Front. Philos. China 2018, 13(3): 377–393
DOI 10.3868/s030-007-018-0029-2

SPECIAL THEME

Paul Thagard

Mind, Consciousness, and Free Will

Abstract This commentary discusses how philosophy and science can collaborate to understand the human mind, considering dialogues involving three philosophers and three cognitive scientists. Their topics include the relation of philosophy and science, the nature of mind, the problem of consciousness, and the existence of free will. I argue that philosophy is more general and normative than science, but they are interdependent. Philosophy can build on the cognitive sciences to develop a theory of mind I call "multilevel materialism," which integrates molecular, neural, mental, and social mechanisms. Consciousness is increasingly being understood as resulting from neural mechanisms. Scientific advances make the traditional concept of free will implausible, but "freeish" will is consistent with new theories of decision making and action resulting from brain processes. Philosophers should work closely with scientists to address profound problems about knowledge, reality, and values.

Keywords cognitive science, consciousness, free will, intuition, materialism, mind, philosophy, science

How can philosophy and science collaborate to understand the human mind? In engaging dialogues, three philosophers and three cognitive scientists discuss crucial topics concerning the relation of philosophy and science, the nature of mind, the problem of consciousness, and the existence of free will. The philosophers are Terry Horgan, Owen Flanagan, and Derk Pereboom, interacting respectively with the psychologist Thalia Wheatley, the neuroscientist Giulio Tononi, and the cognitive neuroscientist Marcel Brass.

In this commentary, I try to clarify the issues discussed by the six discussants, and to sketch my own take on them based on ideas developed in detail in my forthcoming *Treatise on Mind and Society*, a trio of books to be published by Oxford University Press.

Paul Thagard (✉)
Department of Philosophy, University of Waterloo, Waterloo, ON N2L 3G1, Canada
E-mail: pthagard@uwaterloo.ca

1 What Is the Relation between Philosophy and Science?

Happily, none of the participants in the dialogues takes the extreme view that science is irrelevant to philosophy (because philosophy can answer all the important questions by itself), or the equally extreme view that philosophy is obsolete (because all the questions it asks are better addressed by science). Horgan says that philosophy is an empirical discipline, and Wheatley looks to philosophy for broad and important questions. Flanagan and Tononi both address problems about consciousness that combine philosophy and science. However, whereas Pereboom thinks that a solution to the ancient philosophical problem of free will should be at least compatible with scientific findings, Brass doubts that empirical science is relevant to debates about free will.

Philosophy is the attempt to answer general questions about knowledge, reality, and values. Science is the empirical and theoretical study of the natural world. These enterprises would be at odds with each other if philosophy needed to invoke supernatural entities and methods. But I agree with philosophers from Thales to Quine who have maintained that philosophy can be naturalistic, not requiring and indeed rejecting the invocation of ideas and techniques that transcend the world open to empirical investigation.

What then distinguishes philosophy from science? I think that the two main differences are that philosophy is more general and more normative than science (Thagard forthcoming-c). Whereas science investigates particular kinds of things such as stars, chemicals, and genes, philosophy asks questions about existence in general and methods for finding out what exists. Moreover, issues about knowledge and values are inherently normative, concerned with what ought to be, rather than just with what is. Science also addresses normative questions, for example in figuring out how to build better bridges and to improve education, but philosophical fields such as epistemology and ethics are much more intensely normative, constantly evaluating what constitutes knowledge and good conduct.

Philosophy and science need each other, in a kind of interdependence similar to what occurs in a good romantic relationship where members of a couple can be better and happier than they would be alone. Philosophy needs science because it cannot plausibly address general questions about knowledge and reality without a rich understanding of how minds interact with particular parts of the world. Even normative questions about values and obligations are connected with empirical questions about the mental and behavioral capacities of human beings.

Science does not need philosophy for routine investigations in well-trodden areas, but whenever it ventures into new territory, it runs into important questions in epistemology and ethics, concerning what knowledge is, how it grows, and the

moral implications of novel applications. Scientists who dismiss philosophy as superfluous in the area of empirical investigations are usually extremely naïve about their own methodology and the goals of science.

Accordingly, I advocate this *philosophical procedure:*

(1) Identify the most important philosophical issues as questions;
(2) Consider a range of available answers to these questions;
(3) Evaluate these answers based on coherence with scientific knowledge and other defensible philosophical doctrines;
(4) Reach philosophical conclusions by accepting some answers and rejecting others based on coherence with evidence and human goals.

