Review of General Psychology
2018, Vol. 22, No. 1, 85–94

© 2017 American Psychological Association
1089-2680/18/$12.00 http://dx.doi.org/10.1037/gpr0000125

Meaning, Autonomy, Symbolic Causality, and Free Will

Russ Abbott
California State University, Los Angeles

As physical entities that translate symbols into physical actions, computers offer insights into the nature of meaning and agency. Physical symbol systems, generically known as *agents*, link abstractions to material actions. The *meaning* of a symbol is defined as the physical actions an agent takes when the symbol is encountered. An agent has *autonomy* when it has the power to select actions based on internal decision processes. Autonomy offers a partial escape from constraints imposed by direct physical influences such as gravity and the transfer of momentum. Swimming upstream is an example. *Symbols* are names that can designate other entities. It appears difficult to explain the use of names and symbols in terms of more primitive functionality. The ability to use names and symbols, that is, symbol grounding, may be a fundamental cognitive building block. The standard understanding of causality—wiggling X results in Y wiggling—applies to both physical causes (e.g., one billiard ball hitting another) and symbolic causes (e.g., a traffic light changing color). Because symbols are abstract, they cannot produce direct physical effects. For a symbol to be a cause requires that the affected entity determine its own response. This is called *autonomous causality*. This analysis of meaning and autonomy offers new perspectives on *free will*.

Keywords: autonomy, autonomous causality, free will, meaning, symbolic causation

This article is divided into six parts. This introduction provides a roadmap. It also explains why it makes sense to discuss computers in an article in a psychology journal.

Section two discusses how computers construct meaning. I use a player piano to illustrate an approach to meaning that is widely used in computer science: a symbol's meaning is the physical action(s) the system takes on encountering it.

Section three introduces autonomy. Autonomy requires one energy flow to control another. A standard light switch illustrates how one energy flow, the energy that flips the switch, can control another, the current in a circuit. But not all such situations imply autonomy. A system consisting of a switch, a circuit, and a light source has no autonomy. For an entity to qualify as having autonomy it must contain both the controlling and controlled energy flows. Under this definition, the player piano introduced in section two can be understood as having a limited form of autonomy.

In preparation for a discussion of how symbols can act as causes, section four examines the longstanding question of what a symbol is. That section also examines the ability of biological organisms to work with categories, concepts, and individuals. The section suggests that these abilities may be three aspects of a widespread cognitive capability: an ability to work with cognitive entities.

Section five reframes the earlier discussion of autonomy in terms of symbols and causality.

The final section examines the implications of the preceding sections for free will.

Why Discuss Computers in a Psychology Journal?

Newell and Simon (1976) characterize computers as physical symbol systems. As physical devices computers act in the material world; in addition, computers manipulate symbols. This joint capability is fundamental to how symbols get meaning.

The definition of meaning outlined in section 2 will link symbols to actions. Doing so requires a bridge between the abstract and the physical. A physical symbol system provides that bridge. A corollary is that any entity, including human beings, capable of giving meaning to symbols must be a physical symbol system.

A second reason to include a discussion of computers is their transparency. Technology does not yet enable one to follow in detail the functioning of human (or animal) brains, much less their minds. The man-made nature of computers mitigates that problem. Even so, it may not always be feasible to explain how the detailed functioning of a computing system produces its higher level results.

How Computers Transform Symbols to Meaning

This section presents a simplified version of operational semantics (Fernández, 2014), a computer science approach to semantics. I use the player piano as an example.

The modern equivalents of a player piano (electronic keyboards, digital pianos, digital synthesizers, etc.) are physical symbol systems. They can read and manipulate symbols—in the form of digitized musical scores—and transform them into sounds.

This article was published Online First September 21, 2017.

Correspondence concerning this article should be addressed to Russ Abbott, Department of Computer Science, California State University, Los Angeles, Los Angeles, CA 90032. E-mail: RAbbott@calstatela.edu

Assume that the internal logic of a modern player piano is expressed as software. That software includes commands to perform actions that produce sounds. An example might be an action to strike selected strings with felt-clad hammers. (These commands are similar to the more familiar print command, which performs an action that produces output.) Assume those commands are something like `strike_C`, `strike_C#`, `strike_D`, and so forth. (For simplicity, I limit the discussion to a single octave.)

We can make up a simple notation for notes: *C/4* means the note *C* for a quarter note. Using this notation, the first four bars of "Mary Had a Little Lamb" might appear as follows (the lyrics are added for the reader's convenience):

E/4 D/4C/4D/4E/4 E/4 E/2 D/4 D/4 D/2 E/4 G/4 G/2

Ma - ry had a lit - tle Lamb, lit - tle Lamb, lit - tle lamb.

