12

A Framework for Understanding Agency

Kayuet Liu and Hakwan Lau

Abstract

Why do some thoughts feel involuntary and intrusive? When should we hold someone responsible for their actions and thoughts when they all have some basis in the brain? Are we truly free agents when we are bounded by shared values and culture? This chapter presents a framework for how our consciousness of our own intentions and emotions allows us to form causal narratives about ourselves and the world. These narratives determine our sense of agency, and we ascribe responsibility correctly depending on the extent to which one is capable of forming culturally appropriate narratives. Different ways of characterizing consciousness are analyzed, with a focus on one that may prove most useful within the context of understanding individual agency. A variant of the higher-order view of consciousness is advocated that allows us to form causal, albeit imperfect, narratives about ourselves. However, it is because of these imperfect narratives that our understanding of agency and responsibility is formed. Thus, understanding how these narratives come about is an important first step to understanding agency and how some thoughts are considered involuntary and intrusive. Implications of this framework are discussed using examples from mental illnesses, addiction, suicide, and racism.

Do Our Brains Make Us Do It?

Aberrant acts committed by patients with severe mental illnesses (e.g., schizophrenia) are often considered not punishable. By law, if patients have lost their capacity to reason, society should not hold them criminally responsible (Mobbs et al. 2007). However, there seems to be a spectrum of controllability (Moscarello and Hartley 2017) within which we ascribe degrees of responsibility to patients suffering from different mental disorders. Take addiction as a contrasting example. Law aside, members of society often disagree on whether addicts are responsible for their own actions. Some hold that it is the addicts' own "decision" to go down the path of becoming who they are. Some have challenged the notion that addiction is a brain disease, because the neural correlates of addiction are not sufficient to cause addicts to do what they do in

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus.
Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

many environments (Levy 2013). The question, then, is: How "voluntary" are the behaviors of addicts (e.g., seeking out a drug) compared to those of patients with schizophrenia (e.g., talking to an imaginary friend in public), since both have bases in the brain? Indeed, when a healthy student is late for class due to procrastination rather than a traffic jam, this behavior is also based in the brain. In a sense, people's brains are the causes of their behaviors in all these instances. Yet we ascribe agency and responsibility differently in each case.

Perhaps this relates to how we see an individual's agency and responsibility in the context of culture. If someone is brought up in a society in which it is acceptable to take food off each other's plates without first asking, much as we may disapprove of such actions, we may accept that person's behavior more easily than had that person been brought up in a culture where such behavior was sternly forbidden. In this sense, are we truly free individuals, acting voluntarily out of our own desires or judgments? Or are we bounded very much by our shared values, so that our errors may reflect more on the failure of society rather than ourselves?

In this chapter, we do not attempt to solve these difficult questions, but rather hope to provide a useful framework within which they can be addressed. We will argue that the notion of consciousness is crucial to understanding these issues. However, there are many different ways to characterize consciousness. We evaluate a few accounts and focus on one that may prove useful within the context of understanding individual agency.

Classical Literature on Free Will

Traditionally, debates on free will concern whether our actions are predetermined; that is, whether our actions are genuinely *de novo*. The assumption is that, in principle, if we knew all current physical events of the world, together with a complete understanding of all physical laws, we should be able to predict the next events perfectly (Laplace 1951; Hoefer 2016). Within this context of determinism (Hoefer 2016), how can one be an "unmoved first mover" (Strawson 1994; Pereboom 2001)? Are our actions not already fully determined physically, before they actually take place? To the extent that some notion of freedom is possible within the deterministic framework (i.e., compatibilism; Fischer 2006; Nahmias 2016), it cannot be because these actions are random and thus unpredictable. Rather, we are responsible for our actions because we have some degree of autonomous control over our actions. To argue that our actions are genuinely free in this sense, one option is to argue that the physical world is not truly fully deterministic. Some have appealed to findings from quantum physics: that certain future events are not fully determined, even if all current physical events are known (Kane 1996). This debate is important and interesting, but many good reviews of the literature already exist (e.g., Sinnott-Armstrong 2008). Here we focus specifically on considering which cognitive

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus.
Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

architectures may allow some meaningful notion of control to happen, while assuming that classical Newtonian physics is sufficient for understanding cognitive systems such as ourselves. For an in-depth discussion of these issues, with regard to the implications of quantum physics, see Tse (2013).

