Is the experience of agency necessarily retrospective? A predictive approach

Examination number: B113150

MSc Philosophy

The University of Edinburgh

2018

Contents

Introduction

In daily life, we normally assume that our actions are our own, that we cause them, and even that they are the effects of our conscious commands. When asked about the reasons for our actions, we may give answer like "Because I intended to" or "Because I thought it was the right thing to do". Concepts such as free will and intentional action forma part of almost every person's conception of their own activity. Even our legal and most ethical systems presuppose the causal efficacy of mental states in that causation of action, in terms like responsibility and punishment.

However, the demonstration that this ideas are true is far from simple, and in the last decades several scientific studies from the areas of psychology, cognitive science, neuroscience and neurology, have provided evidence that suggests that we may not be the masters of our actions as we once thought. First, the idea of a conscious will that decides on action and starts it has been largely rejected, both from the trenches of science and philosophy. Arguments against the will abound: from the violation of the conservation of energy to the impossibility of interaction between substances. There are some that try to resist them and defend a the position of libertarianism, but the evidence against it is overwhelming.

Benjamin Libet's work in the 1970s and 80s started a tradition of investigation concerning the role of neural activity in intentional behaviour. According to Libet (1985), his experiments showed that conscious intention cannot be considered the cause of voluntary action, because the emergence of the intention relative is consistently preceded by preparatory neural activity in the motor cortex of the brain. Libet argued that this results demonstrated that the neural activity is responsible for voluntary action, and that consciousness may only become aware of the intention once unconscious processes have started it.

Nonetheless, once conscious will is taken out of the picture the main evidence that is presented in its favour still remains. When we act intentionally, we seem to experience, to perceive that our action were caused by our intentions to perform them. To cite a famous example, if I give my arm a command to rise, it rises. This experience, according to libertarianism, is sufficient evidence to assert that it is our commands, our intentions, that cause action.

Even this *prima facie* acceptable evidence has been severely put to question by scientific research. Perhaps the most famous argument is Daniel Wegner's (1999) Illusion argument, which considers that the evidence from experience to support conscious will is illusory. We accept it because, he says, we assume that the experience is an effect of the causation of action by the will. Yet, there is no evidence that indicates that the experience has a causal link to intentional action. Instead, he argues, the experience is nothing but an inference that is made retrospectively, after the action has taken place, based on our continued perception of conscious thoughts followed by actions.

Nevertheless, his argument is not entirely conclusive. In the present work, I shall defend that although there is undoubtedly a retrospective component as a source of the experience of agency (that is, of goal-directed action), that does not imply that it may have another component that is determined by an actual link in the causal chain that leads from thought to action. Moreover, I shall defend that such a link may actually exist, and I will attempt a brief description of what it could be. Hence, in Section 1 I will begin by examining Daniel Wegner's Illusion argument as a token argument of the perspective that the experience of agency is retrospective inference, and argue that he does not demonstrate that the experience of agency is not causally connected to intentional action (sections 1.1 and 1.2). In Section 2 I will review the evidence gathered in the research surrounding the intentional binding effect, which will serve as positive proof that a retrospective component is not sufficient to account for the sense of agency.

Section 3 will be dedicated to a brief description of the Predictive Processing framework, and in 3.1 I will analyse the Comparator model for the experience of agency, which supports retrospective inference and is based of Predictive Processing. Finally, in Section 4 I shall discuss Daniel Dennett's brilliant account on qualia as predictive phenomena (4.1), and in section 4.2 I will explore whether Dennett's ideas may shed light on the nature of the sense of agency.

As an introductory note, I need to clarify my uses of three expressions: 1) the sense of authorship I the sensation that one is the cause of the action that one performs; 2) by the sense of agency I will denote the hypothesised feeling one might experience when an action has been indeed caused by a conscious thought or intention; and 3) under the term experience of agency I will consider every experience about agency that has a retrospective component, regardless of whether a predictive factor may also be involved. Additionally, it may be necessary to make clear that I do not intend to defend the

possibility of the conscious will's causal efficacy, nor even its existence. My contention is that conscious thought may still have a causal role in the production of behaviour, even in an entirely physical world.

1. Daniel Wegner's Illusion argument

Wegner displays a powerful argument against the notion that conscious will is responsible for the causation of actions, based on the analysis of a wide array of empirical data. Nonetheless, the claim I shall contest in the current Section is not that the belief in a conscious will as the source of action is illusory; rather, what I believe is mistaken in Wegner's argument is the conclusion that the sense of agency is not indicative of a causal link between conscious thought and action.

The Illusion argument rests on three main premises: (1) according to Wegner and Wheatley (1999), Benjamin Libet's experiments on the timing of conscious intentions and voluntary actions[1] showed that intentions are not the causes of actions, but that instead actions and intentions are both the products of unconscious brain processes, which may or may not be themselves causally linked. (2) The experience of agency can occur even in absence of a conscious thought as its probable cause, for instance, in cases when action is induced by Transcranial Magnetic Stimulation. (3) Inversely, endogenous actions that could be classified as voluntary are often not accompanied by an experience of agency (e.g. motor automatisms, Ouija board manipulation, hypnotism) (Wegner and Wheatley, 1999, pp. 482-483).