This procedure requires philosophy to interact with science to ensure that its claims about knowledge, reality, and values fit well with findings in the natural and social sciences. But it also recognizes that philosophy goes beyond the investigations of particular sciences because of its need for greater generality and normativity.

This philosophical procedure is similar to scientific methods in attempting to come up with coherent answers to important questions, but differs because philosophy needs to range more broadly across relevant sciences. For example, physicists and biologists investigate particular kinds of objects such as planets and trees, but philosophy pursues challenging questions about the nature of objects in general, requiring it to be informed about planets *and* trees. Moreover, philosophy cannot avoid questions about what kinds of objects *ought* to exist, for example whether it is ethical to create robots that are autonomous and intelligent.

This procedure is incompatible with common philosophical practices. Horgan describes how philosophers are often interested in investigating the working of concepts, which they can do in the armchair by consulting their intuitions concerning thought experiments. He thinks that intuitions can serve as data for theorizing about how concepts work because strong intuitions are widely shared on the basis of conceptual competence. Then philosophical theories can be evaluated as inferences to the best explanation of these data.

This method is based on false assumptions. First, experimental philosophy has found substantial diversity in philosophical intuitions based on ethnicity, gender, personality, philosophical background, and age (Colaço et al. 2014). On the occasions when philosophers do agree in their intuitive reaction towards thought experiments, the agreement is more plausibly explained by common socialization than by shared conceptual competence.

An analytic philosopher might respond that philosophy deals with timeless concepts that should be amenable to the use of intuitions as data. But careful study of the history of philosophy reveals many changes that have occurred in important concepts such as knowledge, for example since ancient Greek ideas about *episteme* and *techne* (Parry 2014).

Second, even if there were widespread agreement among the intuitions of philosophers and ordinary people, there is no reason to believe that the concepts about which they agree are well-suited to describe how the world is and how it ought to be. As the development of science has often shown, it is frequently desirable to jettison old concepts in favors of ones that are parts of theories that provide better explanations of the evidence (Thagard 1992). The method of conceptual analysis is unduly conservative.

Third, the usual practice of conceptual analysis assumes an empirically outmoded theory of concepts as being definable by necessary and sufficient conditions. Hundreds of psychological experiments suggest alternative views of human concepts as being more tied to sets of examples, typical features, and explanations, not the strict definitions that are usually sought as part of conceptual analysis (Murphy 2002). A new theory of concepts shows how to combine all three of these aspects (exemplars, typical features, and explanations) through common neural mechanisms (Blouw et al. 2016).

This theory suggests that conceptual analysis can be performed, not by using intuitions to produce definitions, but by empirically identifying standard examples, typical features, and explanations used in a concept. I call this method "3-analysis," and apply it below to the concept of consciousness.

The point of this exercise is not to establish concepts as fixed manifestations of conceptual competence, but rather to characterize the current state of concepts with a view to considering whether they need to be revised or supplanted. In the philosophy of mind, many concepts are merely holdovers of prescientific views of mind, constituting prejudices and obstacles to future progress. The point of conceptual analysis in any dynamic field is to suggest ways of moving on to better systems of concepts that are part of stronger explanatory theories.

I agree with Wheatley that the proper use of thought experiments in both science and philosophy is to help generate hypotheses that can then be evaluated with respect to evidence (Thagard and Stewart 2014). The philosophical use of thought experiments to justify conclusions on the basis of intuitions should be abandoned. Horgan gives Putnam's famous Twin Earth question about whether XYZ is water as an example of a thought experiment that yields strong intuitions, but this example is scientifically incoherent (Thagard 2012). XYZ is supposed to be different from H_2O but indistinguishable from it, but chemists know that even a slight change in the chemical constitution of water produces dramatic effects. For example, if H_2O were replaced by heavy water with the isotope of hydrogen deuterium (D_2O), then life would not occur. Philosophy needs to be more seriously empirical than the method of intuitions about thought experiments allows.

Pereboom and Brass discuss the possibility of acquiring knowledge of transcendental agents with transcendental freedom of will. From my naturalistic

perspective, such longing for transcendence is wishful thinking left over from traditional ideas about the soul, immortality, and religious responsibility. My philosophical procedure still allows concern with crucial philosophical questions about the nature of reality, persons, and morality, but maintains that the resulting answers will be neither transcendental nor supernatural.

2 What Is Mind?

In their experimental work, scientists do not need to worry about metaphysical questions such as the reality of mind. For example, Wheatley's studies of emotional cognition, Tononi's investigations of sleep, and Brass' experiments on imitation can stand on their own as identifying important phenomena about the operations of human minds. From a theoretical perspective, however, general philosophical questions become important.

