend that claim, I argue, first, that researchers studying consciousness in the 19th century developed a research program that was very similar to the current science of consciousness, and, second, that early consciousness researchers failed to solve the same problems as those that elicit disagreement among contemporary consciousness scientists.

In the second part of this article, I attempt to explain why consciousness scientists have not reached consensus on the overflow debate and on the physical bases of consciousness after all this time. I will hypothesize that the underdetermination of scientific theories of consciousness by the evidence provides a unifying explanation of the difficulties in solving these problems throughout history.

1. The early scientific study of consciousness

In this section, I show that early consciousness researchers developed a research program similar to that of the contemporary science of consciousness.

1.1. The Leibnizian framework and the cognitive problem of consciousness

To present the early scientific study of consciousness, it is useful to start from a disagreement between the Cartesian and the Leibnizian traditions on the nature of the mind, and more specifically, on the existence of *unconscious* mental activities[1]. William Hamilton, one of the first British thinkers to introduce the Leibnizian framework in Britain, describes the difference between these two philosophical traditions:

> The question I refer to is, Whether the mind exerts energies, and is the subject of modifications, of neither of which it is conscious. This is the most general expression of a problem which has hardly been mentioned, far less mooted, in this country [England]; and when it has attracted a passing notice, the supposition of an unconscious action or passion of the mind has been treated as something either unintelligible, or absurd. In Germany, on the contrary, it has not only been canvassed, but the alternative which the philosophers of this country have lightly considered as ridiculous, has been gravely established as a conclusion which the phenomena not only warrant, but enforce (Hamilton, 1836, p.338).[2]

The idea of unconscious perception seemed "unintelligible, or absurd" to philosophers belonging to the Cartesian and British traditions because they saw consciousness as the defining feature of the mind (*mens*, *l'esprit*)[3]. Descartes writes:

> *Thought*. I use this term to include everything that is within us in such a way that we are immediately aware [*conscii*] of it. Thus all the operations of the will, the intellect, the imagination and the senses are thoughts. I say 'immediately' so as to exclude the

[1] Here and below, I rely on Simmons' interpretation (2001) of the differences between the Cartesian and Leibnizian views of the mind. Danziger has also provided a very similar interpretation of the opposition between the Cartesian and Leibnizian traditions (1980).

[2] John Daniel Morell (1862) makes a very similar remark at the beginning of a chapter entitled "Preconscious Mental Activity": "[Cartesians] regard consciousness as wholly inseparable from mental activity. The same principle passed, through Locke, into the modern English school of metaphysics, and became a fixed idea with nearly all English writers on mental philosophy down to comparatively recent times. On the Continent, and especially in Germany, another and altogether different course was pursued. Leibniz denied the Cartesian dogma *ab initia*, and maintained the doctrine of unconscious perception, or latent thought, as a fact which can be verified throughout all the stages of animal life, and in the actual operations of the human mind." (p.34).

[3] For Descartes, the mind (*mens*, *l'esprit*) is distinct from the body, and thought is its essence, or principal attribute.

> consequences of thoughts; a voluntary movement, for example, originates in a thought. (Descartes, 1644/1985, *Principles of Philosophy*, Part I, §9 / CSM II 113)

Our thoughts can be known directly from a first-person perspective, without the need to infer them from our behaviors, and the fact that we are immediately aware of thoughts makes them mental phenomena. From this perspective, the idea of unconscious thoughts, or unconscious perceptions, is meaningless. For if thoughts and perceptions are unconscious, they are not thoughts and perceptions at all.

On the other hand, following Danziger (1980) and Simmons (2001), we could interpret philosophers from the Leibnizian tradition as believing that *representationality* is the mark of the mental[4]. In an often quoted passage, Leibniz develops the view that there are "obscure perceptions":

> at every moment there is in us an infinity of perceptions, unaccompanied by awareness or reflection; that is, of alterations in the soul itself, of which we are unaware because these impressions are either too minute and too numerous, or else too unvarying, so that they are not sufficiently distinctive on their own. (Leibniz, 1704/1996)

Leibniz explicitly opposed the Cartesians by arguing that they had been "taking for nothing the perceptions of which we are not conscious" (Leibniz, 1714/1965, §4). Making full sense of Leibniz's theory of perception is beyond the scope of this article (for an in-depth treatment, See Kulstad, 1990; McRae, 1978). For our purpose, the most important novelty in Leibniz's theory was the distinction between perception and consciousness:

> it is good to make a distinction between 'perception', which is *the internal state of the monad representing external things*, and 'apperception', which is 'consciousness', or the *reflective cognition of this internal state*, which is *not given to all souls, or at all times to the same soul*. (my emphasis, Leibniz, 1714/1965, §4)

Following Simmons (2001, 2011) we can reconstruct the Leibnizian distinction between perception and consciousness to highlight its differences with the Cartesian view of the mind: first, perception is a representational activity; second, consciousness is not an intrinsic property of perception, but results from a reflexive cognition on perceptions; and third, unconscious perception is possible. For our purpose, the most important aspect of this distinction is that consciousness results from what we would regard today as a cognitive capacity operating on representations.

[4] Cartesians also thought that mental states have representational components, but they believed that they could not be reduced to those (Simmons, 1999). Mental states would be incomplete, for Descartes, if they were only defined representationally, for the experiential character of mental states, which does not seem to represent anything, would be lacking (Simmons, 2001).

A century after Leibniz, many theories of consciousness developed in the Leibnizian tradition testify of a concern for the description of cognitive mechanisms to explain the difference between unconscious and conscious activities of the mind (Romand, 2012). Herbart's theory of consciousness provides a good example (Herbart, 1816/1964). Herbart formalized the concept of "threshold" (*limen*) in psychology and developed what we can interpret as a theory of the cognitive mechanisms by which representations reach the threshold of consciousness or fall below it. According to him, representations could become conscious through a "conflict of the representations" competing in intensity to occupy consciousness. In this competition, representations could gain degrees of consciousness and thus reach the threshold for consciousness, or lose degrees of consciousness and become "repressed" or "consciousless"[5].

Herbart's theory highlights the fact that the Leibnizian framework prompted a wealth of new questions that were not meaningful in a Cartesian framework (Simmons, 2001). For example, trying to explain what makes perception conscious was meaningless in a Cartesian framework, just as trying to develop a theory of the selection of representations for consciousness. Moreover, answering these questions did not require one to take any particular stance on the mind-body problem. For instance, one could evaluate Herbart's theory of consciousness independently of one's stance on the mind-body problem. Similarly, one could remain agnostic on the mind-body problem and yet attempt to solve the problem of knowing which mental activities require consciousness, and which do not. To this extent, the primary benefit of the Leibnizian framework was to separate the mind-body problem from the problem of consciousness, thereby constituting consciousness as an independent target for philosophical and scientific investigation. Consciousness became more than a metaphysical issue, it also became an empirical problem that could largely be answered independently from metaphysical speculations. Insofar as it was distinct from metaphysical speculations about consciousness, the problem that interested researchers of the 19th century could retrospectively be called the "cognitive problem of consciousness". Reconstructed in contemporary terms, the problem was that of knowing which cognitive systems and capacities are essential for consciousness, and by which operations these systems transform unconscious representations into conscious representations.

Except maybe for proponents of the Integrated Information Theory, who do not posit the problem of consciousness in representational terms (Tononi et al., 2015), contemporary consciousness scientists address the cognitive problem of consciousness. For example, the differences between the global workspace theory of consciousness (Dehaene & Changeux,

[5] For more on Herbart's theory of consciousness, see Boudewijnse et al. (1999), and Kim (2015).

2011) and higher-order theories of consciousness (Lau & Rosenthal, 2011) can be reconstructed as differences on which cognitive systems are supposed to be essential for consciousness, and by which operations they contribute to consciousness. On the global workspace theory, representations compete for attentional resources and entrance into a "global workspace", the role of which is to broadcast the winning representations to a variety of cognitive modules. On this view, the "global broadcast" of a representation is the operation by which an unconscious representation is transformed into a conscious representation. On the other hand, higher-order theorists, such as Lau & Rosenthal (2011) believe that global broadcast often happens unconsciously, which could indicate that this is not the operation by which representations become conscious. For them, consciousness depends on a cognitive system charged to differentiate signals from noise: when a sensory representation is targeted by a higher-order representation which tags it as being a reliable signal rather than noise, that sensory representation becomes conscious. These two theories of consciousness certainly do not answer the mind/body problem, but rather, are interested in solving the cognitive problem of consciousness. As such, as I have shown, at least some early consciousness researchers and contemporary consciousness scientists were interested in the same questions. I will now show that they attempted to answer these questions with remarkably similar research programs.

1.2. Bracketing off the metaphysical

In the 19th century, more than a century after Leibniz, scientists could attempt to explain what we would now regard as *cognitive* problems of consciousness without having to find a solution to the mind body problem.

The fact that most researchers of the 19th century recognized the existence of something similar to what would be called the "explanatory gap" a century later (Levine, 1982) contributed to the separation between, on the one hand, metaphysical problems related to consciousness, and, on the other, the cognitive problems that researchers could attempt to answer scientifically. The "explanatory gap" is the intuition that physical accounts of subjective experiences are unable to explain the connection between physical facts, such as brain states, and our subjective experiences. On this view, we could demonstrate that physical states correlate with subjective experiences, but we would be unable to explain *why* those physical states correlate with *these* subjective experiences, and not with different types of experiences, or no experiences at all. A century before Levine coined the term

"explanatory gap", John Tyndall (1872) provided an illuminating description of a very similar problem[6], which he called an "intellectually impassable chasm":

> Granted that a definite thought, and a definite molecular action in the brain, occur simultaneously; we do not possess the intellectual organ, nor apparently any rudiment of the organ, which would enable us to pass, by a process of reasoning, from the one to the other. They appear together, but we do not know why. Were our minds and senses so expanded, strengthened, and illuminated, as to enable us to see and feel the very molecules of the brain; were we capable of following all their motions, all their groupings, all their electric discharges, if such there be; and were we intimately acquainted with the corresponding states of thought and feeling, we should be as far as ever from the solution of the problem, "How are these physical processes connected with the facts of consciousness?" The chasm between the two classes of phenomena would still remain intellectually impassable. Let the consciousness of love, for example, be associated with a right-handed spiral motion of the molecules of the brain, and the consciousness of hate with a left-handed spiral motion. We should then know, when we love, that the motion is in one direction, and, when we hate, that the motion is in the other; but the "WHY?" would remain as unanswerable as before. (Tyndall, 1872, p.95)[7].