With the evidence from premises (2) and (3) Wegner and Wheatley (1999) intend to demonstrate that the experience of agency is not a reliable evidence of mental causation, because (from 2) it is not necessary that intentions cause actions for a sense of agency to arise, nor are voluntarily caused actions sufficient to bring it about (3). Thus, and since unconscious processes constitute an additional probable cause for both intentions and actions, Wegner and Wheatley (1999) believe that we are faced with "the psychological equivalent of the third variable problem of analysis […] [because] we can never be sure that our thoughts cause our actions, as there could always be unconscious causes that have produced them both" (Wegner and Wheatley, 1999, p. 482).

Wegner and Wheatley (1999) adopt this line of reasoning and propose that intentions and actions are indeed caused by separate unconscious processes that could share causal connections. That is, that while the independent unconscious processes that bring about intentions and actions

[1] During the 1970s and 80s, neurophysiologist Benjamin Libet carried out a series of experiments whose objective it was to measure the timing of a conscious intention to perform an action in relation to the neurophysiological processes that cause the action. As a result, he found that awareness of an intention is consistently preceded by an increase of preparatory neural activity in the motor cortex (called Readiness Potential), about 350 milliseconds prior to the action. Libet claimed these findings suggested that the brain has settled on a course of action before the intention to perform it is brought to consciousness, and so consciousness is not the cause of action, except for the possibility of vetoing it at the last moment (Libet, 1985). Both Libet's methodology and his interpretations of the results have been widely discussed and challenged, for instance, by Daniel Dennett (1992). I will briefly describe his critiques in Section 2.

may share causal paths, their effects are not causally connected. This, however, opens a rather important question: If there is no causal path that leads from thought action, how is it that we come to believe that such a path exists? In other words, why do we experience our actions as caused by our thoughts?

Following Hume, Wegner and Wheatley (1999) argue that it is not a causal link between thought and action that causes us to perceive causation, but that we attribute causation because we recurrently have experiences of thoughts that are prior to, consistent with, and which appear to be the exclusive cause of our actions. In light of this constant concurrence we retrospectively infer mental causation. Hence, agency is experienced when a thought occurs prior to an action, although not too far in advance nor to close to the action; when a thought is semantically associated with the action, and the strength of the experience varies according to the awareness of probable alternative causes other than the thought.

Crucially, Wegner and Wheatley (1999) argue, interference or manipulation of these variables can cause anomalous experiences of agency. For example, priming a subject with a consistent image prior to an action has been found to strengthen the experience, while studies in confabulation show that subjects tend to retroactively attribute themselves thoughts that justify choices they did not originally make. On the other hand, failure to perceive priority, exclusivity or consistency may lead a person to attribute endogenously generated actions to external agents. Dowsers, for instance, believe that external forces direct their dowsing rods towards the locations where water can be found, when in reality the shifts in the rods' orientation is caused by imperceptible, automatic movements of the dowsers' hands, called motor automatisms (Wegner and Wheatley, 1999, pp. 483-487).

1. 2 The negative argument

In summary, the Illusion argument consists of two parts: a negative one, designed to show that the experience of agency is not a product of mental causation of action and, hence, an unreliable indicator of the latter; and a positive one, which describes the alternative mechanisms that would be responsible for the belief that conscious thoughts cause action. So, the negative part is comprised of a representative argument for mental causation, and its rejection. This argument would have roughly the following form: if conscious will exists and causes action, then, whenever a voluntary action is performed, it is accompanied by an experience of agency; thus, the experience is an indication that an action has been consciously willed. The strategy to reject it is

relatively straightforward, it suffices to prove that the consequent of the conditional is false in order to conclude that conscious will does not cause behaviour.

As discussed above, Wegner and Wheatley (1999) provide evidence that supports the claim that mental causation of action is neither necessary nor sufficient to elicit an experience of agency. However, an analysis of may show that it may not warrant such a strong conclusion.

Concerning the contention about necessity, one must consider that most evidence comes from experimental conditions that may affect the results in different manners. First, in cases like the *I Spy* experiment (Wegner and Wheatley, 1999, pp. 487-489), the experimenters force a post-action judgement of the experience of agency, thus introducing a retrospective component that is not part of the experience of the action. Because of this judgement it is unclear whether the phenomenon being evaluated in the experiment is the actual experience of the action or the judgement itself, which is inevitably retrospective. Moreover, also in the *I Spy* experiment as well as others cited (Wegner and Wheatley, 1999, pp. 483-487), the experimental conditions need that the subjects believe, before the actions that are being evaluated are carried out, that they indeed have control over the outcomes of said actions. Such a belief may either affect the post-action judgement of the experience or the experience itself, especially because the conditions are designed so that the subject has no clear indication that her action is not the actual cause of the outcome. Such examples could themselves be thought of as experimentally induced illusions of agency, that do not necessarily indicate the illusory nature of the experience or sense of agency in general. Nor should they be taken as evidence that the sense of agency is merely inferred, for they could be accounted for as the result of predictive processes without the need of a recourse to inference (this point will be further discussed in Sections 3 and 4).