The Consciousness Requirement

Shifting from the traditional focus on whether our actions are predetermined, it has been argued that for individuals to be held responsible for their actions, they need to be conscious of certain relevant events (Caruso 2012; Levy 2014). Intuitively, this makes sense: it seems unfair to hold one fully accountable for something that one isn't even aware of having done, nor having even remotely contemplated doing. Incidentally, there is experimental evidence that this "consciousness requirement" is in line with our folk psychological concepts of free will and responsibility (Shepherd 2017). Accordingly, many have focused on the question: To what extent do *conscious* mental states truly cause behavior (Pockett et al. 2006)? This diverges from the traditional question of determinism, because this new question is still meaningful even if consciousness itself were fully deterministic (Nahmias 2014). So long as the (deterministic) conscious processes in the brain causally influence our behaviors, there may still be an important sense in which we have control over our actions.

There are, of course, critics of the view that consciousness is relevant. For example, Smith (2005) claims that forgetting a close friend's birthday (i.e., something that one does not consciously choose to do) does not eliminate the responsibility of failing to call or send a card. We will not go into the details here (for further discussion, see Caruso 2012), but an important lesson to derive from these exchanges is that the arguments often depend on which specific notion of consciousness is at play. We will focus on this issue in the next few sections.

The Surprising Power of the Unconscious

The question of whether the conscious processes in the brain are causal to behaviors may seem trivial (Baumeister et al. 2011). Although we all feel that our conscious thoughts and decisions have causal efficacy, several lines of empirical studies seem to challenge this intuition. First, studies of unconscious priming, mostly coming from the area of social cognition, show that our actions and decisions may be influenced by unconscious cues (e.g., words or symbols irrelevant to the primary tasks at hand, or the gender or age of the experimenter), the meaning of which we are not fully aware (Bargh et al. 2012). If true, these findings may show that our actions are not as fully consciously determined as we thought. Some have called into question, however, whether these findings can be replicated (Chabris et al. 2019), and one could argue that the relevant cues are merely unattended but not truly subliminal.

Another line of studies focuses on the well-known Libet clock paradigm (Libet et al. 1983), in which subjective estimates of action onset and conscious intentions (i.e., the time span during which individuals feel that they are about to make an action) are reported. The actions concerned are typically simple motor movements, such as spontaneous flexing of the wrist or pressing of a button. Preceding these simple movements, it has been reported that the brain activity arises well before the onset of conscious intention (Deecke et al. 1969; Libet et al. 1983; Lau et al. 2006; Soon et al. 2008). These findings have stimulated many debates. Simple computational models have also been proposed, suggesting that the findings are exactly what we should expect since neuronal processing is noisy (Nikolov et al. 2010; Schurger et al. 2012).

The conclusion from these studies seems to be that our conscious intentions are preceded by unconscious brain activity. Some have taken this to mean that our conscious intentions are "determined" by the preceding unconscious activity, yet this interpretation is unwarranted. In most cases, what was shown is simply a weak statistical correspondence between the unconscious activity and the subsequent intention. In any case, whether conscious intentions are determined is beyond the scope of our current interests.

The Libet studies also led to another finding: the time between our conscious intention and actual action execution may be too small for meaningful causation to take place (Lau et al. 2006). Further, based on early neuroimaging studies (Lau et al. 2004), the corresponding putative "intention" areas of the brain can be targeted with magnetic stimulation (Lau et al. 2007): this showed that the reported onset of conscious intention can be influenced by stimulation even *after* the action is completed. Subsequently, other studies have also used psychophysical methods to produce similar results (Banks and Isham 2009).

Just because intentions may be subsequently revised, however, does not rule out the possibility that pre-revised versions of intention may occur prior to action. Ultimately, the actions involved in the Libet studies are simple and inconsequential, so that even if conscious intentions do not cause them immediately, this does not rule out that intentions may be important for subsequent and more complex behavior.