Here, Tyndall presents the intuition that, even if we had a perfect description of all brain processes, and even if we were "intimately acquainted with the corresponding states of thought and feeling", we would still be unable to explain how consciousness emerges from physical processes. However, just as the explanatory gap does not prevent the existence of contemporary consciousness science, this "intellectually impassable chasm" did not stop the emergence of the early scientific study of consciousness. Many physiologists readily admitted that the metaphysical problem of consciousness could not be solved, while arguing that they could still study the physiological "conditions of consciousness". For example, Maudsley (1887) writes:

> It is certain that by no exercise of consciousness of which we are capable can we explain what it is in itself (…). The aim of sober inquiry is, therefore, to search and, if possible, find out the conditions of consciousness – the conditions, that is to say, under which it arises, varies, sinks and lapses. (p.489)

In a similar vein, Herzen (1886) writes:

> What is consciousness, such that it manifests itself only when the nervous centers are functioning in a particular way? And why does it manifests itself only in these conditions?

[6] On Tyndall's views on materialism and the mind, See Barton (1987).

[7] This thought experiment echoes Leibniz's "mill analogy", supposed to show, through a similar thought experiment, that machines (or brains) cannot have mental states (Leibniz, 1714).

> The essence of consciousness is as inaccessible as the essence of everything else: do we know what is matter, or what is force? We don't. There are primordial, irreducible, unexplainable facts that we have to accept as they are, and it would already be a significant endeavour if we could specify the conditions under which they appear. That's all we can ask to science. (my translation, Herzen, 1886, p.5).

As illustrated here, the expression "conditions of consciousness" was used as a metaphysically neutral term allowing scientists to make progress on the cognitive problem of consciousness while avoiding to set foot into the morass of metaphysical speculations[8]. In this respect, the research strategy developed in the 19th century is strikingly similar to that of contemporary consciousness science.

Indeed, one of the most important goals of the science of consciousness today is to discover the "neural correlates of consciousness" (Chalmers, 2000; Crick & Koch, 1990). Neisser (2012) is particularly explicit on the role of the term "correlate" in the contemporary study of consciousness: "A notable claim on behalf of the correlate idea is that the neutral language frees us from philosophical disputes over the mind/body relation, allowing the science to move independently" (p.681). Researchers of the 19th century adopted a very similar method to get rid of philosophical disputes over the mind/body relation. This way, they could attempt to discover the physiological differences between cases in which subjects are conscious, and cases in which they are unconscious, as summarized by Herzen:

> If we admit that there is consciousness in some cases, and not others, we are forced to admit that *there is a difference in the conditions of the phenomenon*. We must now try to know when and why (or rather, in what circumstances) the activity of nervous centers is unconscious. (my translation, Herzen, 1886, p.5)

Just as in contemporary consciousness science, the search for the physiological conditions of consciousness played a heuristic role: the hope of early consciousness researchers was not only to find the physiological conditions of consciousness, but also to use physiological data to improve, confirm, or falsify theories of consciousness. Researchers taking part in the early study of consciousness were sometimes philosophers or psychologists, but they all shared a common interest in physiology. Henry Charlton Bastian made this very clear:

> Consciousness being the indispensable basis of all real knowledge, surely no subject can be more interesting than an enquiry – merely tentative though it may be – as to its nature and mode of evolution, including as this does a consideration of the question as to what parts of our organism gave rise by their activity to this universal condition of sentiency.

[8] For more on the strategies used by early neurophysiologist to avoid metaphysical debates about the relation between the mind and the brain, see Chirimuuta (2017).

> But the subject is as difficult and as subtle as it is interesting – and is rendered all the more complex because it has been so often written about by men who, though great philosophers and abstract thinkers, have not always possessed an adequate knowledge of Physiology, wherewith to test the possible truth or falsity of their theories. (...) The more it receives a strictly scientific treatment, starting from a basis of physiological data, the more hope will there be for the stability of the super-imposed theories. (Bastian, 1870, p.502)

Similarly, in contemporary consciousness science, the primary role of the project that aims at finding neural correlates of consciousness is to help researchers devise better cognitive theories of consciousness (Dehaene & Changeux, 2011; Lamme, 2010; Lau & Rosenthal, 2011). For example, physiological data indicates that unconscious processing of sensory information is restricted to sensory areas, whereas conscious processing of sensory information is distributed throughout the cortex (e.g., Dehaene et al., 2001; Fisch et al., 2009; Gaillard et al., 2009). Proponents of the global workspace theory of consciousness use this physiological data to support their cognitive theory of consciousness, namely, that consciousness depends on the global broadcast of information to a variety of cognitive modules throughout the cortex (Dehaene & Changeux, 2011).

So far, I have shown that the early scientific study of consciousness was similar to the current science of consciousness in three respects: first, researchers were primarily interested in what I called the "cognitive problem" of consciousness, rather than the mind/body problem, or the metaphysical problem of consciousness. Second, they used a metaphysically neutral language and attempted to find the physiological conditions of consciousness. Third, the search for the physiological conditions of consciousness also had the same heuristic purpose as the contemporary search for the neural correlates of consciousness. From these similarities between the early study of consciousness and contemporary consciousness science, I suggest that we may learn from the difficulties of the early study of consciousness to inform our contemporary practices.

I will now provide evidence that early consciousness researchers faced the same obstacles as contemporary consciousness scientists, thereby motivating the claim that disagreement on the existence of unnoticed and unreported perception, and on the physical basis of consciousness, is a feature of the scientific study of consciousness throughout its history[9]. In Section 3, I explain why a lack of consensus on these questions has long been a

[9] In the next sections, I will emphasize that early consciousness researchers disagreed on most issues related to consciousness. This should not obscure the fact that, throughout the 19th century, a "consensus practice" progressively emerged (Kitcher, 1995). That is, early consciousness scientists

feature of the scientific study of consciousness by hypothesizing that the science of consciousness has been underdetermined by the evidence throughout its history.

2. Debates in the early study of consciousness

I provide two examples of early scientific debates on consciousness: first, on the existence of unnoticed and unreported perceptions and thoughts, and second, on the identification of the physiological conditions of consciousness. By emphasizing the similarities between these debates and contemporary issues in consciousness science, my goal is show that consciousness researchers have repeatedly failed to reach consensus on the same questions.

2.1. What kind of consciousness?

G. H. Lewes probably did the best job at synthesizing the variety of uses of the term "consciousness" in the second half of the 19th century. He complained that "Whoever reflects on the numerous ambiguities and misapprehensions to which the term Consciousness gives rise in philosophical discussion will regret that the term cannot be banished altogether. But since it cannot be banished, our task must be the attempt to give it precise meanings." (Lewes, 1879, p.143). Lewes begins by noting that one acception of the term is "synonymous with Feeling". He then empresses to distinguish further between feeling and "sentience". Sentience is defined as a physical state which can, but needs not, give rise to conscious perception. For example, physiologists knew that the *feeling* associated with visual perception did not originate from the eye as a *sentient* organ, because, as Flourens remarked (1842, p.24), the eyes of decerebrated animals continued to react to light variations and the optic nerve continued to be excitable while the animal did not perceive anything. Consequently, an organism could continue to be sentient without perception.

After defining consciousness as feeling and distinguishing feeling from sentience, Lewes writes that another sense of "consciousness" is to use it "synonymously with Cognition and with Attention. According to this view, to be affected and not to know that we are affected is to be unconscious of the affection: to attend to the affection is to be conscious of it" (1879, p.145). We find a similar distinction in multiple works by different authors, such as Thomas Laycock, who distinguished between "consciousness as feeling" and

differed on questions that they all equally considered as meaningful with a shared commitment to using empirical data to answer these questions and refine their theories.

"consciousness as knowledge", which he also called "cognitional consciousness" (Laycock, 1860, p.141); or Hamilton, who emphasized a dispute between philosophers who defined the term as "knowledge" and many others who "defined the term as a feeling" (Hamilton, 1836, vol. 1, p.191)[10].

The distinction between "consciousness as feeling" and "cognitional consciousness" mirrors Block's distinction between "phenomenal consciousness" and "access consciousness" (Block, 1995). "Phenomenal consciousness" refers to "phenomenality", or the "what it is likeness" of our experiences (Nagel, 1974), the fact that they *feel* a particular way to us. "Access consciousness" refers to the fact that some representational contents are poised for direct use in reasoning, speech, rational action and subjective reports. Hence, it seems that early consciousness scientists worried about the ambiguity of the term "consciousness" and developed very similar distinctions 150 years before contemporary consciousness science[11]. In the next section, I will show that, in the 19th century, the distinction between "consciousness as feeling" and "consciousness as knowledge" led early consciousness researchers to disagree on the existence of unconscious thoughts and perceptions. I will also provide evidence that contemporary consciousness scientists still struggle with similar problems as a result of the distinction between phenomenal consciousness and access consciousness.

2.2. The early overflow debate: unconscious perception and thoughts

[10] However, other authors, as James Mill, considered that the word "consciousness" was strictly synonymous with "feeling": "To say I feel a sensation is merely to say I feel a feeling (...) And to say I am conscious of a feeling is merely to say that I feel it." (1869, vol. 1, p.224). Similarly, Alexander Bain argues that "the knowledge or attention, although an accompaniment of the state, is not its foundation (...) It is most accordant with the facts, to regard Feeling as a distinct conscious element, whether cognized or not, whether much or little attended to in the way of discrimination, agreement or memory" (Bain, 1884, 94).

[11] The distinction between these two types of consciousness also played a role similar to its contemporary role in early debates over the richness of consciousness (for a review of current views on this topic, see Cohen et al., 2016). The question was to know how many objects one could be conscious of at once. Hamilton wrote: "By Charles Bonnet the mind is allowed to have a distinct notion of six objects at once; by Abraham Tucker the number is limited to four; while Destutt-Tracy again amplifies it to six." (1836, p.254). These approximations are strikingly similar to our current knowledge of working memory limitations (e.g., Cowan, 2000). Laycock argued, however, that Hamilton's observation did not "seem to meet the question" (1860, p.153). He defended that Hamilton and others were conflating two questions concerning the richness of consciousness. According to him, the question was "not how many objects you may think you see, but how many objects you can be conscious of at once" (1860, p.154). On Laycock's view, Hamilton *thought* he was able to see six objects at once, but this observation didn't answer the question of whether he was able to *be conscious*, in the sense of "consciousness as feeling", of more than six objects at once.

Due to the ambiguity of the term "consciousness" the expressions "unconscious perception" or "unconscious thoughts" can either mean that some thoughts or perceptions affect our behavior without us *knowing* that they do; or, that some thoughts or perceptions influence our behavior without us *feeling* that we have those thoughts and perceptions.