Furthermore, in the cases of priming in the *I Spy* experiment and of the TMS-induced experience of agency, the experimenters could actually be "tapping" into the non-retrospective processes that could be responsible for the sense of agency. The case of priming could indicate that having a prior representation that is consistent with an impending action is an actual requisite for a sense of agency to be produced. Similarly, the fact that Transcranial Magnetic Stimulation of the motor cortex may produce a sense of agency could point to the location where the experience is generated.

On the other hand, the assertion that mental causation of action is not a sufficient cause of the sense of agency can be subjected to similar objections. For example, in the cases of hypnosis and action projection (Wegner and Wheatley, 1999, p. 482), the fact that they are driven by strong suggestion (in the former) and the belief that the subject is not the sole cause of the action (in the latter) may suggest as well that prior high-order expectations play a part in shaping the experience (for further discussion, see Section 3).

1.3 The positive argument

To sum up, the positive part of the Illusion argument, that the experience of agency is achieved through a retrospective inference from the perception of a constant concurrence of an action and a prior and constant though that appears to be the exclusive cause of the former, is based on two main claims: (1) the experience of agency is not caused by a causal connection between thought and action, nor is it a perception of such a link; and (2), unconscious processes are the actual causes of both conscious thought and action, and while they may themselves be causally connected, they are not consciously perceived.

The truth of claim (1) was the subject of the previous Section, as it is the conclusion of the negative side of the Illusion argument. Therefore, I now turn to the analysis of claim (2). It rests on two further contentions: (a) if (2) is true, then the experience of agency is not trustworthy evidence that action is caused by thought; and (b), the results of Libet's experiments indicate that "brain events cause intention and action, whereas conscious intention itself may not cause action" (Wegner and Wheatley, 1999, p. 481). Independently of the reliability of Libet's experiments (see for example section 2), I believe (b) is highly problematic. Wegner and Wheatley (1999) seem to imply that the distinctions between actions and their neural causes, on one side, and between intentions and their neural causes, on the other, are analogous. Yet, the similarity is far from evident. Whereas action can be distinguished from its neural causes, as the former is the latter's overt consequence and involves other non-neural body systems, the difference between a conscious thought and its unconscious causes is not as readily available.

Although they may appear to be, from the manner in which they are made manifest to us, two entirely different phenomenon, under our modern scientific perspective they are both neural events. Their relationship need not be analogous to that of causality, but supervenience**,** and they (unconscious cause and conscious thought) could be identical.

Yet, when Wegner and Wheatley talk about "the brain events that cause intention" it is unclear whether they are referring to an unconscious brain event that *precedes* and causes the brain event that is the conscious thought or intention, or to the sub-personal neural event that *underlies* it and to which it is token-identical. Firstly, it should not be surprising that conscious brain events are caused by unconscious brain processes, nor that unconscious processes may mediate a possible link between conscious thought and action. It cannot be conceived from the fact that conscious brain events may have unconscious brain events as causes, that conscious thoughts are effectively removed from the causal chain that leads to action.

Second, if the claims is that it is sub-personal and unconscious brain events that underlie conscious thought cannot be the causes of action, it may still refer to one of two positions. On the

one hand, it may be argued that conscious thoughts are epiphenomenal. Yet, this position seems to confuse what is expected to be causally efficacious for intentional action, the intention itself or the experience of the intention that accompanies it. As Bayne (2006) points out: "I might experience my hand as rising in virtue of my intention to raise it, but I do not experience my hand as rising in virtue of my *experience* of my intention to raise it" (p.177). So even if we granted that the experiential features are epiphenomenal, as long as the underlying causes of conscious thought are what is doing the work, we may not infer that they are not causally connected to action. On the other hand, the claim may simply be they conscious thought (and its underlying neural causes) and the neural processes that lead to action are two sets of brain events that do not maintain causal interactions; that is, that the neural processes that constitute the conscious experience of intention are causally isolated from the neural causal pathway that lead to action. This, however, is a strong empirical claim that would require equally strong empirical evidence if it is to be accepted. Such evidence is not addressed by Wegner.

With the previous discussion I hope to have given enough reason to believe, not that mental causation is indeed a necessary and sufficient cause of the sense of agency, but that the opposite, that it is neither sufficient nor necessary, is not sufficiently proven by the Illusion argument. If this is true, then much of the evidence Wegner and Wheatley (1999) provide could be interpreted as instances of induced illusion of an otherwise not illusory experience.