Importantly, Wegner and colleagues conducted studies outside of the context of the simple Libet paradigm, concluding that our conscious intentions may be generally illusory; that is, not causal to immediate actions (Wegner and Wheatley 1999; Wegner 2002). Likewise, it has been shown that unconscious information can influence more complex tasks, such as preparing to answer one type of question over another (Lau and Passingham 2007; Rahnev et al. 2012) or response inhibition (van Gaal and Lamme 2012).

In other studies, it was shown that unconscious information in the brain can facilitate different forms of associative learning (Taschereau-Dumouchel et al. 2018b), which in some cases revealed therapeutic potential. For instance, using a technique called multivoxel neuro-reinforcement (Watanabe et al. 2017), one can pair unconsciously the representations of a spider with monetary reward so

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus.
Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

that the subject will subsequently show reduced physiological threat-related responses to images of spiders (Taschereau-Dumouchel et al. 2018a). In other studies, using a similar experimental setup, powerful forms of reward-based learning have been shown to take place unconsciously (Cortese et al. 2019). In addition, subjective confidence regarding one's ability to discriminate certain stimuli can also be changed with this unconscious association method (Cortese et al. 2016).

In summary, unconscious cognitive processing seems very powerful indeed, especially regarding its ability to form statistical associations and to influence subsequent behavior (via priming). Does this mean that consciousness plays no causal role and has no function? The answer is not so straightforward. Lau (2009) argues that although unconscious processes are powerful, this does not mean that there is necessarily no room for consciousness to add further benefits. To discuss meaningfully the theoretical possibilities of the role for consciousness, we need to distinguish between different notions of consciousness.

Pure "Qualia" versus Sheer Cognitive Power

In the studies mentioned above, *unconsciousness* typically refers to stimuli or processes of which the subjects are unaware; that is, subjects do not know that they take place. So, in a sense, *consciousness* is just the opposite: subjects are aware of the relevant events. There is, however, a tradition in philosophy that analyzes consciousness as concerning pure qualitative experiences (Nagel 1974; Levine 1983; Block 1995). In some cases, philosophers invite us to consider that molecule-by-molecule functional duplicates of ourselves may lack such qualitative experiences altogether (Chalmers 1996). While such conceptual possibilities are intriguing, this notion of consciousness is not particularly interesting for our current purposes (Levy 2014). If we define consciousness as having no functional consequence, of course, it could play zero causal role.

On the other end of the spectrum, one could characterize consciousness as the capacity to access information and use it for purposeful behaviors. On one view, this form of consciousness, called access consciousness, is always supported by strong, stabilized signals broadcast throughout the different systems in the brain (Dehaene 2014; Dehaene et al. 2017). With this notion of consciousness, it is highly likely that consciousness will be causally relevant for important decisions in everyday life (Levy 2014). Once again this seems to depend, somewhat circularly, on the definition of consciousness we choose to adopt. Of course, if consciousness is *by definition* characterized by global, complex, and elaborate processes, it is likely functionally important. But does this ignore how much we seem to be able to accomplish unconsciously?

A Middle Ground?

Because of the above considerations, a notion of consciousness that is relevant for our current analysis should ideally allow for some functional consequences in principle, without *assuming* so from the outset (Rosenthal 2012). Is it possible that there is a view that mental event is conscious in the phenomenological sense while whether it contributes functionally to our rational thinking and agency ascriptions remains an empirical matter? One such possible account is the *higher-order view of consciousness*, which can be traced back at least to John Locke and Immanuel Kant (Lau and Rosenthal 2011). According to this higher-order view, mental representations of events in the world are by themselves unconscious. We can call these first-order representations. They only become conscious when they are *meta*-represented by higher-order representations. That is, higher-order representations *about* the first-order representations are necessary for making the content of the latter conscious. On this view, we can see why powerful forms of unconscious processing are possible: the same first-order representations with the same functional capacities can be conscious or unconscious, depending on the presence of the relevant higher-order representations (upon which one forms beliefs and more complicated narratives). Yet why do we need higher-order representations for first-order representations to become conscious? Traditionally, the arguments come from philosophical analysis (Rosenthal 2004): if one is aware *of* being in a certain mental state, one must represent oneself as being in that state (via the higher-order representations). Below we will elaborate further what this means in terms of cognitive architecture. Criticisms of this well-known theory, and their replies, have been extensively reviewed (Rosenthal 2005; Brown et al. 2019b).