Music is produced when our device reads such a score and translates its symbols into physical acts.

How Symbols Are Converted to Actions

Converting symbols to actions sounds like an impossibility; symbols are abstract, and actions are physical. Yet it is commonplace for computers to control physical devices on the basis of symbolic computations. A player piano is such a device.

The software representing the player piano's logic will consist of two primary components. An outer layer iterates through the score symbols. For each symbol it calls a subcomponent, perhaps called `play_score_symbol`, to convert that symbol into an action.

The job of `play_score_symbol` is to do whatever the symbol given to it means. The body of `play_score_symbol` might look something like the following (note that == tests for equality):

```
if score_symbol == C then strike_C;

if score_symbol == C# then strike_C#;

if score_symbol == D then strike_D;

...
```

In other words, for each `score_symbol`, perform the action associated with that symbol. In doing so, the player piano gives each `score_symbol` a meaning by linking it to a specific action.

Note how arbitrary the association of symbols with actions are. Had `play_score_symbol` associated symbols with commands to strike keys other than those shown, the symbols would have had those other meanings—at least for this player piano.

One might object that these symbols have well-established meanings independent of this player piano and that if `play_score_symbol` had associated symbols with commands to play notes that were not consistent with those well-established meanings, `play_score_symbol` would be wrong.

I think that calling it *wrong* is not the best way to look at it; I prefer *inconsistent*. To communicate, people must agree about what symbols mean. There must be, as Baumeister and Monroe (2014) would probably argue, a cultural aspect. If a symbol means the same thing to multiple people, that is because they agreed to make it have that meaning. But these cultural agreements do not *create* the meanings. I doubt that anyone would argue that the symbol *C* has a meaning independent of how people (or player pianos) interpret it. A shared culture allows us to make the meaning one person gives a symbol consistent with the meaning others give that symbol. But I would not say that culture—absent the people, if that is even meaningful—provides a meaning on its own. In other words, the meaning is not "out there" waiting to be discovered. It must be created by each individual—and then synchronized among individuals so that they can communicate with each other.

Actions Are Not Necessarily External

For a computer (or our player piano) to take an action does not always imply that the action is evident to an outside observer. Actions often change the internal state of the system. For example, the discussion above ignored the time value of symbols. (Recall that *C/4* means a *C* for a quarter note.) We did not discuss how the system treats quarter notes differently from, say, half notes. One way for the system to do that would be to store, for each note being played, the time duration over which the sound should be produced. Storing that information involves a change to the internal state of the system. That too is a physical act. But it is not one that an outside observer can easily see.

I selected a player piano as an example because much of what it does results in externally observable actions (i.e., the production of sounds). In contrast, most of the actions performed by most computing systems involve internal state changes. These actions are still physical: The physical state of the computer changes. But these state changes are not easily observed from outside. Yet making these internal state changes is part of a symbol's meaning.

Operational semantics consist of just this: characterizing the external actions and internal state changes that occur when a symbol or symbolic construct is interpreted.

Autonomy

Although it may seem strange to say so, a player piano as just described exhibits a simple form of autonomy. It determines, through its internal logic, what a symbol means. A symbol cannot *force* a meaning on the player piano. The meaning of a symbol is determined internally.

I realize how strange that may sound. But look again at the code above. It is the code that selects for each symbol what the player piano's response will be. Contrast this with how a billiard ball responds when hit by another ball. The ball that was hit has no choice about how to respond.

One might argue that the software inside a player piano is fixed and that therefore the way the software interprets symbols is fixed—so there is no room for autonomy. That, I claim, is another question. Autonomy turns on whether an entity follows its own internal rules, fixed or not.

The rest of this section examines autonomy in more detail.

Switches and Two-Level Energy Flows

Move a light switch to the *on* position, and an associated light source (i.e., bulb, LED, etc.) emits light—assuming the circuit is intact, the light source is functioning, current is supplied to the circuit, and so forth. Move it to the *off* position, and the light source stops emitting light.

A switch has the crucially important property that it allows one energy flow to control another. A light switch controls current in a circuit. The energy used to flip the switch (i.e., the energy expended by a person when flipping the switch) is not the same energy flow as the energy in the circuit. Although not sufficient for autonomy, this two-level mechanism is an important prerequisite.

It is worth stopping to appreciate the power of switches. Switches enable human beings to use their own energy to control devices they could not otherwise control. Much of civilization depends on devices that are powered by nonhuman energy but controlled by human beings. And of course the transistor, the fundamental computer building block, is at heart a switch.