2. Intentional binding as a measure of agency

As well as line of research dedicated to the operationalisation of intentional action, neuroscientist Patrick Haggard's work on the intentional binding effect can partly be understood as an attempt to avoid some of the methodological errors present in Libet's investigations, for he himself worked in that line of research, contributing with some findings of his own (see, for example, Haggard and Libet, 2001). As mentioned in the previous Section, Benjamin Libet carried out a series of experiments with the goal of determining the sequence of neurophysiological events that take place when a person forms an intention or makes a decision to perform an action. Libet's subjects were therefore connected to an electroencephalogram and an electromyogram, which would provide readings of the brain's activity and the movements of their arm's muscles, thus establishing a measurement of the events in an objective timeline. Meanwhile, the subjects were instructed to perform a flick of their wrist or finger, taking care to notice the moment when they first became aware of the intention to do so, via a clock mounted in front of them. These reports from the subjects provided the second, subjective timeline of events. The results of both timelines were compared afterwards, showing that brain activity (a "Readiness Potential") consistently started about 350 milliseconds before the time in which the subject's reported having become aware of the intention to act. Libet interpreted these results as evidence that consciousness does not start voluntary action, but rather, it is caused by unconscious neural processes (Libet, 1983).

Libet's work has had many supporters (including Haggard himself), but it has also gathered a large amount of detractors. Among them, Daniel Dennett's critiques are especially relevant for Haggard's later choice of experimental designs. Dennett and Kinsbourne (1992) argue not against Libet's results or interpretations thereof, but against his methodology. According to them, the flaw of Libet's experimental design rests on the assumption that subjective and objective time are both linear and commeasurable, maintaining a one-to-one relationship in the order of events. Yet there is evidence that demonstrates that the sequence of events as experienced in consciousness does not necessarily match the objective sequence of events (for instance, the "phi phenomenon" (Dennett and Kinsbourne, 1992, p. 186)). So, rather than reflecting the objective order of events, the order in which they are experienced is determined by rules relevant to the brain's own functioning and cognitive processes.

2.1 Experiments, methodology and results

Thus, Haggard set out to devise a set of experiments that avoided the pitfall of Libet's temporal commensurability assumption, but that could still provide insights into the sequence of the events that make up the experience of voluntary action. So instead of basing their analysis on a comparison between objective and subjective timelines, Haggard *et al.* (2002) used "interval estimates experiments" (p. 382) to determine whether perceived times of actions and their effects vary between voluntary and involuntary action. In this manner, experienced time is not compared to objective time, but to additional instances of perceived time, allowing more appropriate inferences concerning the phenomenology of agency.

The first experiment comprised three operant tasks. In the first one, subjects were instructed to perform a voluntary action by pressing a button, which was followed by an auditory stimulus (a tone) 250 milliseconds after the action. In the second task, subjects received stimulation in the motor cortex by means of Transcranial Magnetic Stimulation (TMS), which induced a twitch of the finger, and was also followed by a tone 250 milliseconds afterwards. Finally, in the last task the subjects heard the activation of the TMS, which stimulated the parietal cortex and could not cause any twitching, and heard the tone 250 milliseconds later. In all three tasks subjects were instructed to provide estimations of the moments in which they became aware of the voluntary actions, the TMS-induced twitches, the "sham TMS" sound, and the tones, using a clock (Haggard *et al*., 2002, p. 382).

Furthermore, as a baseline to which the latter reports could be compared, subjects were asked to time the occurrence of voluntary actions, TMS-stimulation, "sham TMS", and tones, independently from one another and from the operant conditions. The baseline results "indicated a roughly accurate awareness of the voluntary action, delayed awareness of the involuntary TMS-induced twitch, and intermediate values for sham TMS and for auditory tones" (Haggard *et al*., 2002, p. 382).

The results of the operant condition show evidence of "large perceptual shifts" (Haggard *et al*., 2002, p. 382) in the timing of the couples action-tone, and TMS-tone, whereas the estimations of sham TMS-tone occurrence did not vary, relative to the baseline condition. When a TMS-induced twitch was followed by a tone, the twitch was perceived as occurring earlier in time and the tone later than the baseline. As for the voluntary task, the reported estimations located it closer to the tone (that is, later than the baseline) and, conversely, the tone was perceived to have happened closer to the action (Haggard *et al*., 2002, pp. 382-383). Haggard *et al.* (2002) called this effect "intentional binding", for only intentional action seems to produce this attraction in the experience of action and its effect.

In the second experiment Haggard *et al.* (2002) tested the effects that temporal variations between action and their effects could have on the intentional binding. The tasks consisted, as in the previous experiment, of intentional actions followed by auditory stimuli, this time 250, 450, or 650 milliseconds after the action. Subjects had to estimate the occurrence of the tones both in trial blocks in which the interval remained constant (onset at 250 milliseconds only, for example), and three blocks of randomized combinations of the intervals. The results suggest that there is indeed a correlation between the time of onset of the effect of an action and intentional binding, for intentional binding was stronger (action and effect were perceived closer in time) for the fixed blocks than for the randomized blocks, and also stronger for shorter intervals compared to longer intervals (Haggard *et al.*, 2002, pp. 383-384). Thus, the authors concluded that both temporal contiguity and predictability are determining factors for the biding effect (Haggard et. Al, 2002; Moore and Obhi, 2012).