Concordance with Current Science

Just because a philosophical notion of consciousness exists and can serve our purpose does not mean that we should adopt it. Fortunately, there is considerable empirical support for the higher-order view of consciousness (Lau and Rosenthal 2011). Neuroimaging studies have shown that subjective awareness of visual stimuli is associated with brain activity in high-level cognitive regions (e.g., in the prefrontal cortex), even under highly controlled experimental conditions in which the subjects are not merely processing the stimuli in simple (first-order) tasks (Lau and Passingham 2006). Also, neurological patients with selective damage of their visual cortex may lose the relevant subjective visual experiences but not their ability to correctly "guess" the identity of the relevant stimuli (Weiskrantz 1997); when visual stimuli are presented to intact parts of their visual cortex (as compared to the "blind" regions) leading to a conscious experience, higher activity in the prefrontal cortex was also found (Persaud et al. 2011). Disruption of activity in this prefrontal brain region selectively impairs one's ability to introspect whether one has successfully perceived the stimuli, without impairing (first-order) perception itself or memory (Rounis et al. 2010; Fleming et al. 2014). These findings are somewhat difficult to explain if we assume that consciousness is just strong, global information processing.

While these findings are reviewed in detail elsewhere (Lau and Rosenthal 2011), of particular interest to neuroscience is our emerging ability to assess, on a person-by-person basis, the efficiency of the relevant higher-order mechanisms (Baird et al. 2013b; McCurdy et al. 2013; Vaccaro and Fleming 2018). As we will see below, analysis of such individual differences is likely the key to understanding breakdowns of agency and responsibility.

Another line of support of the higher-order view comes from modern studies of artificial intelligence (Lau 2019). It is generally accepted that for neural network models to perform well, they can benefit from having the capacity for "predictive coding," in which a model can generate exemplar images (e.g., of cats or monkeys) top down. This improves the model's ability to classify images (e.g., of cats or monkeys). Training such a generative network, however, can be time consuming. To accelerate this process, another network called the discriminator, whose job is to detect forgeries (Creswell et al. 2018), can be trained to discern whether an image is genuinely from the world or created by the generative network (forgery). When we pit these two networks against each other, so that the discriminator wins a point for correctly identifying a forgery and the generative network wins a point by getting away with it, these two networks grow together not unlike rivaling siblings: they both become highly efficient in a relatively short time.

In the context of the human brain, it has been suggested that similar mechanisms of predictive coding occur. Neurons in the visual cortex may fire when a cat is presented to the subject, but similar neural representations are also involved when we imagine or remember a cat. The brain must be able to tell what causes these same neural representations to be activated in each instance. Given that these top-down generative mechanisms seem to be so efficient, it is likely that through development they are aided by the existence of a discriminator-like mechanism. Such a mechanism can determine whether a visual representation is generated by oneself or triggered externally by actual objects of perception. Plausibly, this mechanism can also tell when the same visual neurons may just be firing because of spontaneous noise. This conceptualization fits well with the higher-order view, in the sense that the output of this putative discriminator is akin to the higher-order representations necessary for conscious perception to occur. Normal conscious perception happens when the discriminator decides that a certain visual representation is truthfully representing the external world right now (Lau 2019).

Formation of Rational Beliefs

In part, based on the findings concerning the surprising power of unconscious processing, it has been argued that higher-order representations may have little

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus.
Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

additional utility besides making the relevant first-order content conscious (Rosenthal 2012). Using the computational interpretation above, however, we can identify some plausible key functions. In particular, one such function may be the formation of rational beliefs. By rational, we merely mean that these beliefs are *subjectively justified* in the sense that upon introspection, it should seem to the *subject* that the beliefs are *reasonable*. In general, it seems reasonable to believe in what we consciously see. Even when we have other reasons to believe that our perceptual system may be at fault—such as during lucid dreaming (Baird et al. 2019), at a live magic show, or after having knowingly ingested hallucinogens—there is still a strong temptation for us to believe what we see. According to the above interpretation of the higher-order view, this is because when we consciously see a cat, we have a first-order representation of a cat as well as a higher-order representation which states that the relevant first-order representation is a truthful representation of the world right now. The derivation from there to the belief that "there is a cat" is akin to a matter of syllogistic inference, so naturally such a belief may seem rational from the outset.