Most people in the world today are dualists, believing in mind separate from body in a way compatible with prevalent religious beliefs about life after death. With the rise of computational explanations of mind in the 1950s and 1960s, many philosophers and psychologists came to endorse functionalism, the view that what matters to mind is not the physical hardware that produces it but the computational functions that turn inputs into outputs. The dramatic rise of cognitive neuroscience since the 1980s and 1990s has provided new support for mind-brain identity, the view that all mental processes are brain processes.

Horgan talks about the possibility of multiple realizability, which is the idea that a mental state could be instantiated physically in many different ways, for example by a computer rather than by a brain, or by the same brain in different ways. Multiple realizability has long been used as an argument for functionalism, because it seems that if a mental state such as a belief can operate in a robot as well as in a person, then the mental state cannot be identified with any brain state. But conclusions about the nature of mind should not be based on abstract possibilities of the sort generated by thought experiments, but on the rich evidence generated by the relevant sciences. Metaphysics is not science but can build on it, with conclusions subject to revision as evidence accumulates and theories improve. A key question answerable only by advances in science and technology is how similar belief-like processes in robots would actually be to beliefs in humans.

My own take on the current state of cognitive science is that a view I call "multilevel materialism" is more plausible than dualism, functionalism, or the identity theory. The alleged evidence for dualism such as life beyond death is increasingly suspect, and dualism's major current support comes from the problem of giving a materialist explanation for consciousness. Functionalism

flourished when purely computational explanations of mind seemed promising, but cognitive neuroscience has revealed many ways in which hardware really does matter to intelligence, for example in producing energy-efficient, real-time solutions to problems of survival and reproduction.

The identity theory is on the right track because of dramatic progress in connecting mental processes with neural processes. But neural mechanisms are not the only ones important for understanding how minds work, because how people think is also a function of molecular mechanisms such as genes and neurotransmitters, representational mechanisms such as inferences and emotions, and social mechanisms such as how people communicate with each other. So, the best way to explain how minds work is not to focus only on what neurons do but also to consider how neural mechanisms interact with molecular, representational, and social mechanisms. Horgan alludes to the exclusion problem, the concern that causality can operate fundamentally at only one level, but that concern strikes me as more a prejudice than a principle.

The key scientific task is to collect more evidence and build theories that explain them, dealing with mechanisms operating at relevant levels, including theoretical integration across the levels. I suggest ways of doing this in my books *Brain-Mind* and *Mind-Society* (Thagard forthcoming-a; Thagard forthcoming-b), but undoubtedly more sophisticated ways will be identified as understanding of brain and social interactions grows.

The key philosophical task is to identify the metaphysical theory that fits well with these ongoing scientific developments. Multilevel materialism, which asserts that mind is matter and energy operating with multiple mechanisms, seems to me to be the best current way of reconciling scientific progress with the philosophical quest for generality.

3 What Is Consciousness?

The problem of consciousness is pivotal to philosophy of mind, and more generally to metaphysics, epistemology, and ethics. Contrary to behaviorism and other views that try to eliminate consciousness, people do have experiences when they think, perceive, and have emotions and pain. The major stumbling block to a naturalistic, materialistic view of mind is the need to provide scientific explanations of these experiences. Without such explanations, materialism would have to be rejected in favor of alternative metaphysics such as dualism (mind is separate from body), idealism (there is no matter, only mind), or panpsychism (everything has a bit of consciousness and mind in it).

The consequences extend to other areas of philosophy. If materialism is correct, then knowing can be a physical process by which brains interact with the world,

but dualism and idealism make knowing into an entirely different and utterly mysterious kind of process. Ethics and aesthetics are also affected by determination of the nature of consciousness. In French and Spanish, the same word is used for both consciousness and conscience, showing the connection between experience and moral sense. Materialism is compatible with some ethical views such as consequentialism, but raises problems for the existence of free will that is often thought to be important for moral responsibility.

The attempt to give a strict definition of consciousness is futile, but the concept can be informatively characterized by a 3-analysis. The exemplars (standard examples) of consciousness include thoughts, emotions, self-awareness, external perceptions such as sounds, and internal perception such as pain. The typical features of such experiences include awareness, shifts in attention, beginnings and ends of states, and a degree of unity. Consciousness is valuable for explaining people's verbal reports such as when they say they are in pain, behaviors such as grimacing and writhing in pain, and the experiences such as pain that everybody has. The open question is what can explain these conscious experiences.