2.2 Prediction vs. retrospective inference

Since its first description the intentional binding effect has been independently confirmed and further investigated in several experiments (for a review, see Moore and Obhi, 2012). The very nature of the phenomenon indicates that a strong retrospective component is at play, for the acquisition of sensory data concerning the outcome of the action seems to be necessary to drive its perception. Nevertheless, a number of studies suggest that while retrospective inference indeed does play an important role, a predictive component that takes place during action selection and execution also contributes to intentional binding.

For example, Engbert *et al.* (2008) describe the necessity for intentional binding to be triggered by motor commands. Similarly to the seminal study, Engbert *et al.* (2008) instructed subjects to estimate the intervals between an initial triggering event and a later auditory stimulus in four conditions that varied according to the event that triggered the tone: intentional action, when one of the fingers of the subjects was pressed down by a rubber hand moved by a mechanised lever, the rubber hand pressing the button, and an experimenter pressing the button. As a result, it was found that only the condition that involved intentional actions produced a "robust" intentional binding effect. Thus, the authors concluded that " an efferent motor command is necessary for a sense of agency, and that proprioceptive and visual information compatible with the action are not sufficient" (Engbert *et al.*, 2008, p. 701). Significantly, this study shows that mere retrospective inference from action and its effect does not suffice to cause the phenomenology associated with intentional action.

In a later study, Moore and Haggard (2008) assessed the contribution of the retrospective component versus a possible prior-to-action component in the experience of agency. Subjects were asked to estimate the interval between their pressing of a button and a tone. The experimenter varied the probability of the tone actually happening after the action, with trial blocks in which the tone had a probability of 75% or 50% of occurring after each press of the button. The authors report that intentional binding took place in both conditions, although it was stronger in the trials with higher probability. This suggests, according to Moore and Haggard (2008), that the increased probability of the outcome creates a representation with a stronger action-outcome link, which is then used to form predictions about subsequent actions that in turn modulate the perception of time (pp. 142-143). Moreover, this idea is reinforced by the analysis of the experimental data, which showed that a shift forward in the perceived time of action was present in 'action-only' trials (where subjects were asked to estimate only the time of their actions, in absence of a subsequent tone), but only if those trials had been preceded by an action-tone trial.

The research concerning intentional binding allows for some conclusions, although it does not seem to entirely capture the experience of agency. First, it demonstrates that there *is* a phenomenology exclusively associated to intentional action, in the form of shifts in the perceived time of occurrence of both actions and their outcomes, that do not take place when an action is unintended (e.g. TMS-induced action) nor by mere observation of an action and its effects. Hence, agency is experienced and perceived differently than other forms of action and perception.

Furthermore, the studies show that this phenomenology is modulated by both post-action retrospective inferences and pre-action predictive processes. Yet, they are unclear as to how the latter may yield a sense of agency that is different from just the intentional binding effect. Is the idea that the formation of predictive representations of action suffices to yield a sense of agency and to the self-attribution of control and authorship? The findings surrounding intentional binding do not warrant such a conclusion. For example, as Engbert *et al.* (2008) point out, "the sense of agency might arise when sensory information corresponds to the prediction made from motor commands; if what actually happened corresponds to my motor command, then 'I did that'" (Engbert *et al.*, 2008, p. 702). This would mean that, while the phenomenology of perceived time associated to intentional action is partly determined by a pre-action predictive component, the sense of agency itself would still be the direct result of retrospective inference. The analysis of such models of the sense of agency (known as comparator models) is the subject of the following Section.

3. The Predictive Processing framework

The main feature of Predictive Processing (PP) accounts is a radical flip of the more traditional conception of the brain as passively receiving incoming sensory stimuli and forming a model of the external world "by a kind of step-wise build-up moving from the simplest features to the more complex" (Clark, 2015, p. 5). Instead, PP views the brain as a "multi-layered prediction machine", whose function is to build models that best anticipate the information acquired via the senses, where "top-level predictions concern matters that are increasingly discrete and abstract […]. Lower level predictions track states whose spatial or temporal signatures are continuous, local, and more fine-grained" (Clark, 2015, p. 7). A percept is, in this view, the 'winning' hypothesis that matches the incoming stimuli, the brain's best estimation of what is 'out there'.

One important consequence is the central role prior knowledge has in shaping perception. We perceive the world as we expect it to be, however, this does not mean that our model of the external world is not caused by actual objective patterns, since prior knowledge has itself been shaped by our contact with the outside world. A good example is the hollow mask illusion: when observing a rotating hollow mask, its convex side is perceived normally, yet, as it rotates further and the concave side is about to face us, the image "switches" and in lieu of a concave-shaped face, we perceive it as convex, because the brain settles on this hypothesis based on the strong, prior knowledge that faces have convex structures (Clark, 2018, pp. 7-8).