What good is having these subjectively justified beliefs? We do not deny that some beliefs may be unconscious,[1] but as human agents, we form narratives about the world and ourselves, and such narratives matter for our actions (Dweck 1999). In doing so, we tend to try to be coherent (Holyoak and Powell 2016). When we have many beliefs (including background, unconscious beliefs), this coherence is often difficult to achieve. One would therefore do well to include in this rational thinking system only beliefs of which we are reasonably certain. According to this perspective, beliefs that enter our narratives are mostly subjectively justified. In the case of perception, such beliefs can be as simple as "there is a cat." However, we also form beliefs about ourselves, our actions, and emotions, to which we now turn.

Self-Narratives in Agency and Emotions

As in perception, we know that motor representations in the brain also serve multiple purposes. Simple motor commands in the brain (in the primary and secondary motor cortices) are activated when we act as well as when we imagine performing the same action or when we observe others performing the same action. So presumably, as in the case of perception, some discriminator-like monitoring mechanism needs to decide when a motor command reflects what *oneself is intending to perform* (rather than just what one is imagining, or noise). This allows one to form the corresponding belief that "I intend this to happen."

[1] What we are claiming here is that beliefs based on conscious experiences are subjectively justified. Beliefs can, however, be based on unconscious experiences, but they will not be subjectively justified.

Taking an analogy from computers, the ability to form such beliefs seems useful. Let us say that a computer server in a large network sends a command to print ten pages to the printer. Another node in the network complains that the printer queue is now jammed and asks the first server to resolve the problem. It would be quite useful for the first server to know who contributed to the printing in the first place.

This self-directed nature of the corresponding representations is also relevant in emotions. Again, we know that for certain basic emotions, imagining them activates similar neural representations as experiencing them (Reddan et al. 2018). Likewise, when we emphasize and think about others' emotions, similar representations are involved. Therefore, when these first-order emotional states occur, the brain needs to know what the causes are. According to the view advocated here, one consciously experiences emotions when the relevant higher-order state points out that the first-order emotional representation reflects what *one is going through*. Thus, the corresponding belief may be simply that "I am angry" or "I am scared."[2] Having such beliefs may be quite useful in navigating social situations and in explaining to others why we behave a certain way.

Contrasting these with the relatively simple beliefs in the case of perception ("*there* is a cat"; see Figure 12.1), there is a sense that agency and emotions are intrinsically more self-involving. In fact, as Ledoux argues, without some minimal concept of the self, an animal may not experience basic emotions (e.g., fear) at all (LeDoux and Sorrentino 2019).

Why does one need to form these rational beliefs about oneself, which requires that the relevant first-order states be made conscious via meta-representations? The proposal is that first-order processes are powerful: we can use them to learn about statistical associations between events. However, mere associations are not coherent *narratives*. Moreover, narratives are stories in which events are *causall*y related. When we say that Julius Caesar invaded a certain country *because* he was angry, we mean that his emotion *caused* certain behavior. When he decided to invade, presumably he saw himself as an agent who was causally *responsible* for the decision. Inferring about causality is, however, notoriously hard. It is technically a challenging problem from a statistical point of view (Pearl and Mackenzie 2018). Without controlled experiments in which we can manipulate the putative causes while holding all other things constant, assumptions need to be made and heuristics involving counterfactuals may need to be invoked (Bond et al. 2012; Chambon et al. 2018). As such, interpretations matter; there may be more than one way to tell a story based on the same facts. With these imperfect narratives, we form a quasi-rational understanding of why we behave a certain way, and we provide socially acceptable justifications of our actions, based on folk psychology.

[2] Forming such beliefs does not necessarily involve language ability.