This 3-analysis handles Flanagan's worries about sense and reference of the concept of consciousness. The exemplars are indicators of the range of reference of the concept, which includes all the experiences that people have that are conscious: thoughts, emotions, perceptions, and so on. They are not things, like tables and trees, nor simple events such as sneezes. Rather, they are processes consisting of a series of interconnected events, ranging over seconds and minutes. The scientific task is to explain how such processes result from underlying mechanisms, such as the interactions of neurons.

The sense of the concept of consciousness comes from its interconnection with other concepts. Consciousness is a kind of mental process that is associated with typical features such as experience, awareness, and shifts in attention. The development of science will likely lead to alterations in the sense of this concept through the realization that consciousness results from brain mechanisms. Over time, there can also be conceptual changes involving reference, if it turns out that consciousness is not just a property of humans and other animals on our planet, but eventually extends to novel entities such as robots and aliens from outer space, on the basis of evidence such as complex behaviors.

Dualists want to keep consciousness outside the realm of scientific explanation, maintaining that science can never address "what it is like" to be conscious. But this vague question can be broken down into much more precise questions about the character of particular conscious experiences. For example, when people have emotions, they undoubtedly have feelings, but these feelings operate along various dimensions. Some emotions such as happiness feel good, while others such as fear feel bad; and emotions come in varying intensity, from mild

contentment to ecstatic joy. Science can aim to explain these variations and many other important aspects of emotions and other manifestations of consciousness.

Flanagan and Tononi agree that science should be able to deal with consciousness, but Flanagan rightly does not endorse Tononi's approach using Integrated Information Theory. At first glance, IIT sounds promising, for it is a characteristic of human consciousness that we take in information from many sources including external and internal perceptions and integrate them together into unified experiences. For example, when you play a musical instrument, you tie together physical movements, sounds, touches, and what you see. So, information integration is a characteristic of conscious experience, but Tononi's attempt to generalize this into an explanatory theory falls short in many ways (Thagard and Stewart 2014; Thagard forthcoming-a).

In the dialogue, Tononi says that an experience is identical with the conceptual structure which is the maximum of intrinsic cause-effect power of a certain form. But how can an experience, which is an event, be the same thing as a structure, which is not an event? Flanagan points out the difficulty of understanding Tononi's concept of intrinsicness, and cause-effect power is equally mysterious. You might think that in Tononi's book and numerous published articles such ideas are clarified, but if you go through them carefully you will discover that obscurity builds on obscurity. Tononi never succeeds in characterizing either information or integration at a level sufficient for good scientific theorizing.

Scientific rigor often comes from mathematical formulations, and IIT might seem to reach the standard because its key quantity, PHI (Φ), is a measure of information integration that is defined mathematically. But a careful analysis of PHI shows that it is not efficiently computable beyond a small number of elements, and therefore is useless for characterizing consciousness in brains with their billions of neurons.

Moreover, interpreting consciousness as information integration extends consciousness to far too many entities. Tononi admits that even simple photodiodes do a small amount of information integration, and it is clear that cell phones do a huge amount (Tononi 2008, 236). Smart phones tie together many sorts of information acquired by Wi-Fi, camera, microphone, GPS, and touchscreen. But there is absolutely no behavioral evidence or theoretical argument that a smart phone has even a tiny amount of consciousness.

Tononi's view sounds a bit like panpsychism, because it extends consciousness beyond the minds of humans and similar animals. But he does not actually claim that everything has some consciousness, because not everything is capable of integrating information. Nevertheless, IIT is ontologically excessive in its attribution of consciousness to many entities that show no behavioral evidence of being conscious. Such behavioral evidence is not limited to humans, who have the added benefit of being able to make reports about their experiences, but

extends also to many animals, including mammals, probably birds, and maybe even reptiles and fish. All of these have brains with large numbers of neurons that integrate diverse inputs and produce complex behaviors. So it is a reasonable conjecture that consciousness as it currently exists in the universe results from neural mechanisms.

There are currently two advanced neurocomputational theories of consciousness available, and undoubtedly more will be developed. Stanislas Dehaene (2014) proposes that consciousness is the global availability of information encoded and broadcast in a neuronal workspace, providing an impressive array of relevant evidence such as brain scans in support. This theory is far clearer than Tononi's, and is supported by computational models that show it can be made mathematically precise.