PP features both top-down and bottom-up processes for perception: as each layer's function is to estimate the state of the layer below, there is a continuous top-down flow of predictions informed by prior knowledge; yet, when predictions fail to match the incoming data, prediction error signals are sent up the hierarchy. The system's response then may vary according to the 'weight' assigned to the error signals, from altering the models to eliminate the errors to altering the incoming data so that it matches the existent model. This represents one of the greatest upshots of PP, because it articulates context, prior knowledge, perception and action in , and sheds light on the common cognitive processes that realise perception and action (Clark, 2018, pp. 7-9).

3.1 The Comparator model

The comparator model of the experience of agency can be roughly understood as a more modern version of Wegner's matching model, in which a representation that anticipates the action and its outcome is retrospectively matched with the actual outcome and thus the experience is generated.

Analogously, in the comparator model outgoing motor commands are integrated into a predictive model of the action and its expected outcome, which consists of the sensory feedback -from both the body and the environment- when the action is carried out. Once the action is performed, the received sensory feedback is compared to the predictive model. When the sensory data and the model match no prediction errors ensue, and action and outcome are attributed to the person, producing an experience of agency. However, if they do not match the experience of agency is diminished or not generated at all, prompting a search for causes other than one's action or a revision of the predictive models (Haggard, 2017, pp. 201-202; Haggard and Chambon, 2012, p. 390).

While, following Haggard (2017), the comparator models may "successfully explain the phenomena of 'non-agency'" (p. 202), such as the feeling that something has gone wrong when an action does not produce the intended result, it is not very effective when trying to account for the more 'positive' aspects of the experience of agency. According to Haggard (2017), and in contrast with the account of perception propounded by Clark (2015), in comparator models "the content of perception is given by the difference between actual and predicted feedback", so that "we only perceive what we cannot predict" (p. 202). This view effectively bars the possibility of a positive experience of agency, since it would correspond to a match between sensory feedback and predictive models that would not generate the prediction errors necessary to give content to an experience of agency.

Yet, we do seem to have positive experiences of agency, a "'buzz' of agency", as Haggard (2017) calls it, independently of the mechanism that brings the about (predictive or retrospective). Nonetheless, we do not expect every action to elicit an experience or sense of agency. When we perform a routine activity, like walking, for instance, we do not necessarily feel that the movement of our limbs is the product of a choice, a command, or a representation that prefigures our goal. This does not mean the brain does not produce such a representation, but that we are not conscious of them. At most, what we experience is being the authors of our actions and that we are in control, and that looks questionable as well. When walking, we do not experience ourselves as controlling the action of walking, we merely do it, and the knowledge of control and authorship underlies our activity. In line with what the comparator model suggests, we experience no "non-agency". In general, save for some pathologies, there is no need to self-attribute authorship. It is only when errors are detected , either in the action itself or its expected outcome, that a question about "who did that?" makes sense. And this question is usually preceded by a revision or adjustment of the actions to fit the goals.

So, it seems plausible that, instead of constant positive experiences, control and authorship of its own action are a default assumption for the brain, as Haggard and Chambon (2012) suggest. In this manner, the absence of prediction error described by the comparator model

may have a place in the experience of agency -other than enabling a post action inference about agency-, that better matches the sense of authorship. The sense of authorship could be thus negatively defined as the default assumption under which the brain carries out action, and which is only when evidence that indicates otherwise is met. Hence, the sense of authorship is not normally felt, because authorship is presupposed in every action, and it may emerge only when the question of "who did that?" is posed.

4. The sense of agency as prediction of action

4.1 Dennett on qualia

In his 2015 paper *Why and How Does consciousness seem the way it seems?*, Daniel Dennett presents a theory concerning how subjective features of experience (also known as qualia) may be realised by the brain's predictive functions. The issue of for qualia in terms of physical (neural) processes has been famously called by Chalmers (manuscript) the "Hard Problem" of consciousness, for they are conceived in philosophical tradition as irreducible and entirely subjective. In this context, Dennett's account may shed light on how the sense of agency (another feature of subjective experience) may be explained in terms of neural processes. Dennett's argument (2015) reads as follows:

1. There is no double transduction in the brain.

2. Therefore there is no second medium, the medium of consciousness or, as I like to call this imaginary phenomenon, the *ME*dium.

3. Therefore, qualia, conceived of as states of this imaginary medium, do not exist.

4. But it seems to us that they do.

5. It seems that qualia are the source or cause of our judgments about phenomenal properties.... but this is backwards. If they existed, they would have to be the *effects* of those judgments.

6. The seeming alluded to in (4) is to be explained in terms of Bayesian expectations."

(In Clark, 2018, p.6)

So, Dennett (2015) starts by examining what he calls the problem of "Double Transduction". According to him, if consciousness were to exist as a medium different that the physical, the sensory data that is collected from the environment would have to suffer a first transduction into "spike trains" of neural activity so that it is processed by the brain, and be transduced once more for the benefit of consciousness (and yet again, a third time, in order to communicate the judgements of consciousness back to the brain). However, Dennett (2015) argues, "biology has been thrifty on us" (p.2) in that, through the trial and error process of evolution, it has made Double Transduction entirely unnecessary by arranging us in such a way that spike trains are sufficient to generate all

aspects of behaviour (premise 1). Thus (2), the medium of consciousness lacks any function and is, in consequence, unnecessary as well.