Switches and Autonomy

Switches do not on their own imbue something with autonomy. Whether or not a light source emits light is controlled from outside the light source. Neither the light source nor its circuit has anything to say in the matter. Neither has the means to determine when the switch is flipped. Autonomy arises when an entity includes switches within itself and when it is organized to use those switches to control the expenditure of its energy resources.

To make this point clearer, I will examine two biological examples, both involving *E. coli* cells. One does not involve autonomy; the other does.

François Jacob and Jacques Monod (Jacob & Monod, 1961) famously discovered gene switches. For their work they were awarded the 1965 Nobel Prize in Physiology or Medicine. In *E. coli*, genes involved with the metabolism of lactose are expressed only when lactose is present. (The presence of glucose is also a factor, but for our purposes that can be ignored.) The presence of lactose (actually a byproduct of lactose) flips a switch for the expression of the lactose metabolism genes.

The switch mechanism involves a protein (called a repressor), which normally attaches to the DNA immediately prior to the relevant genes. The repressor blocks RNA polymerase from reading and transcribing the gene. When present, the lactose byproduct attaches to the repressor and pulls it off the DNA. This unblocks the RNA polymerase access to the genes. The genes are then read and transcribed. This mechanism fits our switch pattern. The presence of lactose unblocks the chemical energy that enables RNA polymerase to crawl the DNA strand and read and transcribe genes.

This mechanism does not reflect autonomy. There is no decision-making process within the bacterium that controls the expression of the genes. A gene switch is a chemical reaction that functions in much the same way as a light switch. In effect, lactose reaches into the cell and flips a switch.

In contrast I would credit *E. coli* with autonomy in how it navigates a liquid nutrient-rich environment (see Hu & Tu, 2014). Each *E. coli* cell moves through its environment by alternating runs and tumbles. When the cell's flagella rotate counterclockwise a run occurs; the cell moves in a straight line. (The name notwithstanding, *E. coli* flagella are stiff rather than whip-like; when rotating counterclockwise they produce an outboard motor effect.) When the flagella rotate clockwise a tumble occurs; the cell moves randomly and reorients itself for the next run.

E. coli internal chemistry is capable of sensing a difference in nutrient concentration between its front and rear. That difference determines for how long a run persists. If the cell is moving along a nutrient up-gradient, runs continue longer; otherwise, they stop sooner. The cell then tumbles and reorients itself (randomly) for the next run. The effect is to move the cell toward a nutrient source.

This too is a switch. The chemical decision-making process is one energy flow. It controls another energy flow, which powers the flagella.

I credit *E. coli* with autonomy because it contains *both* the decision-making process *and* the energy flow controlled by that process.

Symbols

Consider the following line from the program snippet shown earlier:

```
if score_symbol == 𝒞 then strike_C;
```

Every word-like element in that line, “`if`,” “`score_symbol`,” “`==`”, “𝒞”, “`then`,” “`strike_C`,” represents a symbol. Recall that many of these appear multiple times in the program. Each appearance of one of these elements represents the same symbol. That raises two questions: (1) What does one mean by “the same symbol?” (2) What is that “same symbol” entity? This issue is not unique to computer science. The same question arises in every academic discipline that uses symbols. If a symbol, say, *x*, appears multiple times in a mathematical expression, it is generally understood that each appearance represents “the same” symbol.

The question of what a symbol is has a long philosophical history. See, for example, Balaguer, 2016; Bricker, 2016; Floridi, 2017; Nelson, 2016; Orilia and Swoyer, 2016; and Reicher, 2016. Space limitations preclude a review of this literature.

I find it useful to think of symbols as similar to names. For example, consider the name *John*. One might find that name in a baby name book. One might select that name for one's own baby—after which *John*, when used in the appropriate context, refers to that child. The name itself, however, is independent of any person. If two people are both named *John,* it is said that they have “the same name.” No one seems confused by statements of this sort. Yet I suspect it would be very difficult to provide a rigorous completion for the sentence “A name is ___,” where the blank is filled in by properties that characterize names.

Newell and Simon (1976) noted that a symbol has two properties: It can be manipulated as an entity, and it is capable of designating something else. Names also have those properties—although Newell and Simon do not note the similarity. This is what Helen Keller realized in her flash of insight. She called it the mystery of language. Things have names; names can refer to things (Keller, 1903). Interestingly, Keller experienced her realization as like a memory. “I felt a misty consciousness as of something forgotten—a thrill of returning thought.” It is as if the ability to connect names (or more generally symbols) to things comes built into our brains.

I realize that in saying a symbol is a name, I have not solved the problem of what a symbol is. I have only reformulated the question to ask what a name is.