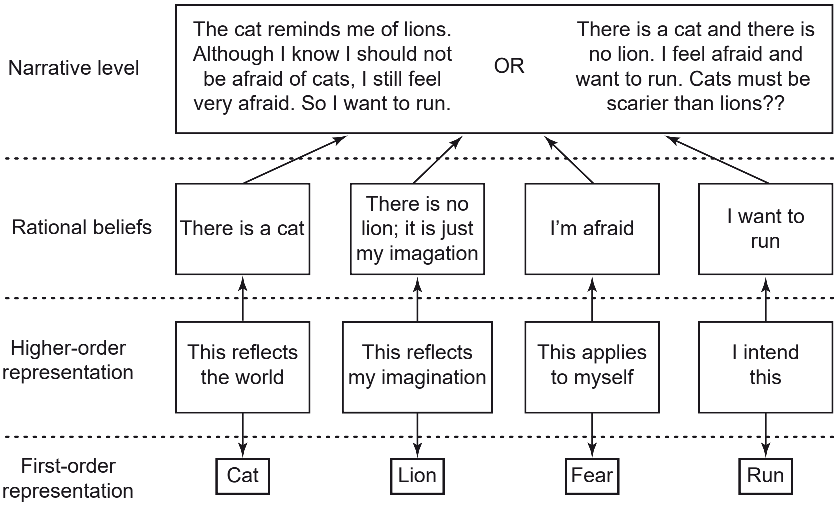
Narrative level
The cat reminds me of lions. Although I know I should not be afraid of cats, I still feel very afraid. So I want to run.
OR
There is a cat and there is no lion. I feel afraid and want to run. Cats must be scarier than lions??
Rational beliefs
There is a cat
There is no lion; it is just my imagation
I'm afraid
I want to run
Higher-order representation
This reflects the world
This reflects my imagination
This applies to myself
I intend this
First-order representation
Cat
Lion
Fear
Run

Figure 12.1 From first-order representation to self-narrative. Higher-order representations relate mental states to oneself (e.g., this reflects my imagination) upon which rational beliefs are formed. Different narratives, however, can be formed depending on how one relates the beliefs.

Some Related Views

The process of beliefs and subsequent narrative formation that we propose here is quite similar to the accounts of the post hoc explanations made by Gazzaniga's left-brain interpreter (Wolman 2019). To clarify, the above view does not mean that consciousness is the same as self-narrative (or self-consciousness). The view holds that the necessary and sufficient conditions for conscious experiences involve having the corresponding first- and higher-order representations. It is only by having these representations that we can form rational beliefs, based on which we form these causal narratives, while trying to be as internally coherent as we can.

This view links consciousness with rationality. Therefore, it is related to other models of rational decision making. In the Two Systems framework, championed by Daniel Kahneman (2011), first-order representations may roughly correspond to processes in System 1 (fast), with subsequent processes belonging to System 2 (slow). In reinforcement learning, there is a well-known distinction between model-based and model-free learning and decision making (Dayan and Berridge 2014). Roughly, higher-order mechanisms may relate better to model-based processes, whereas the first-order mechanisms may map to model-free statistical associations.

Our goal is not to replace or compete with these views. They are independently, empirically well supported, but they may serve different purposes. For example, although there may be a sense that the Two System approach or model-based versus model-free distinction may map onto conscious versus unconscious processes, such mapping is not intended to be clear cut. The higher-order view, on the other hand, is a theory of consciousness per se with direct empirical evidence. In the present context of understanding agency, the consciousness requirement is an important component of the overall argument. Accordingly, it is not clear if one should be absolved of the relevant responsibility just because decisions are made with System 1 (fast) thinking or model-free learning.

Another distinction is that one may conceive of the Two Systems framework as representing two parallel processes. The first- versus higher-order model here, however, stipulates that the two mechanisms are in a hierarchy. This hierarchical nature may have important consequences for treatment of mental disorders; intervention at the first level will causally impact on the higher level (Taschereau-Dumouchel et al. 2018b).

Understanding Mental Illnesses

This self-narrative account of agency may help us understand why some believe that patients suffering from severe mental illnesses may be less deserving of punishment than the unpunctual student, even though in both cases brain activity causes the relevant behavior. In the case of a patient with schizophrenia, the very basic mechanism of higher-order perceptual reality monitoring may be at fault. Thus, the patient may be unable to distinguish self-generated inner speech from externally triggered voices (e.g., from "God"), occurring from a breakdown between the bottom two levels in Figure 12.1. The patient may not be able to tell if an action is voluntarily produced or controlled by aliens. As such, the entire self-narrative system may well disintegrate. It is probably not fair to hold such patients accountable for their own behavior, if they are not correctly aware of who and what events caused these actions in the first place.