Accordingly, one can distinguish between three different views on unconscious thoughts and perception in the 19th century: (1) unconscious thoughts and perceptions do not exist at all, in any sense of the term "conscious", a view that I call the Cartesian view. (2) Unconscious thoughts and perceptions exist *only* in the sense that we sometimes do not *know* that we have some thoughts and perceptions when we have them, although having those thoughts and perceptions still *feels* like something. Following a similar contemporary view (Block, 2007), I call this the "overflow" view. (3) Unconscious thoughts and perceptions exist in both senses of the term "conscious", a view that I call the "Leibnizian view". I now provide a brief review of arguments in favor and against these views.

Unconscious thoughts were supposed to play a role in explaining a variety of phenomena, among which, the association of seemingly unrelated ideas[12]. The following case, provided by Hamilton, was often discussed:

> Suppose, for instance, that A, B, C, are three thoughts, that A and C cannot immediately suggest each other, but that each is associated with B, so that A will naturally suggest B, and B naturally suggest C. Now it may happen, that we are conscious of A, and immediately thereafter of C. How is the anomaly to be explained? It can only be explained on the principle of latent modifications. A suggests C, not immediately, but through B; but as B (...) does not rise into consciousness, we are apt to consider it as non-existent. (...) One idea mediately suggests another into consciousness, the suggestion passing through one or more ideas which do not themselves rise into consciousness. (Hamilton, 1836, p.352-353)

To account for this kind of cases and avoid appealing to unconscious thoughts, those who did not accept the Leibnizian framework, such as Carpenter (1874, Chapter XIII) and John Stuart Mill (1865), readily admitted the existence of "unconscious *cerebration*". According to

[12] In his chapter on "preconscious mental activity", Morell argues in favor of unconscious perceptions and thoughts by using an inference to the best explanation based on several cases that unconscious thoughts and perceptions are supposed to explain, among which: "After puzzling over a difficult problem a long time, and leaving it unsolved, we not unfrequently find, on taking it up again, that the materials have rearranged themselves in our minds, so that the solution is perfectly easy. (...) Secondly. One idea will sometimes suggest another, which had, as far as we know, no previous connection with it. Thirdly. *Habits*, when fully acquired, will come into operation, under proper circumstances, quite unconsciously. (...) Fourthly. Cases of this kind often occur. We write a letter and despatch it. Two or three days after we remember that we have made an error in the statement, or spelt a word incorrectly. At the time, the error was committed unconsciously; by a latent process that error is brought, perhaps, some days after, into the sphere of consciousness." (1862, p.37).

them, *the brain* could influence behavior without us knowing or feeling its influence, but the idea that *the mind* could do so was meaningless. For proponents of this Cartesian view, an idea suggested another through the influence of an unconscious cerebration, but no activity of an unconscious *mind* needed to be involved. J. S. Mill writes:

> I am myself inclined to agree with Sir W. Hamilton, and to admit his unconscious mental modifications, in the only shape in which I can attach any very distinct meaning to them—namely, unconscious modifications of the nerves (...) it may well be believed that the apparently suppressed links in a chain of association, those which Sir W. Hamilton considers as latent, really are so; that they are not even momentarily felt; the chain of causation being continued only physically by one organic state of the nerves succeeding another so rapidly that the state of mental consciousness appropriate to each is not produced. (1865, vol. 2, p.22)

Two main reasons led to the early demise of the unconscious cerebration view, as explained by Harald Höffding:

> Instead of speaking of unconscious thought or unconscious feeling, it would be safer—if we wish to avoid all hypotheses— to speak with Carpenter and John Stuart Mill of unconscious cerebration, were not this expression unsuitable, as suggesting, in the first place, the mistaken notion that there may be consciousness of cerebration, properly so called, and because, in the second place, it might appear to affirm that there is nothing at all in unconscious activity related to what we know in ourselves as conscious states. (1891, p.81)

First, it appeared senseless to talk about unconscious cerebrations, because cerebrations could never be conscious in the first place. As Lewes wrote: “We ought never to apply the negative to phenomena of an order which does not admit its positive. No one, indeed, would think of calling a machine unconscious or a dog inhuman; but we may call a man inhuman, and a sentient act unconscious.” (1879, p.151). Second, and more importantly, unconscious activities seemed to be *of the same kind* as conscious activities, such that talking about unconscious cerebrations instead of unconscious mental activities amounted to “an exclusion of the mind from the highest functions of the mind”, as argued by Henry Holland (cited in Ireland, 1875, p.380)[13].

Instead of appealing to unconscious cerebrations, some researchers opposed the Leibnizian view by positing *conscious* activities of the mind that were unattended,

[13] By the end of the 19th century, the expression “unconscious cerebration” disappeared. The last proponents of this view complained that “the expression “unconscious cerebration” is one rarely seen in the contemporary literature. It is hardly to be found in the indices to treatises on psychology, and even Baldwin's 'Dictionary' fails to assign it a separate caption.” (A. H. Pierce, 1906).

undiscriminated, or rapidly forgotten. To the best of my knowledge, Condillac was the first to use this strategy against the Leibnizian doctrine:

> I distinguish between two kinds of conscious perceptions: those that we remember, and those that we immediately forget. (…) I think that we are always conscious of the impressions that we receive in the soul; but, sometimes, those perceptions are so subtle that we forget them immediately." (my translation, Condillac, 1743, sect. ii, §6)

Condillac's view, on which consciousness overflows the limits of memory, gained considerable influence in the 19th century[14]. The most important proponent of the overflow view was surely Lewes, according to whom "oblivescence is no proof of insentience" (1879, p.148):

> That we forget feelings immediately [after] they have passed is not an argument against their having been felt. We forget myriads of feelings, even energetic feelings, experienced a year ago, a week ago, an hour ago. Some which passed but a minute ago — visceral sensations, sights, sounds, touches — are beyond recall. Who will say that these were organic states but not feelings? To be conscious of performing an act, and to be conscious of having performed it, are two different mental states (1879, p.166).

A large number of philosophers and physiologists, such as François Achille Longet (1842), Daniel Noble (1858, p.96-97), Alfred Vulpian (1866) or William Ireland (1875) agreed. According to proponents of the overflow view, Leibniz's "obscure perceptions" were unnoticed, undiscriminated, but nonetheless *felt* sensations[15]. In that sense, these authors thought that we could be conscious of more than we could consciously discriminate at one moment or could remember having been conscious the moment after. For instance, Henry Calderwood uses a simple example to support the view that we have more sensations than we can consciously discriminate:

> Let the whole ten fingers be moved over the same surface at the same moment, and we fail to distinguish ten distinct sensations. The failure in discrimination does not occur because there are not ten distinct impressions, with ten distinct molecular changes in the brain, and then distinct sensations, but because we have not discriminating power

[14] For example, Hamilton (1836, p.339), Laycock (1860, p.183) and Dunn (1858, p.90) refer to Condillac's view on the existence of unconscious activities of the mind.

[15] There is, however, a crucial difference between this early "overflow" debate, and the current overflow debate (Block, 1995; Block, 2007). The early version of this debate focused on the question of knowing whether unconscious perception exists or not. As such, accepting the overflow view was a way of denying the existence of unconscious perception. On the other hand, in the current overflow debate, proponents of the overflow view generally accept the existence of unconscious perception (e.g., Block, 2016). Instead, they distinguish between unaccessible, phenomenally conscious contents, and unconscious contents (which are, by definition unaccessible too). I thank an anonymous reviewer for drawing my attention to this difference between the early, and contemporary overflow debates.

> enough to deal with so many. In this way it happens that multitudes of impressions are made on the sensory nerves which are never noticed by us. The failure in this case to keep the distinction sharply confirms the view that the discriminating power is quite distinct from that which determines the existence of the sensation. The nerve fibres can do more work than the discriminating power at our command can interpret. (Calderwood, 1879, p.221).

Early consciousness researchers posed the debate over the existence of unnoticed sensations in terms that were very similar to those of the contemporary overflow debate, initiated by Block (1995). For example, Lewes writes:

> At any given moment you are unconscious of feelings in your finger-tips and toes, nay, unconscious of having those parts, a momentary attention suffices to raise a vivid consciousness of fingers and toes. Were these feelings non-existent (...) and only called into existence by an increased innervation of the parts consequent on the act of attention? Or were they existent, but obscured by the predominance of other stimulations? (Lewes, 1879, p. 186).[16]

He answered the question by arguing that "unless some sensation were already there, no effort of attention could evoke it" (p.186). Otherwise, according to Lewes, one would need to suppose that attention could somehow "create" sensations. This argument was unsuccessful because proponents of the Leibnizian view argued that, in these cases, attention could be directed towards *unconscious* sensations. Höffding is particularly clear on this:

> In like manner, when we listen in a state of abstraction to someone speaking to us, we may not until long afterwards become conscious of what he has said. It is only by the express direction of attention that the impressions unconsciously received are here raised above "the threshold." That we are able to remember something is therefore no decisive proof that we consciously apprehended it at the time of its occurrence. By connection with that which has been consciously apprehended, even an unconscious impression may be called to memory. (Höffding, 1881, p.76)

Other authors such as Hamilton (1836), or Bastian (1869, 1870a), straightforwardly rejected the distinction between consciousness as feeling and consciousness as knowledge, by arguing that consciousness always involves cognition. On Bastian's view, we cannot attend to a sensation without *knowing* that we have that sensation or at least knowing that the sensation has a particular quality. If we ignored everything concerning our sensations, we would have no reason to attend to them. Hence, to the extent that unnoticed sensations are

[16] Compare this with Schwitzgebel's question "Do you have constant tactical experience of your feet in your shoes? Or is experience limited to what's in attention?" (2007).

available for attention, they must somehow be *known* to have particular qualities. For this reason, Bastian argues:

> Mr. Bain stops short of the truth when he says "the lowest or more restricted forms of sensation does not contain any element of knowledge." It does not contain knowledge, it is true, in its highest sense, involving affirmation and belief, but as a state of consciousness, it is inseparable from knowledge in its essence, which implies *discrimination of difference or agreement*. We, in common with others, would rather believe that no sensation, not even the simplest, can exist without the element of cognition being at the same time present in consciousness. (Bastian, 1869, p.214)[17]

The problem with the overflow view, according to Bastian, is that

> any sensation, however simple, can only be recognised as such—can only be revealed in consciousness—inasmuch as it represents a certain quality or qualities, by which it can be differentiated from or classed with previous states of feeling. Therefore even the most simple sensation does necessitate the existence of intellectual activity, since discrimination is the most fundamental mode of intellect. (p.214)

According to Bastian, for a sensation to have the quality that it has, one needs to be able to discriminate *that* sensation from other sensations. On his view, the capacity to discriminate between different sensations requires some *knowledge* of those sensations. Hence, every conscious sensation must somehow be known, for, otherwise, the sensation would not have the particular quality that it has. And if sensations must be known to have the particular qualities that they have, consciousness as feeling requires consciousness as knowledge. Consequently, consciousness as feeling cannot "overflow" the limits of consciousness as knowledge.