The inference to (3) is relatively simple: if consciousness as a second medium does not exist then its states or objects -that is qualia- cannot exists either. Nevertheless, (4) it may still be questioned "Why does it seem to us that qualia exist? How is it that we make judgements about cuteness, whiteness, etc., if the objects to which they are supposed to refer do not exist?" Dennett's Humean response performs what Clark (2018) calls a "720 Triple Flip" on qualia: (5) the existence of qualia in the world is not the cause of our judgements about them, rather qualia are "projected" out to the world by us as a result of judgements about our experience; therefore, qualia are not the causes, but the effects of our judgements.

Thus, Dennett (2015) sets out to give an account of qualia in terms of Predictive Processing. Qualia are, he argues, second order predictions. So Dennett (2015) posits the existence of a second layer of prediction: in addition to the brain's predictions about "what is out there", it also produces predictions about our dispositions towards what we might expect to find in the world. Subjective properties are, thus, Dennett argues, the expectation of sets of dispositions towards objects we predict to encounter. Dennett (2015) gives an illustrative example: we are not moved to "cuddle or protect, nurture, kiss, coo over" (p.5) a baby in virtue of a perception of its cuteness, instead it is the expectation of these dispositions when we expect to encounter an infant that is interpreted as cuteness. Moreover "(w)hen… it to have" (p. 5), which gives rise to the belief that subjective properties actually exist in the world.

In addition says Dennett, analogous accounts may be given for the rest of subjective properties; that is, accounts about how neural processes produce conscious experience.

4.2 A predictive account of the sense of agency

There are several lessons to learn from Dennett's account of qualia that are pertinent to the matter at hand. The first is that from the dispelling of consciousness as a medium it

does not follow that conscious experiences are illusory. What it does imply is that the theoretical constructs built around the *MEdium* in order to explain in order to explain experience are false. Analogously, the illusory nature of conscious will does not necessarily render the sense of agency invalid, but invites us to a find a new explanation in more correct terms.

Second, all features of subjective experience can and should be described in terms of "spike trains". Third, and most important, qualia are the effects of judgments, but these judgements are made as predictions, not retrospectively. In contrast, a Wegnerian account of quails would probably identify them as the effect of an inference about the constant concurrence of a felt disposition and an external object. This does not mean that retrospective inference does not play an overall important role, just that it is not immediately responsible for bringing about experience. Its role is that of confirmation and reinforcement of judgements that have been made as predictions, as the Comparator model has it (Section 3). But perhaps this is a disservice to Wegner, after all his work shed light on the retrospective aspects of the experience of agency and helped in dispelling the notion of conscious will. Moreover, the theory that allowed Dennett's account wasn't yet available at Wegner's time. Thus, what follows is an attempt to apply that theory to the sense of agency.

Dennett's story relies heavily on the prediction of possible dispositions for action towards expected external objects; that is, the prediction of possible ways to interact with the objects based on the affordances that the objects offers. So the brain does not only predict possible states of the world, but its own possible, future action as well. But what role do these predictions have in bringing about the sense of agency?

Here an analogy with Dennett's third lesson is crucial. Like qualia, the sense of agency must be the effect of a judgement, and just as subjective properties do not result from a retrospective judgement concerning dispositions and external objects, the sense of agency may not be a retrospective inference involving the experience of a thought and an observed action. The important part here is that all elements of the judgement need to be internally available to make a prediction. Yet, the elements of the judgement that causes the sense of agency are not necessarily the same. It would be strange (and possibly dysfunctional) that the feelings of action emerge in virtue of the sole predictions of action. In the same manner that qualia are the effects of a judgement about dispositions towards objects, it seems reasonable to suppose that the sense of agency is the result of a

judgement about action. I believe the most likely candidate, as a judgement about action, is action selection. Thus, just as qualia are not the cause of judgements about them, but the effects of judgements about dispositions for action, the sense of agency does not cause our brain to select an action, but is the result of a predicted action having been settled upon.

Yet, one problem still remains. As we defined it, the sense of agency is the "feeling" that our conscious thoughts cause our actions. But are predicted dispositions for action conscious states? Dennett (2015) does not say much about how experiential features may become conscious and an attempt to unveil the mechanisms that produce experience vastly exceeds the limits of this essay. Suffice it to say that in Predictive Processing accounts (Clark, 2015; and section 3) perception is achieved when the brain settles on a prediction. In terms of action, a prediction may become consciously perceived when it has been selected for execution (but not by execution itself). Nevertheless, we must remember that the objective of perception is not to be "presented" to consciousness (Dennett and Kinsbourne, 1992), but that its form, content, level of abstraction, etc. depend on rules determined cognitive requirements, and not those of consciousness. This may be the reason why sometimes we experience mere "glimpses" of the actions we are about to perform. Another possibility is that the clarity and amount of content of a perception of action varies according to whether it comes from a higher- or lower-level prediction. Thus, the execution of a plan may cause a stronger sense of agency than routine activities such as walking.