Another alternative that goes into far more detail about how neural mechanisms encode information is the semantic pointer theory of consciousness (Thagard and Stewart 2014; Thagard, forthcoming-a). Semantic pointers are neural processes that explain cognition and emotion as results of the binding of representations that include sensory inputs (Eliasmith 2013). For example, the concept *cat* is not just verbal, but operates in the brain as a pattern of neural firing that incorporates inputs that include sight (the cat's color), sound (the cat's meow), and touch (the smooth fur). Like neural processes in general, semantic pointers are rarely conscious, but some break through to consciousness when they outcompete other representations for limited resources of attention. According to this theory, the main mechanisms responsible for consciousness are: (1) neural firings that result from sensory inputs and internal brain processes, (2) binding of these firings into more complicated patterns that can function as symbols, and (3) competition among semantic pointers for attention.

Unlike Dehaene's theory, semantic pointers can explain why experiences differ. Neural firings result from sensory inputs, other neural firings, and internal binding processes that produce new neural firings with emergent properties such as being able to function like symbols. Seeing a cat is different from hearing a trumpet, and the imagined experience of seeing a cat playing the trumpet is different still, all because of the production of different neural firings and bindings into semantic pointers. Such perceptions are different from emotional experiences, which result from different kinds of binding of (1) neural firings for internal bodily perception such as rapid heartbeat and (2) cognitive appraisal of the relevance of the situation to the goals of an organism.

The semantic pointer theory of consciousness can explain many other aspects of conscious experience, such as why there are shifts in attention, why consciousness stops and starts, and why it is often unified. It is only a step on the road to a full scientific theory of consciousness, but shows good prospects for a mechanistic, materialist theory of consciousness. In contrast, centuries of dualism

and idealism have added little to the understanding of consciousness.

Flanagan suggests that it is an advantage of IIT that it is not biologically chauvinistic, allowing consciousness for any system with a high PHI value. But there is nothing chauvinistic about providing an exclusively biological explanation for a phenomenon that so far has only been found in living animals. If robots or space aliens turn up with good evidence of conscious experience, for example through their verbal reports and complex behaviors, then we can appropriately ask whether their consciousness results from mechanisms similar to those in a biological theory of consciousness. Our theories of consciousness may then be revised accordingly.

Chris Eliasmith and his research group are already working with robots that simulate neural firings and bindings, and it would not be hard to incorporate semantic pointer competition as well. But there is no reason to suppose that current robots are conscious, because nothing in their behaviors suggest that they are. For example, they show no evidence of pain or other experiences, and they make no reports of conscious experiences.

Even though semantic pointers are currently running on special computer chips that have a loose approximation to human brains and control the motions of robots, these robots and chips are still enormously different from human bodies. Neural firing results from much more than electrical excitation and inhibition in artificial neurons. Whether a neuron fires is affected by chemical signals through more than 100 different neurotransmitters, chemical signals from glial cells, numerous hormones, and interactions with the immune system. I suspect, therefore, that if robot consciousness develops it will be very different from human consciousness.

Tononi says that consciousness is intrinsic existence, existing in and of itself, whereas the physical world exists extrinsically, derivative of our existence. This idealist metaphysics is utterly at odds with what science has learned about the development of the universe. When the Big Bang took place more than 13 billion years ago, there were no particles, let alone molecules. It took billions of years for life to begin on our planet, and billions more before animals evolved with sufficiently complex sensory systems, behaviors, and nervous systems to be credibly judged as conscious. All currently conscious beings, including people and many other animals, lack Tononi's mysterious intrinsic existence: We all result from biological mechanisms of evolution, reproduction, genetics, digestion, metabolism, and so on. Tononi's intrinsic/extrinsic distinction is a remnant of prescientific prejudice that yearn to place humans at the center of a heartless universe.

It is possible that we will eventually have to take seriously the "mysterian" view that humans simply are not smart enough to figure out how consciousness works. But it would be ridiculous to adopt it before centuries of investigative

effort. Science is much younger than philosophy, which goes back a few thousand years. Physics only began seriously in the 17th century, and biology in the 19th century Cognitive neuroscience, the field most relevant to providing mechanistic explanations of consciousness, is only a few decades old.

Nevertheless, there has been substantial progress concerning the neural and molecular basis for complex thinking, including perception, learning, emotion, and problem solving. I predict that in a matter of decades rather than centuries the problem of consciousness will join other previously mysterious phenomena such as planetary motion, the origin of species, and the operation of life in the pantheon of scientific accomplishments. Philosophy cannot lag behind in mystical ignorance.