It seems that Lewes had anticipated this argument, however. He notes that "the term Cognition is ambiguous" (Lewes, 1879, p.183). By the idea that consciousness involves cognition, one could mean either that when one is conscious of something, there is "a recognition by the Ego of its own operations"; or that "consciousness [is] discriminated feeling". Lewes thought that the only sense in which cognition was relevant to consciousness was in the latter sense of the term. He agreed that consciousness involved cognition, in the sense that it required some capacities to discriminate sensations. However, he refused the claim that consciousness involved the *recognition of oneself* as discriminating sensations. Hence, according to Lewes, one can have *conscious* sensations, which result from operations of cognitive capacities, *without knowing* that one is currently exerting these cognitive capacities or has these sensations. By making the distinction between cognition at

[17] For the record, this article appeared in the very first volume of the journal *Nature*, in 1869.

a sub-personal level (the discriminations involved in sensations), and cognition at a personal level (the recognition by the Ego of its operations), Lewes escaped the conclusion that consciousness as feeling should be reduced to consciousness as knowledge.

Considering that all participants in this debate were empirically minded, one natural way of settling the problem could have been to appeal to empirical results demonstrating unconscious perception. Gustav Fechner was probably the first researcher attempting to provide empirical evidence of unconscious perception. He reported a number of cases in which unseen objects could give rise to after-images (Fechner, 1860)[18]. According to him, if unseen objects could have conscious *effects*, it necessarily meant that they were seen unconsciously (Romand, 2012). Fechner's inquiries later influenced the first real experiments attempting to demonstrate unconscious perception, most notably Peirce and Jastrow's experiment (1884). However, even if researchers were aware of these experiments, they probably wouldn't have settled the matter. Indeed, it seems that Peirce and Jastrow interpreted their finding that subjects (i.e., themselves) could make sensory discriminations without knowledge of making these discriminations correctly, as indicating that some *differences in sensations* could fail to elicit *a sensation of difference*[19]. Nonetheless, they note that a failure to have a sensation of difference does not necessarily mean that there was no actual difference in conscious sensations. Peirce and Jastrow insist that their result indicates that there is no least perceptible difference in sensations, as supposed by Fechner, but only perceptible differences in sensations that fail to elicit sensations of differences. As such, researchers could have interpreted Peirce & Jastrow's early experiment in various ways, and proponents of the overflow view wouldn't have been convinced that this experiment could demonstrate unconscious perception.

Given the disagreements just exposed, it is safe to say that, by the end of the 19th century, the debate over the existence of unconscious perception was still unsolved. My aim is not to provide a complete analysis of the discussion over unconscious thoughts and perception in the 19th century, but only to provide an overview of the debate to justify that a

[18] To the best of my knowledge, the only researcher taking this case into account was Höffding: "Fechner relates (*Elements of Psychophysics*, Vol. ii, p. 432), that one morning in bed he was surprised by having a white image of the stove-pipe when he closed his eyes. As he lay with his eyes open and speculated, he had seen before him, without being conscious of it, a black stove-pipe with a white wall as background, and what now made its appearance was the negative after-image of this. The physical excitation had thus been of such a nature that the visual sensation *might have* arisen; but the attention being otherwise engaged, what appeared to consciousness was not the sensation itself, but only the more impressive after-image" (1891, p.384).

[19] Peirce and Jastrow write that "the quantity which we have called the degree of confidence was probably the secondary sensation of a difference between the primary sensations compared." (1884, p.82).

comparison between problems faced by early consciousness researchers and those of contemporary consciousness scientists is meaningful.

Today, most researchers accept the existence of unconscious perception. Peters & Lau (2015) found that cognitive scientists who participated in a survey on unconscious perception "reported believing that subliminal processing exists (94%)". However, they also found, in the same survey, that only 36% of participants believed that the existence of unconscious perception had been unequivocally demonstrated in the empirical literature. Although scientists have developed a wide variety of empirical methods for assessing unconscious perception, debates over the existence of unconscious perception continue (e.g., Peters et al., 2017). Similarly, the overflow debate, revived by Block in the 1990s, remains unsettled. It seems that most researchers recognize that solving this puzzle is first and foremost a methodological and conceptual challenge rather than a matter of acquiring more data. For instance, in a review of the current methods used to address the question of whether phenomenal consciousness overflows cognitive access, Phillips (2018) concludes that "given our present data and methods, not only do we not know whether consciousness requires cognition, we do not know how to find out" (p.7). Hence, it seems that, after 150 years, no consensus has been reached on very similar problems, namely, on the existence of unconscious perception and on whether there can be unnoticed, unremembered or unreportable perception. I conclude that lack of consensus on these problems is a feature of the study of consciousness which, accordingly, calls for an explanation.

Some researchers have argued that physiological data could help solving this debate (Lamme, 2010; Block, 2007). As it turns out, physiologists of the 19th century thought so too, but the problem still couldn't be solved. In the next section, I provide an overview of the debate on the physiological conditions of consciousness.

2.3. The physiological conditions of consciousness

Psycho-physiologists of the 19th century could not snap colored pictures of the brain with functional MRIs. However, they could lesion, decerebrate and decapitate many non-human vertebrates, and see what happens in each case. Unfortunately for the animal kingdom, that's what they did.[20]

[20] Early consciousness scientists, and chief among them, David Ferrier, were targeted by anti-vivisectionist groups for both ethical and religious reasons (Finn & Stark, 2015). Indeed, the anti-vivisectionist movement was as much concerned by animal suffering as by the "cold, proud, atheistic spirit that distinguishes modern investigators" (Clarke, 1888). Ferrier's investigations on localization of brain functions, and the research of the early study of consciousness, were perceived

Pflüger's experiments on decapitated frogs will provide a useful starting point for our discussion, as these experiments signed the beginning of a debate that would occupy consciousness scientists for fifty years (Pflüger, 1853; Klein, 2017). Here is the case of the decapitated frog, as described by Ferrier:

> When a drop of acetic acid is placed on the thigh of a decapitated frog, the foot of the same side is raised, and attempts made with it to rub the part. On the foot being amputated, and the acid applied as before, the animal makes a similar attempt, but failing to reach the point of irritation with the stump, after a few moments of apparent indecision and agitation, raises the other foot, and attempts with it to remove the irritant. (Ferrier (1876), p.20)

Klein (2017) provides two additional historical cases with decapitated frogs:

> a brainless frog will swim if dropped in water (Lewes 1877, 190). If completely submerged, it will swim to the surface. And not only that; if one impedes the emerging, pithed frog by putting an inverted jar in its path, the frog will not easily be trapped. It will actually re-descend until it can swim out of the jar, and then will swim up to the surface (Goltz 1869, 70). This is an astonishing sequence of behaviours for an animal that lacks a brain. (Klein, 2017, p.7).

Now, here is the vexing question that caused so much debate throughout the 19th century:

> What is the nature of the impression which is the immediate antecedent of this responsive activity? Is it a purely physical phenomenon, or has it likewise a subjective side? In other words, are these actions merely reflex or excito-motor, or are they the result of sensation properly so-called? If we define sensation as the consciousness of an impression, it will be seen that the problem to be solved is, whether consciousness is an accompaniment of the activity of these centres (Ferrier, 1876, p.40)

There are three main kinds of responses to this question (summarized in Figure 1): first, one can argue that consciousness "accompanies" the activity of the spinal cord. Lewes (1873, 1879), Herzen (1886), Pflüger (1853), Foster (1890), and Schiff (1858) supported this view. Second, other researchers defended that consciousness accompanies the activity of the midbrain or the thalamus, such that *decorticated* (i.e., without a cortex) animals are conscious, but not decapitated animals. Carpenter (1874), Dunn (1858), Vulpian (1866), Longet (1842) and Noble (1858) defended this view. Third, some claimed that the cerebral cortex was necessary for consciousness, such that neither decorticated or decapitated

as providing support for materialist views of the mind. After the passage of the 1876 Cruelty to Animal Act, anti-vivisectionsts prosecuted Ferrier in 1881. The prosecution failed and Ferrier received important support from the scientific community, emphasizing that Ferrier's work was crucial for developing surgery using his functional maps of the brain.

animals were conscious. Flourens (1842), Bastian (1870), Ferrier (1876), and Maudsley (1867) championed this view[21].

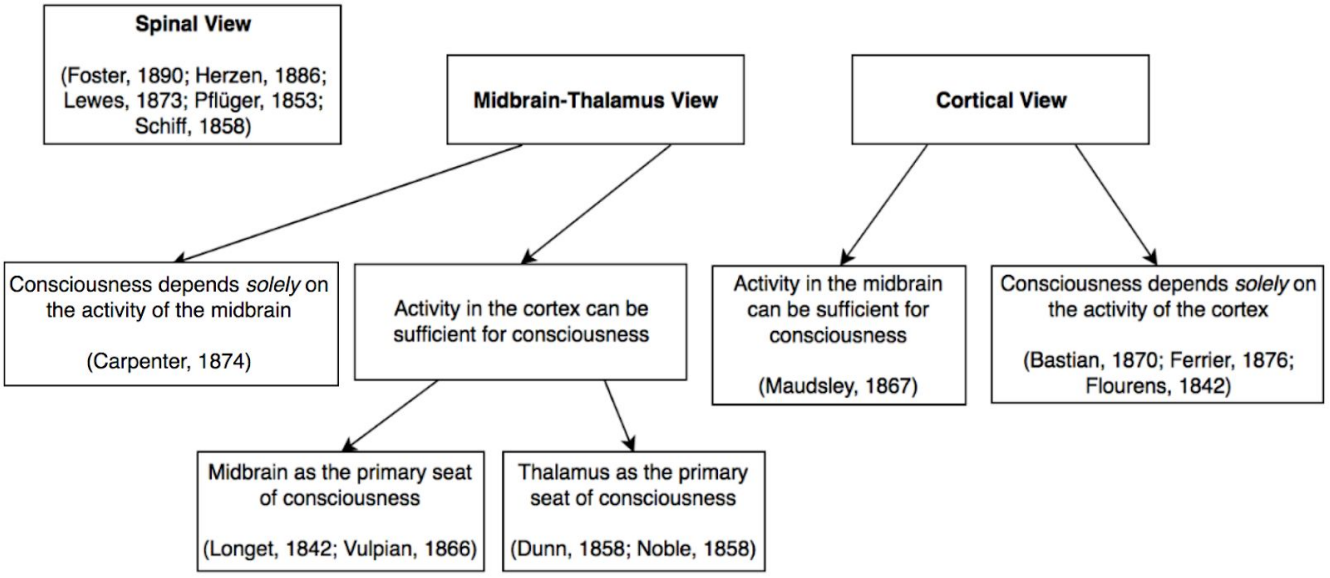

Figure 1. Summary of the variety of views on the physiological conditions of consciousness.