In the case of the unpunctual student, the higher-order system is presumably intact. What might have caused the (mildly) delinquent behavior may be a momentary overemphasis on the value of not having to rush or an attraction to another activity. These values represented in the first-order system are no less brain based and, in a sense, they too cause the resultant behavior. With an intact higher-order system, however, one should be able to appreciate that these first-order values are problematic and that one would be guilty all the same for acting a certain way.

What about cases of addiction and substance abuse? Between severe mental illness and everyday cases of delinquency, there likely lies a spectrum. In some cases of addiction, one may suffer first-order malfunctioning such that a substance may be associated with an unrealistic expected level of immediate reward, even when one is well aware that it cannot be good in the long run. In some cases, this higher-order mechanism may well be relatively intact, so one may be accountable for not recognizing the situation as such. However, there may also be cases where such a higher-order system is compromised, due to chronic abuse, which is known to impair the brain circuitry responsible for high-level control (Baler and Volkow 2006). Still, even when the higher-order system is intact, there are cases where alternative actions are not perceived as feasible or attractive, as suggested by the Rat Park studies (Alexander et al. 1978; Solinas et al. 2008); in such cases, the best solution may well lie in social policy rather than neurobiology (Hart 2013). Like others (e.g., Levy 2013), to make the correct judgment we think that we need to be able to adjudicate between the different cases in terms of both the specifics of the brain impairment as well as the environment. What we want to emphasize is that the distinctions between higher-order representations, beliefs, and self-narratives are crucial to the notion of agency.

A relevant intermediate condition to consider may be obsessive-compulsive disorders. Here, intrusive thoughts may primarily arise due to a malfunction at the lower levels; for instance, one may register the scene of an unclean bathroom as extremely threatening. Among these patients, some may genuinely believe that this is the case at the self-narrative level. Other patients, however, may recognize that the unclean bathroom is actually not that harmful, yet the visceral experience may be too much for them to overcome. In other words, patients may differ in terms of whether the conscious experience ultimately affects the healthy functioning of the entire higher-order system.

Because of these considerations, we call for more effort to subclassify the various disorders, including anxiety and depression. Is a certain patient suffering from malfunctioning at the first-order or higher-level, or both? Importantly, as LeDoux and Pine (2016) have argued, these different etiologies may need to be targeted independently. This is complicated by the fact that a disorder at one level may influence another level, as they are interconnected. To provide comprehensive treatment, one useful strategy may well be to target both the higher and lower levels (DeRubeis et al. 2008). Recognizing which level is the source of the problem for a particular individual will likely help finesse this process.

Thus, we argue that mental disorders are brain disorders and to understand agency and its breakdown we need to look carefully at the individual's condition, using the first- versus higher-order framework. Finding some correlates in the brain for certain misbehavior, however, should not absolve an individual of relevant responsibility. By understanding which brain correlates may be reflecting specific behavioral impairments, the framework advocated here provides a way to identify the theoretically relevant correlates at the different levels.

Understanding Responsibility of Individuals in Society

Mental illnesses can sometimes lead to one of the most devastating consequences: suicide. At times considered to be one of the most personal and existential decisions (Nietzsche 1955), suicide has also been analyzed as a social phenomenon (Durkheim 1951). In Emile Durkheim's classic work on the topic, he argued that many aspects of suicide depend on societal norms and values. If suicide were entirely a matter of personal decision or individual psychopathologies, it would be difficult to account for the stable cross-cultural differences in suicide rates (Durkheim 1951; Liu 2009). In what ways are we to understand suicide as being culturally and socially contextualized? Does the individual not make the decision after all?

The advocated view in this chapter can help us categorize the different ways in which social influences take place. At the first-order level, the statistical regularities of social events are picked up by the individual: How rarely does suicide occur? When the tragic event of suicide occurs, how do others react? As the individual becomes consciously aware of these events and their contingencies, the individual forms rational beliefs about suicide in social settings. Importantly, one also forms narratives about these events, in which one as an agent plays certain causal roles.