4 Do People Have Free Will

Philosophical problems do not arise in isolation from each other. The mind-body problem and the question of consciousness are intimately connected, because whether or not mind is material depends on whether a materialist explanation of consciousness can be given. Both of these problems are highly relevant to the question of free will, concerning whether people's actions are determined or freely chosen. Materialist accounts of mind and consciousness threaten the existence of free will, because if your mind and conscious decisions are just neural processes, free choice may seem like an illusion.

Wheatley says that the concept of free will assumes that people carrying out an action could consciously choose otherwise. For example, when I am about to eat a chocolate bar, I could consciously decide that it has too many calories and choose not to eat it. She doubts that this kind of free will exists, presumably because there are psychological and neural mechanisms that determine what people decide and choose. Horgan points out problems in interpreting counterfactual statements such as "I could have not eaten the chocolate *if* I had chosen."

Pereboom and Brass discuss different aspects of free will, including its connection with moral responsibility as well as the ability to do otherwise. Brass says that neuroscientific experiments do not strongly relate to the philosophical issue of whether free will exists. But he is inclined to believe in the existence of free will because it makes life easier to think of yourself as a free agent. In contrast, Pereboom doubts that people have the sort of freedom required for moral responsibility that involves people deserving to be blamed. He thinks that not believing in free will can make people more compassionate and less reactive to others, making social life easier.

I think that philosophy and science can work together to help resolve the

various problems about free will. First, psychologists and neuroscientists can continue to conduct experiments that reveal details about the mental and neural processes of decision-making. The results of these experiments alone do not answer the question of whether people have free will, but they can provide evidence for theories that do have such implications.

Second, scientists can attempt, possibly in collaboration with philosophers, to develop theories of decision-making and choice evaluated on the basis of their ability to explain the full range of available evidence. In cognitive science, explanations are usually mechanistic, specifying combinations of connected parts whose interactions produce regular changes. For psychology, the parts are mental representations and the interactions are computational processes that change these representations by means of inference. For neuroscience, the parts are neurons and the interactions are excitations and inhibitions based on synaptic connections, leading to new patterns of neural firing including ones that can cause actions. Increasingly, these psychological and neuroscientific explanations are growing together, because of neurocomputational theories of mental representations such as the Semantic Pointer Architecture.

Accordingly, science can continue to develop neuropsychological theories of decision and action. One recent theory models both intentions and emotions as semantic pointers and describes actions as resulting from neural processes in numerous brain areas, including the thalamus, basal ganglia, amygdala, prefrontal cortex, and motor cortex (Schröder et al. 2014). However, this theory has not yet been closely integrated with the semantic pointer theory of consciousness using unified computational models. Once good theories of action, decision, choice, and consciousness are in place, it will become possible to address the fundamental philosophical question about whether people could have done otherwise if they had chosen.

The third step in a philosophical-scientific approach to free will requires the involvement of philosophers, because it deals with general and normative questions that go beyond the experimental results and scientific theories developed in the first two steps. Based on scientific theories that are more solid than currently exist, it should be possible to answer fundamental questions about whether actions are causally determined and whether people have free choice. Then philosophers, with the assistance of cognitive scientists familiar with the relevant theories and experimental results, can address the moral question of whether the resulting account of choice is compatible with the concerns of morality, for example with respect to blame and punishment.

We currently do not have enough scientific knowledge to resolve these issues, but the overall trajectory of scientific progress leads me to make some conjectures. Although there will never be a time when there is an absolutely certain account of the neural mechanisms that produce every human action,

better theories and better experiments are combining to produce much understanding of why people make the choices that they do. Nothing in this trajectory suggests the need to invoke random processes such as those implied by quantum theory, and nothing suggests the need to invoke supernatural entities such as transcendental agents. To advance, we need better theories and computational models, integrating strong theories of action and consciousness.

A likely result is that traditional notions of free will turn out to be as obsolete as religious notions of soul and immortality. Nevertheless, within the evolving mechanistic accounts of human decision, there are different kinds of processes that have implications for a range of ideas about freedom. Brains operating entirely on instinct, such as those of insects, lack even the tiniest amount of freedom. People engaged in automatic behaviors such as breathing, walking, and brushing their teeth are not much better off.

However, a good theory of making decisions has to recognize that people do act differently when they are thinking more consciously and deliberately. For example, people can use implementation intentions to help them to overcome temptation (Gollwitzer 1999). I can consciously plan to apply this rule: If I am offered a chocolate bar, then I will think not to eat the chocolate bar because of its high caloric count. Conscious choices are different from ones based on instinct or automatic behavior. What I am calling the automatic and deliberate modes are what psychologists usually call system 1 and system 2 dual processes, or thinking fast and slow (Kahneman 2011).