I now review several arguments and counterarguments for each of these views, beginning with arguments in favor of the spinal view. First, it is important to note that the type of sensations elicited in the spinal cord are nothing like the sensations we usually have. Lewes writes:

> In saying that the Spinal Cord is *a* seat of sensation, it is not meant that it is *the* seat, nor that the sensations are *specifically* like the sensations of colour, of sound, of taste, of smell; but they are as like these as each of these is like the other. (1873, p.84)

Despite being different from ordinary sensations, they are sensations nonetheless. Moreover, these sensations are supposed to be extremely transient, as Foster writes:

> we may thus infer that when the brainless frog is stirred by some stimulus to a reflex act, the spinal cord is lit up by a momentary flash of consciousness coming out of darkness and dying away into darkness again (1890, p.912).

[21] There are important differences between all these authors, even when I classified them in the same category (Figure 1). For example, Carpenter thought that the midbrain was the only seat of consciousness, while Vulpian and Maudsley believed that activity in the midbrain or the cortex could be sufficient for consciousness. Similarly, Dunn and Todd disagreed with Noble on whether the activity of the striatum could be sufficient for experiencing emotional sensations. Finally, at some points Maudsley seemed to accept that the activity of the midbrain could be sufficient for consciousness, particularly in "lower animals", while Ferrier was categorical on the fact that consciousness depended only on the activity of the cortex. Nonetheless, despite these differences, I believe that the present taxonomy is broadly representative.

Now, there are three main arguments in favor of the spinal view. First, researchers generally admitted that complex behaviors involving adaptation to unusual situations require consciousness. Behaviors of decapitated animals are well-adapted and complex. Therefore, the spinal cord is sufficient for consciousness (Pflüger, 1853). Second, Lewes (1873) argues that nerve tissues in the spinal cord and the medulla oblongata or midbrain are not fundamentally different. If tissues are not radically different, and if mental properties result from properties of the tissues, then mental properties are not fundamentally different between the spinal cord and the midbrain. Hence, if one acknowledges that the midbrain can generate conscious experiences, one must accept the same conclusion for the spinal cord. Third, Herzen (1886) remarks that decerebrated animals continue to react to nociceptive stimuli just as non-decerebrated animals. These "pain-like behaviors" are the only objective signs on which we can rationally base our attributions of pain in typical cases. Consequently, Herzen argues, based on the observation of the same pain-like behavior, it is irrational to attribute pain experiences to non-decerebrated animals while refusing to assign pain experiences to decerebrated animals. If pain-like behaviors count as evidence for pain experiences in the former case, it should also be the case in the latter. Hence, one must conclude that the spinal cord and the midbrain are sufficient for consciousness, at least in the case of pain experiences[22].

To the first argument, proponents of the midbrain-thalamus and cortex views responded that we are not typically conscious of reflex actions elicited in the spinal cord, and these actions are not usually under rational control. Consequently, it seemed that all that was proved by the complexity of behaviors in decapitated animals was that the spinal cord could be sufficient for complex behaviors, but not sufficient for consciousness. After all, Carpenter and others had already shown that a wide variety of complex and seemingly goal-directed behaviors could happen unconsciously, as in the case of eye movements for example (Carpenter, 1874).

To the second argument, one could respond that many behaviors which probably recruited the cerebral hemispheres were *not* conscious (Bastian, 1870), thus proving that activity of "nerve tissues" was probably not sufficient for consciousness. Moreover, following Flourens (1842), Carpenter had shown that an operation could ablate the cerebellum or a disease destroy it without any loss in "sensorial capacity" (Carpenter, 1864). If ablation of the cerebellum did not modify consciousness, then consciousness had to be elicited by a

[22] Very similar arguments have recently been developed in defense of the existence of pain experiences in fish (See, e.g., Tye (2017), and Michel (2018) for a critical analysis of these arguments).

specific kind of activity of "nerve tissues" and not by *any* kind of nervous activity, as supposed by Lewes.

To the third argument, it could be answered, following Goltz's experiments (1869), that decapitated frogs lacked some of the pain-like behaviors that they would have had if they had pain experiences, as did non-decapitated frogs. Indeed, Goltz discovered that brainless frogs placed in water, the temperature of which is slowly raised, do not manifest any pain. Opponents to the spinal view used this experiment to argue that, if a frog's spinal cord were able to feel pain, the frog should have reacted to the painful increase in temperature. Instead, brainless frogs did not exhibit any pain-like behaviors in this situation.

Despite these arguments, proponents of the spinal view maintained that activity in the spinal cord could be sufficient for consciousness. After all, they could counter the existence of unconscious reflex actions elicited in the spinal cord by arguing that these actions are typically accompanied by faint sensations that are unremembered or unattended. Moreover, against Goltz' experiment, Foster (cited in Lewes, 1873) reports that if a decapitated frog has its leg in water while the temperature is gradually raised, it will withdraw its leg. Lewes concludes that "The depressing influence of heat on the Spinal Cord destroys its reflect powers" (1873, p.84), he continues:

> It proves, to my mind, that although the frog remains motionless in the heater water and shows no sign of pain from the stimulus of heat, this is assuredly not because Sensibility in general is gone, but simply because Sensibility to temperature is gone. (p.84)

It seems that, by the end of the 19th century, although the view that the spinal cord acted only automatically became the consensus view among physiologists[23], no *decisive* argument had been given against the spinal view. Boring retrospectively concluded that "there could be no clear decision between Pflüger, who said that the reflexes of the cord should be conscious, and Lotze, who said they were not" (Boring, 1950, p.666).

On the other side of the theoretical spectrum, opponents to the spinal view also disagreed on the physiological conditions of consciousness. I identify three main arguments in favor of the midbrain-thalamus view. First, animals without their cerebral cortices could present complex behaviors. Second, invertebrate animals do not have a cerebral cortex but can perform what seem to be conscious actions (Carpenter, 1874). Third, the midbrain or the thalamus seem ideally suited to integrate information from all the senses to constitute one unified conscious experience because they receive afferent nerves from the sense organs (Dunn, 1858; Noble, 1858).

[23] See Liddell (1960) on the history of the discovery of reflex actions.

Proponents of the midbrain-thalamus view thought that the cortex might be sufficient for consciousness, especially consciousness of abstract thoughts or complex perception, but conscious *sensations* resulted mainly from the activity of the midbrain or thalamus[24]. Accordingly, the role of the hemispheres was to "intellectually develop sensations" and "transform them into ideas" (my translation, Vulpian, 1866, p.672). In other words, the cerebral cortex was responsible for access consciousness, and consciousness of thoughts, but was not responsible for "consciousness as feeling".

David Ferrier begged to differ[25]. He remarked, first, that the supporters of the midbrain-thalamus view accepted that consciousness did not arise in the spinal cord, although the spinal cord could produce complex and adapted behaviors in decapitated animals. Warning us to avoid relying "on appearances alone", he writes:

> the mere faculty of adaptation is not necessarily a proof of consciousness, for, as we have seen, it exists in some degree in the spinal cord, and if it is not regarded as proof of conscious action on the part of the cord, neither can it be taken as such here; for it may be that the more complex adaptation manifested by the mesencephale is simply the result of more complex and special afferent and efferent relations. (p.43)

Based on this argument, either the midbrain-thalamus view had to collapse into the spinal view, or proponents of the mibrain-thalamus view had to provide evidence that there was a difference in kind, and not in degrees, between activity in the spinal cord and the midbrain. As argued by Carpenter, such evidence could come from the "evidently conscious actions of invertebrate animals" (1876, p.44), which do not possess a cortex. Nevertheless, Ferrier responded that

> the ganglia of the invertebrates are not completely homologous with the mesencephalic ganglia of vertebrates, for if they were so, we should expect that not merely sensation, but also the other psychical faculties, should be manifested by vertebrates deprived of their cerebral hemispheres, even though to a less degree. But it is not a difference in degree only which is observed, but a manifest difference in kind. It is probable, therefore,

[24] At some points Carpenter (1874) seems to defend the view that *all* thoughts originate under the form of unconscious cerebrations, and are thus unconscious, until these thoughts are somehow transferred to the midbrain where they can evoke sensations and become conscious. In that respect, Carpenter's view might have been similar to the contemporary view held by Carruthers (2015), according to which all thoughts are unconscious because consciousness is "sensory-based".

[25] Ferrier received the ideal training for participating in the early science of consciousness: he was Alexander Bain's student, and also studied in Wundt and Helmholtz' laboratories. Later, he worked at the same hospital as Hughlings Jackson, who became a friend and mentor (Pearce, 2003). It is no surprise then, that Ferrier became a fellow of the royal society at the age of 33. His first book, *The Functions of the Brain* (1876), is one of the greatest achievements of early neurophysiology, and contains some of the best examples of the scientific rigor of the early scientific study of consciousness.

> that in the ganglia of the invertebrates there are nerve cells which perform, in however lowly a manner, the functions of the cerebral hemispheres in vertebrates. (1876, p.44)

In other words, Ferrier agreed that invertebrate animals are conscious. However, he argued that one could not conclude anything on consciousness in decorticated animals from this, for the mesencephalic structures in humans should not be considered analogous to the brain of invertebrates[26].

Probably the most convincing of Ferrier's arguments against both the midbrain-thalamus and spinal views came from experiments in animals, but also from clinical studies in humans:

> we have experiments of disease which practically detach the hemispheres from their mesencephalic connections, and leave thought and speech intact, so that we can obtain direct testimony as regards the consciousness of impressions. Such an experiment is performed by a lesion of the crus cerebri or of the posterior part of the peduncular expansion (...), phenomena not unfrequently occurring in clinical experience. When this occurs, the individual has absolutely no consciousness of tactile impressions made on the opposite side of his body, *however much he may strain his attention to receive them*. In the mesencephale alone, therefore, sensory impressions are not correlated with modifications of consciousness; whence we must conclude that sensation is a function of the higher centres. (Ferrier, 1876, p.45)

If the midbrain were *sufficient* for consciousness, one would predict that sectioning the connection between the cortex and the midbrain would not affect consciousness. Ferrier remarked that this is not what happens, both in patients with lesions and when the section is experimentally performed in animals. Rather, it seems that subjects are not conscious of tactile sensations, because impressions are not transmitted from the unconscious midbrain to the cortex where they are transformed into conscious sensations.