At the narrative level, interpretation matters. Different stories have different meanings. We interpret victims of suicide as causal agents. Why did so-and-so kill themselves? What drove them to such a desperate decision? Was it moral for them to do so? How does it causally affect their loved ones? These narratives apply to others as well as oneself and are naturally colored by our social understanding of the relevant concepts and implications. As such, societal norms and values influence suicidal behavior. However, we ultimately understand the decision is to be made by the individual concerned, within the given social context (Weber 1930).

Take racism as another example. One may perceive the presence of many youngsters of a certain ethnic group in a neighborhood as being statistically correlated with a higher occurrence of crime. Hypothetically, let us say that this association were statistically true in a certain context. Our unconscious first-order system may be truthfully picking up such an association, yet at the narrative level, where we ascribe causal relations to events, we do not necessarily have to interpret the ethnicity of the relevant youngsters as the *cause* of the prevalence of crime. More plausibly, both the ethnic makeup and occurrence of crimes in the neighborhood could be commonly caused by poverty and other forms of social injustice. Accordingly, the individual is still responsible for making the correct and appropriate interpretation and forming the correct narrative, given the same statistical facts picked up by the first-order system.

In this context, it is worth noting that in overcoming racism, much effort is focused on addressing our various implicit biases. What we have argued for here (excluding any possible statistical biases) is that there is also a level of narration to consider, where cause and effect is up for debate. Just because narratives are subject to interpretation does not mean they are too elusive to be worth studying. They can be changed and clarified through education and social discourse. At times, focusing on this higher level may well be more effective than methods focusing on changing our first-order implicit biases.

Closing Remarks on the Role of Higher-Order Mechanisms in Intrusive Thinking and Treatment

When considering therapeutic intervention in higher- or first-order frameworks, it may be useful to consider B. F. Skinner's perspective on whether the concept of human consciousness plays any meaningful role in science (Blanshard and Skinner 1967:325): "No major behaviorist has ever argued that science must limit itself to public events.…As a behaviorist, however, I question the nature of such events and their role in the prediction and *control* of behavior" (italicized emphasis added). Skinner famously advocated studies in psychology based on the Pavlovian tradition, in which we focus on how an individual learns about the statistical associations between observable events, rewards, and behavioral responses. In our terms, this concerns the unconscious first-order level. Skinner (1971) went so far as to speculate that to engineer a better society, we should focus precisely on these basic mechanisms. Indeed, in health and disease, methods of intervention based on these Pavlovian principles have often been proven effective. However, as we see from the quote above, even a stern behaviorist as Skinner did not completely rule out the possibility that consciousness could ever be studied. The problem is that at this higher narrative level, things are complex and often depend on social factors that affect the various interpretations. Unlike Skinner, however, we are not so pessimistic about the possibility that we can understand this system. We have laid out how higher-order level narratives may be causal and may also inform ourselves about our role as causal agents. In the context of intrusive thinking, this means that whether a thought is considered intrusive or voluntary may ultimately depend on these imperfect narratives. The narratives themselves may be complex, depending on perspectives, but at least we can try to understand the underlying general mechanisms.

To conclude, we have provided some limited evidence, but no proof, that this framework is correct. The issues at hand are of immense historical significance. They concern whether we can understand individuals as rational agents. Prior to World War I and II, many great scholars from disciplines as diverse as sociology, economics, political science, psychology, and neuroscience opined on this topic. However, over the last half century, discussions on consciousness and free will have shifted toward physics. We suspect this may be due to contingent sociohistorical factors, some unfortunate and not necessarily productive. The important question of freedom of the individual, in the context of health and disease, in isolation as well as within social networks, may not benefit as much from insights from physics as from careful analysis of the neurocognitive structure (outlined throughout this volume) that underlies our narratives about ourselves. That psychotherapy can be just as effective as Pavlovian-based methods in the clinic, at least in some cases, suggests that we should not write off the intriguing possibility that our higher-order mechanisms can also be systematically understood, and modified, as needed.

From “Intrusive Thinking: From Molecules to Free Will,” edited by Peter W. Kalivas and Martin P. Paulus.
Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1