In this manner, it can be said that the sense of agency is the effect of the selection for execution of a consciously perceived prediction of action. Paraphrasing Dennett (2015), what it is to experience a sense of agency is to generate the series of expectations about action and their confirmation by action selection. When an action selected for execution is not predicted (such as reactions) or when prediction of action is perceived too weakly or not at all (e.g. routine action), then a sense of agency is not elicited. However, as discussed in Section 2, such actions are still felt as controlled and caused by oneself as not unintended. Once the action has been produced, the prediction about its outcome is either corrected or confirmed, and information about the link between action and outcome is integrated into the brain's statistical knowledge of the world.

5. Conclusions

I have so far argued that the sense of agency is neither illusory nor the product of retrospective inference. I believe I have at least demonstrated the latter, and provided evidence that supports the truth of the former. While retrospective inference does indeed seem to be an indispensable element in the processes that produce an experience of agency, I must be clear now that it is not way sufficient to account for our experiences of the sense of agency.

The positive proposal sketched in Section 4 should be considered that, a first approximation towards an explanation of the sense of agency in terms of second order predictions. It is in no way definitive, but I believe it is plausible and may be the seed of future research. The Predictive Processing framework and Dennett's work on qualia are both very fertile and they will still produce more information that may help in developing a complete description of the sense of agency in term of neural events.

Bibliography

- Bayne, T. (2006) Phenomenology and the Feeling of Doing: Wegner on the Conscious Will. In Pockett, S., Banks, W. P. and Gallagher, S. (Eds.) *Does Consciousness Cause Behaviour*. Cambridge, Massachusetts: The MIT Press.
- Chalmers, D. (manuscript) The Meta-Problem of Consciousness. PhilArchive copy v2: https://philarchive.org/archive/CHATMO-32v2. Retrieved April 10, 2018.
- Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*. *36(3)*, 181-253.
    - (2015) Radical Predictive Processing. *The Southern Journal of Philosophy*. *53*, 3-27.
    - (2018) Strange Inversions: Prediction and the Explanation of Conscious Experience. In Huebner, B. (Ed.) *The Philosophy of Daniel Dennett*. Oxford: Oxford University Press.
- Dennett, D. C. (2013a) Expecting ourselves to expect: the Bayesian brain as a projector. *Behavioral and Brain Sciences. 36(3)*, 209-210.
    - (2013b) *Intuition pumps and other tools for thinking*. New York: W. W. Norton & Company.
    - (2015) Why and How Does Consciousness Seem the Way it Seems? In Metzinger, T. and Windt, J.M. (Eds.) *Open MIND. 10(T)*. Frankfurt am Main: MIND Group. doi: 10.15502/9783958570245
- Dennett, D. C. and Kinsbourne, M. (1992) Time and the observer: the when and where of consciousness in the brain. *Behavioral and Brain Sciences*. *15*, 183-247.
- Engbert, K.; Wohlschläger, A.; and Haggard, P. (2008) Who is causing what? The sense of agency is relational and efferent triggered. *Cognition. 107(2)*, 693-704.
- Haggard, P. (2017) Sense of agency in the human brain. *Nature Reviews Neuroscience*. *18*, 196-207.
- Haggard, P. and Chambon, V. (2012) Sense of agency. *Current Biology. 22(10)*, 390-392.
- Haggard, P.; Clark, S.; and Kalogeras, J. (2012) Voluntary action and conscious awareness. *Nature Neuroscience. 5(4)*, 382-385.
- Haggard, P. and Libet, B. (2001) Conscious intention and brain activity. *Journal of Consciousness Studies*. *8(11)*, 47-53.

- Haggard, P. and Tsakiris, M. (2009) The Experience of Agency: Feelings, Judgements, and Responsibility. *Current Directions in Psychological Science. 14(4)*, 242-246.
- Jensen, M.; Di Costa, S.; and Haggard, P. (2015) Intentional binding: a measure of agency. In Overgaard, M. (Ed.) *Behavioral Methods in Consciousness Research*. Oxford: Oxford University Press.
- Libet, B. (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*. *8*, 529-566.

- Moore, J. W. and Haggard, P. (2008) Awareness of action: Inference and prediction. *Consciousness and Cognition. 17*, 136-144.

- Moore, J. W.; Lagnado, D.; Deal, D. C.; and Haggard, P. (2009) Feelings of control: Contingency determines experience of action. *Cognition. 110(2)*, 279-283.

- Moore, J. W. and Obhi, S. S. (2012) Intentional Binding and the sense of agency: A review. *Consciousness and Cognition. 21(1)*, 546-561.
- Wegner, D. and Wheatley, T. (1999) Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist*. *54(7)*, 480-492.