Psychology currently lacks a theory of the neural mechanisms that distinguish the two modes, but the Semantic Pointer Architecture suggests how to build one. Usually, the brain operates in automatic mode, with semantic pointer operations of binding, inference, and action carrying out ordinary activities of perception, problem solving, learning, and behavior. Sometimes, however, semantic pointers such as images, concepts, and rules cross a threshold of firing activity that produces conscious experience. The critical transition to consciousness results from emotional evaluations of the importance of those semantic pointers to the goals and current activities of the brain, outcompeting more peripheral representations. The concepts, rules, and images in consciousness can then have a bigger impact on behavior than unconscious processes. This impact, however, can be eliminated if other semantic pointers outcompete the concepts and rules that contribute to good decisions.

The difference between automatic and deliberate modes is not sufficiently great to allow attribution of free will in the traditional sense to human behavior, for two reasons. First, given neural mechanisms for intention, action, emotion, and consciousness, these deliberate decisions are still causally determined through neural processes. Second, people rarely choose to make their decisions consciously and deliberately rather than automatically. It is an open question why

people, who normally carry out most of their behaviors automatically, are sometimes spurred to think more consciously and deliberately about what they do; but there is no reason to believe that people actually choose to operate in deliberative mode when they want to.

Nevertheless, the deliberative mode opens up the possibility of considering reasons furnished by others who express moral blame or approval. If you are actually thinking about what you are doing, then you have the capacity to take into account what your moral codes require and what other people advise you to do. This capacity is demolished if you are subject to mental illnesses such as schizophrenia that make you delusional, or if you are subject to intense coercion from other people. But without illness and coercion, people have the capacity to think consciously about what they are doing in ways that can be influenced by moral considerations.

Such deliberation is not free in the transcendental sense that can only be provided by a supernatural soul, but it can make major differences in how people interact with each other. In my *Treatise on Mind and Society*, I say that this sort of will is approximately free, or *freeish*. In accord with Wheatley's injunction, I am not redefining the concept of free will but rather introducing a new concept more in line with scientific findings. Freeish will is not strong enough to justify the kind of responsibility and punishment based on desert that Pereboom rejects. But it fits well with a social, consequentialist account of blame and punishment aimed at improving people's overall behavior. Because people sometimes use a conscious, deliberative mode of thinking, holding people responsible for their actions can help to produce a better society.

Given these developments in the cognitive neuroscience of decision making, what should we make of the question of whether people sometimes could have chosen otherwise? I agree with Horgan that the traditional philosophical way of dealing with counterfactuals in terms of possible worlds semantics is not helpful, because we know nothing about whether there are possible worlds in which I rejected the chocolate instead of gobbling it.

A much more plausible account of counterfactuals comes via mechanisms (Thagard forthcoming-c). Counterfactual statements such as "if I had dropped the glass, it would have broken" are neither true nor false, because they do not directly correspond or fail to correspond to anything in the world. Nevertheless, they can be plausible or implausible based on underlying mechanisms, which in this case include the force of gravity and the molecular forces holding the parts of the glass together. Given enough force applied to a weak structure, we can confidently assess that the glass will break when dropped.

We do not yet have enough detailed knowledge of how brains make decisions to assess counterfactuals about people's choices and actions, but some relevant information is available. When people are stressed, tired, or hungry, they are

more likely to operate in the automatic mode that does not employ the additional concentration and effort needed for conscious deliberation. Therefore, if I am in one or more of the states of stress, fatigue, and hunger, then I probably would not have been able to decide otherwise than to eat the chocolate bar; such choices are automatic. On the other hand, when I have the energy and concentration to operate consciously in deliberate mode, then it is more plausible that I could have chosen not to eat the chocolate bar and therefore to have acted differently.

Does this mean that people have free will at least some of the time, when they have the mental resources to operate consciously and deliberately? The answer is no, for the reasons already mentioned: we cannot choose what modes we are operating in, and even the conscious, deliberate mode is still carried out causally by brain mechanisms. Nevertheless, the counterfactual test of whether you could have done otherwise is an indicator of the attenuated version of freedom that I have called freeish will, which suffices for consequentialist notions of moral responsibility.

5 How Can Philosophy and Cognitive Science Move Forward?

The procedures I recommended for dealing with problems of mind, consciousness, and free will generalize into a full answer to the question of how philosophers can work with scientists in the attempt to understand fundamental issues about knowledge, reality, and morality. Philosophers should neither proclaim their autonomy and superiority in addressing issues in epistemology, metaphysics, and ethics; nor should they capitulate to scientists such as Stephen Hawking who declare the irrelevance of philosophy (Hawking and Mlodinow 2010). Instead, philosophy and science can collaborate to address issues of fundamental importance to the future of human beings.