Finally, Ferrier was also well known for following Fritsch and Hitzig's experiments (1870) on electrical stimulations of the brain to discover the specific functions carried out by various brain areas, and applied these methods to identify the physiological conditions of consciousness. Among other observations, he did not find any specific modifications of behavior by stimulating the optic thalamus, but he remarked that stimulation of the angular

[26] Bastian had a similar response (1870), although a bit more complex, based on Spencer's evolutionary reasoning applied to psychology (Spencer, 1868). Bastian thought that, as they develop throughout evolution, activities of cerebral mechanisms become automatic and unconscious. Accordingly, the activity of the brain of invertebrates could be conscious, while, in the course of human evolution, activity in the human midbrain progressively became automatic and unconscious, leaving conscious activity to the newly acquired cerebral cortex.

gyrus could cause “confusion in vision” (Ferrier, 1875, p.425)[27]. Ferrier’s innovative experimental techniques yielded new arguments in favor of the cortex view: vision could be modified by electrical stimulation of the cortex, independently from the midbrain or the thalamus.

Despite Ferrier’s genius, early consciousness researchers could not uncontroversially identify the physiological conditions of consciousness. His work on electrical stimulation was criticized by Lewes, arguing that there was no proof that the electrical current stimulated the cortex instead of merely passing *through* the cortex (Lewes, 1876, p.74). Moreover, the argument according to which the spinal cord is unconscious because subjects lose tactile sensations when a lesion separates the spinal cord from the brain had already been answered by Schiff (1858), who argued that, in these cases, consciousness was split, such that there could be one consciousness in the spinal cord and another in the brain.

Since the early study of consciousness, contemporary consciousness scientists have not reached consensus on these questions, and they still hold, with variations, the different views represented in the 19th century. Some philosophers and consciousness scientists hold panpsychist or almost-panpsychist views, according to which a wide variety of entities that we do not usually consider conscious, such as, for instance, sets of inactive logic gates, are conscious (Aaronson 2014a,b; Chalmers, 1996; Tononi & Koch, 2015). The midbrain view is still held by Merker, who argues that neural activity in subcortical structures of the brain is sufficient for consciousness (Merker, 2007). The local recurrency theory of consciousness is also quite similar to the midbrain-thalamus view, in that its proponents highlight the difference between conscious sensations and conscious access to these sensations, leading them to hypothesize the existence of consciousness in “sensory modules” (Block, 2007; Lamme, 2015). Finally, others believe that consciousness is dependent upon the activity of the prefrontal cortex, either to globally broadcast contents throughout the brain (Dehaene & Changeux, 2011), or to create higher-order representations (Lau & Rosenthal, 2011) (on this issue, see also Odegaard et al., 2018; Michel & Morales, forthcoming). The fact that some theories on which a wide variety of entities are conscious can co-exist with theories on which only some cognitive mechanisms in the prefrontal cortex

[27] Ferrier’s identification of the parietal cortex as the seat of vision was wrong. The localization of the area of vision led to a controversy between Ferrier and Munk (Fishman, 1995; Glickstein, 1985). The debate was quite tense, to the point that James wrote: “'the quarrel is very acrimonious; indeed the subject of localization of functions in the brain seems to have a peculiar effect on the temper of those who cultivate it experimentally (...) Munk's absolute tone about his observations and his theoretical arrogance have led to his ruin as an authority” (James, 1890, p.46). Despite being impolite, Munk was right: the occipital cortex is the seat of vision.

generate consciousness is surely the sign of an absence of consensus on the neural correlates of consciousness.

Following the brief historical review of the search for the physiological conditions of consciousness, and considering contemporary disagreements on the neural correlates of consciousness, I conclude that lack of consensus on the physical basis of consciousness is a long-standing feature of the scientific study of consciousness. Accordingly, as in the case of the overflow debate, this long-standing disagreement in the study of consciousness calls for an explanation.

3. Consciousness science underdetermined

It seems that both the overflow debate and the debate over the physical basis of consciousness have been going on for a while, at least since the 19th century, with little success in settling both of these debates. As such, I have argued that lack of consensus on these problems has been a feature of the study of consciousness throughout its history.

I now hypothesize that the underdetermination of theories of consciousness could be a reasonable explanation of persistent disagreement in the scientific study of consciousness on these problems. After providing support for the hypothesis that the science of consciousness could be underdetermined, I will argue that underdetermination could be particularly problematic and persistent in the case of consciousness science, thereby explaining why underdetermination has not disappeared after all those years.

3.1. Theories of consciousness are preserved come what may

A popular view in philosophy of science is that scientific hypotheses are not tested in a vacuum but within a web of other hypotheses (Duhem, 1962). From this, it follows that a failed prediction leaves open the possibility of rejecting either the hypothesis that one wanted to test or some other background hypotheses. For example, to test a scientific hypothesis which makes a specific prediction about the temperature of my cup of coffee, I need to hold some beliefs about the reliability of my thermometer. Now, if my prediction fails, I could reject either the hypothesis that I wanted to test or the background hypothesis that my thermometer is reliable. What I should do in this case is *underdetermined* by the evidence (Stanford, 2017; Turnbull, 2018).

In consciousness science, theories of consciousness make predictions about what should or should not happen in certain conditions. For example, Lewes' theory of

consciousness makes the empirical prediction that subjects will continue to have tactile experiences when their midbrain is separated from their cortex, while Ferrier argues that it should not be the case. This prediction is made with the background hypothesis that subjects can typically report having subjective experiences when they have them. As seen in Section 2.3, when one sections the connection between the midbrain and the hemispheres, subjects do not report having any tactile experiences. Instead of rejecting his hypothesis, Lewes could reject the background belief that subjects can typically report having subjective experiences when they have them and argue that the case described by Ferrier is a case of consciousness without the possibility to report. Similarly, we saw that Fechner predicts that conscious after-images can be elicited by unseen objects, while Vulpian or Ireland make the opposite prediction. When Fechner reports a case in which an unnoticed object causes an afterimage, Ireland can simply reject the hypothesis that subjects typically remember what they consciously see, and argue that the object was actually consciously seen, but immediately forgotten, such that it created the illusion that the object was not seen. In these cases, scientists reject background hypotheses to save their preferred views of consciousness from being falsified.

Underdetermination was an important problem in the early study of consciousness. We also have reasons to believe that underdetermination is just as pervasive in the contemporary science of consciousness. Here, I provide just two telling examples of underdetermination in contemporary consciousness science:

First, Aaronson (2014a,b) provided (what he thought was) a decisive argument against the integrated information theory (IIT) of consciousness, destined to show that, according to the theory, a set of inactive logic gates can be conscious. But there is more: he demonstrated that, according to IIT, a set of inactive logic gates could have an arbitrarily high level of consciousness, provided that it is large enough. As a response, proponents of IIT accepted Aaronson's counterintuitive conclusion that a set of inactive logic gates could be conscious (Tononi, 2014). Accepting that sets of inactive logic gates are conscious, of course, seems to violate a large number of background hypotheses. IIT could be saved from Aaronson's challenge precisely because there was the possibility, for proponents of IIT, of rejecting one or several background hypotheses rather than the core of the theory. As such, proponents of IIT can continue to support the view that the physical basis of consciousness is the integration of information.

Second, in the debate over the existence of conscious feelings in fish, opponents of the existence of consciousness in fish argue that fish cannot be conscious, since they do not

have a cortex, and a cortex is necessary for having conscious experiences (Rose, 2014; Key, 2015, 2016). On the other hand, proponents of the existence of consciousness in fish have shown the existence of a wide variety of behaviors typically considered to be related to consciousness in these animals. For example, fish attempt to avoid stimuli that could damage their bodies, and these responses are reduced when fish are administered analgesics (Sneddon et al. 2003). Similarly, fish injected with harmful chemicals will move from enriched environments to barren tanks if the latter are filled with analgesics, thus indicating that fish search to relieve the pain (Sneddon 2011). The fact that fish exhibit these behaviors is thought to reinforce the view that fish have conscious experiences, at least conscious experiences of pain (see, e.g., Tye, 2017). In order to preserve the hypothesis that a cortex is necessary for the existence of conscious experiences, opponents to the existence of consciousness in fish reject the background hypothesis that pain-relieving behaviors indicate the presence of conscious experiences (for a review of this debate, see Michel, 2018).

The list of rejected background hypotheses in the study of consciousness could go on and on. To some extent, the short history of the study of consciousness I provided is just the story of background hypotheses rejected by consciousness scientists to save their preferred views of consciousness. To put it bluntly: if you don't like an empirical result suggesting that an entity is unconscious when your theory says that it should be conscious, there is always a way of arguing that this entity, which does not seem conscious, is in fact conscious, by rejecting some background hypotheses. Quine's claim that hypotheses can be preserved "come what may" seems especially well adapted to the study of consciousness (Quine, 1951). That theories of consciousness can be saved "come what may" might come from the fact that, in a sense, no hypothesis is held sacred in consciousness science[28]. Indeed, the history of the study of consciousness seems to indicate that no background hypothesis is too precious to reject: the hypothesis that systems that perform no interesting functions are unconscious, the hypothesis that conscious organisms know when they have experiences, or even the hypothesis that consciousness cannot be split when the brain is sectioned from the spinal cord. The problem is that, if no hypothesis is too costly to reject, everything is permitted: there is always a way of saving one's preferred theory. Consequently, I believe that the hypothesis according to which the science of consciousness has long been underdetermined by the evidence could provide a unifying explanation of the long-standing disagreement on both the identification of the physical bases of consciousness and the overflow debate.

[28] Even the hypothesis that phenomenal consciousness exists (see, e.g., Frankish, 2016).

Nevertheless, one could argue that scientific underdetermination is typically *transient,* because all responses to disconfirming evidence are typically not equally rational or supported by the evidence, or, at least, rarely remain so over long periods of time (Laudan, 1990; Laudan & Leplin, 1991). As such, it could seem implausible that underdetermination has persisted in the study of consciousness for more than 150 years. Moreover, this argument could also support the claim that, although underdetermination has been persistent in the history of the study of consciousness, there is no reason for thinking that it will persist in the future. If so, underdetermination in the study of consciousness would not be particularly problematic. I answer this argument in the last section by providing some reasons for thinking that underdetermination in the science of consciousness might be a persistent phenomenon.

3.2. Problems with detection rules

I now hypothesize that the reason why consciousness science could have remained underdetermined over such a long period of time is that consciousness scientists often disagree on what I call "detection rules". Here are several examples of detection rules: "subjects can typically report what they are conscious of", "when subjects are conscious, they know that they are", "when subjects are confident that they perceive something, they have a conscious perception of that thing". In consciousness science, detection procedures, that is, the type of procedures by which scientists produce judgments on the subjects' consciousness or unconsciousness of some contents during their experiments, use detection rules. For example, if a scientist makes use of a detection procedure relying on subjective reports to judge that a subject is conscious of a stimulus, her detection procedure is using the detection rule: "subjects can typically report what they are conscious of".