The weakest form of collaboration consists of philosophers just keeping track of what happens in science and then trying to use empirical findings and scientific theories to become more effectively general and normative. But I urge philosophers to be more active in several ways. The burgeoning field of experimental philosophy shows that philosophers can conduct their own experiments on topics neglected by psychologists (e.g., Sytsma and Buckwalter 2016). A few philosophers are even conducting neuroscientific experiments using brain scanners. Theorizing based on experimental findings can always be conducted in the armchair, although considerable effort is required to develop methods such as computer modeling that are an important part of current theory in psychology and neuroscience.

Philosophers daunted by the difficulty of acquiring the practical knowledge of

how to conduct experiments and computer simulations can pursue the invaluable path of collaborating with people in the sciences who have the skills. It is an oddity of philosophical practice that philosophers rarely collaborate (Thagard 2006). Almost all scientific articles are now co-authored, whereas few philosophy articles are. I think that the individualism of philosophy is a remnant of outmoded ideas of philosophical method based on pure reason, an inherently solitary approach. You do not need collaborators to consult your own intuitions. Once the barrenness of a priori thought experiments is appreciated, philosophers can open up to collaboration with other philosophers and with scientists in relevant fields.

Scientists acquire the practical knowledge of how to collaborate as part of their graduate school education, whereas philosophers and other humanists are usually expected to go it on their own. In my own work at the intersection of philosophy and cognitive science, I have benefited enormously from working with psychologists and computer scientists. Experimental philosophy is also often collaborative, including fruitful interconnections with psychology. There is great potential for advancing philosophy in league with science through individual and collective work.

References

Blouw, P., Solodkin, E., Thagard, P., & Eliasmith, C. 2016. "Concepts as Semantic Pointers: A Framework and Computational Model." *Cognitive Science* 40: 1128–62.

Colaço, D., Buckwalter, W., Stich, S., & Machery, E. 2014. "Epistemic Intuitions in Fake-Barn Thought Experiments." *Episteme* 11.2: 199–212.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York: Viking.

Eliasmith, C. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Gollwitzer, P. 1999. "Implementation Intentions: Strong Effects of Simple Plans." *American Psychologist* 54: 493–503.

Hawking, S. S. W., & Mlodinow, L. 2010. *The Grand Design*. New York: Bantam.

Kahnemann, D. 2011. *Thinking Fast and Slow*. Toronto: Doubleday.

Murphy, G. L. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Parry, R. 2014. *Episteme* and *Techne*. *Stanford Encyclopedia of Philosophy*. Retrieved on August 6th, 2018 from https://plato.stanford.edu/entries/episteme-techne/

Schröder, T., Stewart, T. C., & Thagard, P. 2014. "Intention, Emotion, and Action: A Neural Theory Based on Semantic Pointers." *Cognitive Science* 38: 851–80.

Sytsma, J., & Buckwalter, W. (eds.). 2016. *A Companion to Experimental Philosophy*. Oxford: Wiley-Blackwell.

Thagard, P. 1992. *Conceptual Revolutions*. Princeton: Princeton University Press.

Thagard, P. 2006. How to Collaborate: Procedural Knowledge in the Cooperative

Development of Science." *Southern Journal of Philosophy* 44: 177–96.
Thagard, P. 2009. "Why Cognitive Science Needs Philosophy and Vice Versa." *Topics in Cognitive Science* 1: 237–54.
Thagard, P. 2012. *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*. Cambridge, MA: MIT Press.
Thagard, P. 2014. "Thought Experiments Considered Harmful." *Perspectives on Science* 22: 288–305.
Thagard, P., & Stewart, T. C. 2014. "Two Theories of Consciousness: Semantic Pointer Competition vs. Information Integration." *Consciousness and Cognition* 30: 73–90.
Thagard, P. (forthcoming-a). *Brain-Mind: From Neurons to Consciousness and Creativity*. Oxford: Oxford University Press.
Thagard, P. (forthcoming-b). *Mind-Society: From Brains to Social Sciences and Professions* Oxford: Oxford University Press.
Thagard, P. (forthcoming-c). *Natural Philosophy: From Social Brains to Knowledge, Reality, Morality, and Beauty*. Oxford: Oxford University Press.
Tononi, G. 2008. "Consciousness as Integrated Information: A Provisional Manifesto." *Biological Bulletin* 214: 216–42.