Not all detection procedures are created equal, because detection procedures use different detection rules. For instance, a detection procedure using the rule that subjects are conscious of a stimulus if their pupil reacts to the stimulus might not be as good as a detection procedure using the rule that subjects are conscious of a stimulus when they can report being conscious of it. Some detection rules seem more *reliable* than others. Here, I will consider that a detection rule is reliable if a detection procedure using that rule has a disposition to produce a large proportion of true judgments about the presence or absence of consciousness (of something) in a subject.

Detection rules have a special role in consciousness science, because one must rely on detection procedures to test hypotheses about consciousness. Indeed, theories of

consciousness typically make predictions about whether subjects are conscious or unconscious of contents in certain situations. For example, the global workspace theory predicts that, if a subject does not attend to an object in a crowded setting, she will be unconscious of that object (Dehaene & Changeux, 2011). One cannot verify whether the subject is conscious or unconscious of the stimulus in this situation without being able to detect the presence or absence of consciousness of a stimulus in that subject. Hence, when scientists test their hypotheses, they typically do so by relying on detection procedures using detection rules. As such, detection rules can be viewed as tools connecting empirical facts with hypotheses about consciousness. For example, Ferrier tested whether subjects would still be conscious of tactile sensations when a lesion separates the spinal cord from the brain. He observed that, in these conditions, patients claim that they do not have tactile sensations. Here, a patient's claim that she does not have tactile sensations can bear on Ferrier's hypothesis about the patient's *consciousness* of those sensations only if Ferrier uses a detection rule connecting subjective reports of having no sensations with the subject's (un)consciousness of those sensations. Such a detection rule could be: "when subjects report having no conscious experiences, they do not have conscious experiences". This example illustrates that, without detection rules, empirical facts remain silent about the subjects' consciousness. Hence, scientists must typically use detection rules when they test hypotheses about consciousness.

The problem is that, if detection rules are typically involved when scientists test hypotheses about consciousness, it is also typically possible for scientists to reject certain interpretations of the results by rejecting the detection rules used to test these hypotheses. In the case of Ferrier's experiment just mentioned, proponents of the spinal view could claim, and *did* claim, that the rule "when subjects claim that they do not have conscious experiences, they do not have conscious experiences" is unreliable. That is, they claimed that detection procedures using that detection rule fail to detect the presence or absence of consciousness, at least in that specific case. Consequently, because they rejected the detection rule used by Ferrier, proponents of the spinal view could reject his interpretation of the results and save their preferred hypothesis, namely, that the spinal cord could still elicit unreported conscious tactile experiences. Hence, in consciousness science, it is typically possible to save one's preferred hypothesis in front of (seemingly) disconfirming evidence by rejecting the detection rule used to test it. Consequently, the systematic possibility of rejecting detection rules is a *prevalent* factor of underdetermination in consciousness science.

Nevertheless, the systematic possibility of rejecting detection rules, by itself, does not explain why underdetermination is persistent throughout the history of the study of consciousness. Indeed, I have suggested that scientific underdetermination is usually transient because all responses to disconfirming evidence are typically not equally rational. To put it bluntly: there are *cheap* and *costly* ways of saving hypotheses. Remember the thermometer example: if, on the basis of some hypothesis, I predict that the temperature of a certain object should be of exactly 20°C, and it turns out that my thermometer indicates 18°C, I can either reject my hypothesis, or save my hypothesis by claiming that the measurement procedure relying on a thermometer is unreliable. Presumably, if I don't have any independent reasons to doubt that the measurement procedure is reliable, and if I can use a wide variety of thermometers, all indicating 18°C, the most rational thing to do is to reject my hypothesis. Indeed, in that case, rejecting the reliability of all measurement procedures using thermometers would be extremely *costly*, because a wide variety of theories and everyday practices depend on the reliability of these procedures. On the other hand, if I had good reasons to believe that measurement procedures using thermometers are often unreliable, saving my hypothesis by rejecting the reliability of these procedures would be cheap.

I have shown that disagreement over which detection rules to use is quite common in consciousness science throughout its history. For example, in the overflow debate, proponents of the view that consciousness overflows reportability argue that the rule "when subjects are conscious of something, they can report being conscious of that thing" is not reliable, while opponents to that view claim that it is. If scientists disagree on which detection rules to use, rejecting them is less costly than if they were largely accepted within the scientific community. In other words, disagreement over which detection rules to use lowers the price of rejecting them. Consequently, I hypothesize that, because consciousness scientists often disagree on which detection rules are reliable, rejecting these rules typically comes at almost no price. In a slogan: no detection rule is too big to fail. As a result, what one should do in front of (seemingly) disconfirming evidence is underdetermined: if one gives more credence to one's hypothesis than to the reliability of the detection rule used to test it, rejecting the detection rule to save one's hypothesis is *prima facie* not an irrational thing to do. Hence, the rejection of detection rules could foster underdetermination because it provides a cheap way of saving one's preferred hypotheses.

One solution to settle disagreements over which detection rules to use could be to attempt to demonstrate the reliability of detection rules themselves. To do so, one could provisionally take a detection rule as a hypothesis and attempt to confirm it. For example, a

confirmation of the hypothesis that “when subjects perform a rational action based on a sensory cue, they are conscious of that cue” could allow us to justify using the corresponding detection rule. In that case, rejecting this detection rule would be more costly.

However, a problem emerges when one attempts to increase the price of rejecting detection rules in this way. Indeed, I argued, first, that consciousness scientists use detection rules to test hypotheses, and, second, that consciousness scientists typically disagree on which detection rules to use. As such, testing a hypothesis corresponding to a detection rule will necessitate the use of other detection rules, and will itself be open to underdetermination. To illustrate this, imagine that we want to confirm the following hypothesis: “when subjects have conscious visual sensations, they can report having those sensations”. To do so, we need a way of knowing when subjects are conscious or unconscious of sensations that would not rely on reports. In turn, this implies relying on a detection rule, the use of which might itself turn out to be contentious. Consequently, the underdetermination that pervades when testing hypotheses in consciousness science might also apply when detection rules themselves are taken as hypotheses. Many authors, as Cohen & Dennett (2011), or Kouider et al. (2010), doubt that the overflow debate could be solved empirically for precisely this reason: it is unclear what detection rule could be used to assess the presence or absence of consciousness in order to test the hypothesis that subjects can report what they are conscious of. Hence, taking detection rules as hypotheses to show that these detection rules should be considered reliable is unlikely to provide a straightforward way of increasing the cost of their rejection. And if raising the cost of rejection of detection rules is quite difficult, underdetermination is all the more persistent.

In sum, I have suggested that underdetermination in consciousness science is persistent, first, because the systematic possibility of rejecting detection rules is a *prevalent* factor of underdetermination, second, because these rules are *cheap to reject*, and, third, because it is *difficult to show that detection rules should be considered reliable*. These three factors combined lead to a situation in which one can *typically* save one’s preferred hypothesis by rejecting a detection rule *at almost no price,* with *little hope for agreement* on detection rules. I suggest that, in such situation, one should expect underdetermination to be persistent in consciousness science. In turn, I have argued that persistent underdetermination could be a unifying explanation for long-standing lack of consensus in the field.

To be clear, I do not claim that there is nothing that scientists could do to reduce underdetermination in the future, or that they could not ultimately come to an agreement on

which detection rules should be considered "too big to fail"[29]. Instead, my claim is that the underdetermination of the study of consciousness by the evidence is a good explanation of the long-standing lack of consensus on problems like that of finding the physical basis of consciousness or knowing whether subjects perceive more than they can report.

Conclusion

Throughout the 19th century, scientific debates surrounding consciousness have remained unsettled. The lack of consensus observed in the contemporary science of consciousness is also a feature of the early study of consciousness. I hypothesized that long-standing disagreements in consciousness science could be explained by the underdetermination of theories of consciousness. Indeed, the history of the study of consciousness leaves us with the impression that the early science of consciousness was like a game whose rules could be changed at will, and in which theories could be preserved "come what may". I suggested that the contemporary science of consciousness does not reach consensus either, which might indicate that our contemporary theories could be similarly underdetermined by the evidence. Finally, I hypothesized that debates over which detection rules to use could explain why underdetermination has been so persistent in consciousness science.

Acknowledgments

I thank Anouk Barberousse, Pascal Ludwig, Keith Frankish, Adrien Doerig, Hakwan Lau, Liz Irvine, Mazviita Chirimuuta, and two anonymous reviewers for their helpful comments.

[29] As an anonymous reviewer remarked, there *is* some agreement on elementary detection rules in consciousness science, for instance, the rule that positive introspective reports of the kind "I saw the stimulus" should be interpreted as indicating consciousness of the stimulus, in normal conditions. As such, there is at least a relative consensus that positive introspective reports are reliable. The problem here is to find the best way for the subjects to make these reports (Michel, forthcoming; Sandberg et al., 2010), such as to avoid report biases (Irvine, 2012). Perhaps developing so-called "bias-free" detection procedures based on introspective reports would lead to further progress in consciousness science (see, e.g., Peters & Lau, 2015).

References

Aaronson, S. (2014a). Giulio Tononi and Me: A Phi-nal Exchange. https://www.scottaaronson.com/blog/?p=1823.

Aaronson, S. (2014b). Why I am not an integrated information theorist (or, the unconscious expander). www.scottaaronson.com/blog/?p=1799.

Bain, A. (1884) *Mental and Moral Science*, vol. 1. 3rd ed. London: Longmans, Green & Co.

Barton, R. (1987). John Tyndall, pantheist: A rereading of the Belfast address. *Osiris*, 3, 111–134.

Bastian, H. C. (1869). Sensation and Perception I. *Nature*, *1*, 213–214.

Bastian, H. C. (1870a). Sensation and Perception II. *Nature*, *1*, 309–311.

Bastian, H.C. (1870b). Consciousness. *Journal of Mental Science*, 15, 501-523.

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.) New Jersey: Prentice-Hall.

Boudewijnse, G. J., Murray, D. J., & Bandomir, C. A. (1999). Herbart's mathematical psychology. *History of Psychology*, *2*(3), 163–193.

Block, N. (1995). On a Confusion about a Function of Consciousness. *The Behavioral and Brain Sciences*, *18*(2), 227–247.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *The Behavioral and Brain Sciences*, *30*(5–6), 481-499; discussion 499-548.

Block, N. (2016). The Anna Karenina Principle and Skepticism about Unconscious Perception. *Philosophy and Phenomenological Research* 93 (2):452-459.

Calderwood, H. (1879). *The relations of mind and brain*. London: Macmillan & Co.

Carpenter, W. B. (1864). *Principles of human physiology*. London: John Churchill.

Carpenter, W. B. (1874). *Principles of mental physiology*. London: Henry S. King & Co.

Carruthers, P. (2015). *The Centered Mind*. Oxford: Oxford University Press.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chalmers, D. (2000). What is a Neural Correlate of Consciousness ? In T. Metzinger (Ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Issues*. MIT Press.

Chirimuuta, M. (2017). Hughlings Jackson and the "doctrine of concomitance": mind-brain theorising between metaphysics and the clinic. *History and Philosophy of the Life Sciences*, *39*(3), 26.

Clarke, J. (1888). *Monkeys' brains once more: Schaefer v. Ferrier.* London: Victoria Street Society United with the International Association for the Protection of Animals from vivisection.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience ? *Trends in Cognitive Sciences*, *20*(5), 324–335.

Condillac, E. B. (1743). *Essai sur l'origine des connaissances humaines*. Amsterdam: Pierre Mortier.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–185.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences*, *16*(3), 241–262.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758.

Dehaene, S., & Changeux, J. P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227.

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*(1947), 76–84.

Descartes, R. (1985). *The Philosophical Writings of Descartes*. Translated by John Cottingham, Robert Stoothoff, Dugald Murdoch and Anthony Kenny (vol. 3). 3 vols. Cambridge: Cambridge University Press. **[CSM]**

Duhem, P. (1962). *The aim and structure of physical theory.* (P. Wiener, Trans.). New York: Atheneum.

Dunn, R. (1858). *An Essay on Physiological Psychology*. London: John Churchill.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Ferrier, D. (1876). *The Functions of the Brain*. London: Smith, Elder and Co.

Finn, M. A., & Stark, J. F. (2015). Medical science and the Cruelty to Animals Act 1876: A re-examination of anti-vivisectionism in provincial Britain. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *49*, 12–23.

Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M.Y., Kramer, U., Fried, I., Malach, R. (2009). Neural “Ignition”: Enhanced Activation Linked to Perceptual Awareness in Human Ventral Stream Visual Cortex. *Neuron*, *64*(4), 562–574.

Fishman, R. S. (1995). Brain wars: Passion and conflict in the localization of vision in the brain. *History of Ophtalmology*, 89(1-2), 173-184.

Flourens, P. (1842). *Recherches expérimentales sur les propriétés et les fonctions du système nerveux dans les animaux vertébrés.* Paris: J-B Baillière.

Foster, M. (1890). *A textbook of physiology*, London: Macmillan & Co.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*. 23(11-12), 11-39.

Fritsch, G., & Hitzig, E. (1870). Über die elektrische Erregbarkeit des Großhirns. *Archiv für Anatomie, Physiologie und Wissenschaftliche Medicin*. 300–332.

Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., Cohen, L., Naccache, L. (2009). Converging Intracranial Markers of Conscious Access. *Plos Biology*, *7*(3), 0472–0492.

Glickstein, M. (1985). Ferrier’s mistake. *Trends in Neurosciences*, *8*(C), 341–344.

Hamilton, W. (1836). *Lectures on Metaphysics and Logic.* 4 vols. Edinburgh: William Blackwood.

Herbart, J. F. (1816/1964). *Lehrbuch der Psychologie*. In K. Kerbach & O. Flügel (Eds.). *Johann Friedrich Herbart’s Sämmtliche Werke*. Aalen: Scientia Verlag.

Herzen, A. A. (1886) *Les Conditions physiques de la conscience*, Genève: H. Stapelmohr.

Höffding, H. (1891). *Outlines of Psychology*. Tr. by M. E. Lowndes. New York: Macmillan.

Hughlings Jackson, J. (1931). Selected writings of John Hughlings Jackson. In J. Taylor (Ed.), *On epilepsy and epileptiform convulsions* (Vol. 1). Birmingham, AL: Gryphon Editions.

Ireland, W. (1875). Can Unconscious Cerebration be proved? *Journal of Mental Science, 21*(95), 366-387.

Irvine, E. (2012). *Consciousness as a scientific concept*. Springer.

James, W. (1890). *Principles of Psychology*. New York: Henry Holt.

Key, B. (2015). Fish do not feel pain and its implications for understanding phenomenal consciousness. *Biology and Philosophy*, *30*(2), 149–165.

Key, B. (2016). Why fish do not feel pain. *Animal Sentience*, *3*, 1–33.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14*(7), 301–307.

Kitcher, P. (1995) *The Advancement of Science Science Without Legend, Objectivity Without Illusions.* Oxford: Oxford University Press.

Kim, A. (2015) Johann Friedrich Herbart, *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.).

Klein, A. (2017). The curious case of the decapitated frog: on experiment and philosophy. *British Journal for the History of Philosophy*, *26*(5), 890–917.

Kulstad, M. (1991) *Leibniz on Apperception, Consciousness, and Reflection*. Munich: Philosophia Verlag.

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220.

Lamme, V. A. F. (2015). *The Crack of Dawn*. *Open MIND* (Vol. 22).

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373.

Laudan, L. (1990). Demystifying Underdetermination, in *Scientific Theories*, C. Wade Savage (ed.), (Series: Minnesota Studies in the Philosophy of Science, vol. 14), Minneapolis: University of Minnesota Press, pp. 267–297.

Laudan, L., & Leplin, J. (1991). Empirical equivalence and underdetermination. *The Journal of Philosophy*, 88, 449–472.

Laycock, T. (1860). *Mind and Brain or, the correlations of consciousness and organisation*. New York: D. Appleton and Co.

Leibniz, G. (1704/1996). *Leibniz: New Essays on Human Understanding*. Eds. P. Remnant & J. Bennett. Cambridge: Cambridge University Press.

Leibniz, G. (1714/1965). *Principles of nature and of grace, based on reason.* In P. and A. M. Schrecker (Eds.), *Monadology and other philosophical essays.* Indianapolis: Bobbs-Merrill.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, *64*, 354–361.

Lewes, G. H. (1873). Sensation in the spinal cord. *Nature*, *9*, 83–84.

Lewes, G. H. (1876). Ferrier on the brain. *Nature*, *15*, 73–74.

Lewes, G. H. (1879). *Problems of Life and Mind* (Third Series). London: Trübner and Co.

Liddell, E. G. T. (1960). *The discovery of reflexes*. Clarendon Press.

Longet, F. A. (1842). *Anatomie et physiologie du système nerveux de l'homme et des animaux vertébrés*. Paris: Fortin, Masson et co.

Maudsley, H. (1867). *The Physiology and Pathology of the Mind*. New York: D. Appleton and Company.

Maudsley, H. (1887). The physical conditions of consciousness. *Mind*, *48*, 489–515.

McRae, R. (1976). *Leibniz: Perception, Apperception, and Thought.* Toronto: University of Toronto Press.

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30(1), 63–81.

Michel, M. (2018). Fish and microchips: on fish pain and multiple realization. *Philosophical Studies*.

Michel, M. (Forthcoming). The Mismeasure of Consciousness: A problem of coordination for the Perceptual Awareness Scale. *Philosophy of Science*.

Michel, M., & Morales, J. (Forthcoming). Minority Reports: Consciousness and the Prefrontal Cortex. *Mind & Language*.

Mill, J. (1869). *Analysis of the phenomenon of the human mind*, 2nd Edition, London: Longmans, Green & Co.

Mill, J.S. (1865). *An Examination of Sir William Hamilton's Philosophy*. Boston: William V. Spencer.

Morell, J.D. (1862). *An introduction to mental philosophy, on the inductive method.* London : Longman, Green & Co.

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

Noble, D. (1858). *The human mind in its relation with the brain and nervous system*. London: John Churchill.

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *The Journal of Neuroscience*, *37*(40), 9593–9602.

Pearce, J. M. (2003). Sir David Ferrier MD, FRS. *Journal of Neurology, Neurosurgery, and Psychiatry*, *74*(6), 787.

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *ELife*, *4*(October), 1–30.

Peters, M. A. K., Kentridge, R. W., Phillips, I., & Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, *3*(1), 1–11.

Pflüger, E. (1853). *Die sensorischen Functionen des Rückenmarks der Wirbelthiere: nebst einer neuen Lehre über die Leitungsgesetze der Reflexionen.* Berlin: Hirschwald.

Phillips, I. (2018). The Methodological Puzzle of Phenomenal Consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences*, 3, 73–83.

Pierce, A. H. (1906). Should we Still Retain the Expression “Unconscious cerebration” to Designate certain Processes Connected with Mental Life? *The Journal of Philosophy, Psychology and Scientific Methods*, *3*(23), 626–630.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60,20–43.

Romand, D. (2012). Fechner as a pioneering theorist of unconscious cognition. *Consciousness and Cognition*, *21*(1), 562–572.

Rose, J. D., Arlinghaus, R., Cooke, S. J., Diggles, B. K., Sawynok, W., Stevens, E. D., & Wynne, C. D. L. (2014). Can fish really feel pain? *Fish and Fisheries*, *15*(1), 97–133.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078.

Schiff, M. (1858). *Lehrbuch der Physiologie des Menschen.* Lahr: Schauenburg & C.

Schwitzgebel, E. (2007). Do You Have Constant Tactile Experience of Your Feet in Your Shoes? Or Is Experience Limited to What’s in Attention? *Journal of Consciousnes Studies*, *14*(3), 5–35.

Simmons, A. (2001). Changing the Cartesian Mind: Leibniz on Sensation, Representation and Consciousness. *Philosophical Review*, *110*(1), 31–75.

Simmons, A. (2011). Leibnizian consciousness reconsidered. *Studia Leibnitiana*, 43(2), 196–215.

Sneddon, L. U., Braithwaite, V. A., & Gentle, M. J. (2003). Do fishes have nociceptors? Evidence for the evolution of a vertebrate sensory system. *Proceedings of the Royal Society B: Biological Sciences*, 270(1520), 1115–1121.

Sneddon, L. (2011). Pain perception in fish: Evidence and implications for the use of fish. *Journal of Consciousness Studies*, 18(9–10), 209–229.

Spencer, H. (1868). *Principles of Psychology*. London: Williams and Norgate.

Stanford, K. (2017) "Underdetermination of Scientific Theory", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.).

Turnbull, M. G. (2018). Underdetermination in science: What it is and why we should care. *Philosophy Compass*, *13*(2), 1–11.

Vulpian, A. (1866). *Leçons sur la physiologie générale et comparée du système nerveux*. Paris: Baillière.

Tononi, G. (2014). Why Scott should stare at a blank wall and reconsider (or, the conscious grid). https://www.scottaaronson.com/blog/?p=1823.

Tononi, G., & Koch, C. (2015). Consciousness: Here, There, and Everywhere ? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *31*(1), 12–19.

Tye, M. (2017). *Tense bees and shell-shocked crabs: Are animals conscious?* Oxford: Oxford University Press.

Tyndall, J. (1872) *Fragments of Science*, vol. 2., New York: P. F. Collier & Son.