The question raised above of how to define the word *name* begins to point toward an answer. A name is not defined by a set of properties; it is defined by how it is used. This is different from many nouns: *Fire*, for example, can be defined by a set of properties. Virtually any term can be a name if it is used as a name, that is, as a concise way to refer to something.

Concepts, Categories, and Individuals

I speculate that a wide range of organisms have the ability to work with what might be called cognitive entities (such as concepts, categories, and individuals) and that names and symbols represent an advanced version of this basic cognitive ability.

Concept neurons. Quiroga (2012) has shown that humans have "concept neurons." (See also the discussion in Quiroga, Fried, & Koch, 2013.) Originally called "Jennifer Anniston neurons" and "Halle Berry neurons" (because both actresses were used in these experiments), these neurons are triggered by a wide range of pictures of their subjects, including full face, profile, close-up, full body, and the subject dressed in various outfits. Even more striking, the same neurons responded to the printed names of these actresses! This suggests that these neurons were responding to concepts, not visual patterns.

Concept neurons are also found in nonhumans. Both carrion crows and rhesus macaques have neurons that respond to specific numerical amounts. A particular neuron will respond to both one dot and one beep. Another will respond to both two dots and two beeps. Other neurons will respond to numeric values up to at least four dots and four beeps (Ditz & Nieder, 2015).

Suggestive evidence of conceptualization. Loukola, Perry, Coscos, and Chittka (2017) recently showed that bumble bees are able to learn a task that seems to require conceptualization. To obtain a reward, bees were required to move a ball to a target location. Bees that observed another bee perform the task generally learned the task faster.

The observation scenario involved three balls. The two closest to the target were fixed in place. The demonstrator bee was thus forced to move the most distant ball. When the observer bees were tested, all three balls were mobile. The observer bees moved the ball closest to the target. Loukola put it this way (Loukola et al. 2017, p. 836):

> [T]he bees did not simply copy the behavior of the demonstrator but rather improved on the observed behavior by using a more optimal route.

These results suggest that bees may recognize *ball* as a category and each of the balls as an instance of that category.

Other experiments also suggest animal conceptualization. In an early language learning experiment a chimpanzee constructed the lexigram sequence "coke that is orange" to request an orange soda (Rumbaugh, 1977).

And then there is Alex, Irene Pepperberg's parrot.

> For 25 years, I have taught Gray parrots meaningful use of English speech (e.g., to label objects colors, shapes, categories, quantities, and absence). Using this code, my oldest subject, Alex, exhibits cognitive capacities comparable to those of marine mammals, apes, and sometimes 4-year-old children. (Pepperberg, 2002, p. 83)

Alex could correctly answer questions about individual objects such as "What color?" or "What shape?", about comparisons between objects such as "What's different?" or "What's same?" [with respect to the properties color, shape, and size (large, small)], and about collections of objects such as "How many?". Alex was even able to answer "none" to a "What's different?" question if there were no differences in the trained properties or to a question such as "How many red blocks?" if there were no red blocks in a given collection. (Pepperberg, 1999).

Categorization. One way to study conceptualization is through categorization. Smith (Smith, Zakrzewski, Johnson, Valleau, & Church, 2016) provides a broad survey of animal categorization capabilities. A wide range of nonhuman animals exhibit categorization capabilities. Smith notes that

> Categorization has conferred fitness advantages on vertebrates for hundreds of millions of years. For example, vervet monkeys (*Chlorocebus pygerythrus*) have developed call signs that warn group members to behave appropriately at the sight of martial eagles (*Polemaetus bellicosus*). (e.g., Cheney & Seyfarth, 1990) In a sense, these calls denote or "name" members of the category eagle. (p. 1)

Since the call "names" a category, the category may function very much like a concept.

Discrete entities. Implicit but often unmentioned in studies of categorization is an ability to perceive discrete entities. After all, what are categories other than ways to group discrete entities?

A number of experiments have tested animals' ability to treat entities as persisting over time.

- Many animals achieve full (i.e., Piaget stage IV) object permanence (Gómez, 2005; Zentall & Pattison, 2016).
- Dogs (a) remained near a fallen owner, (b) avoided a deceptive human, and (c) preferred a human that provided valid information about the location of a reward over an uninformative human (Roberts & Macpherson, 2016).

Abstractions

A wide range of animals have evolved the ability to work with such cognitive entities as collections, concepts, individuals, names, and symbols. I would go so far as to refer to this sort of capability as an ability to form abstractions. Consider the significant fitness advantage possessed by entities capable of responding to their environment in terms of abstractions—such as friend versus foe, food versus predator, danger versus safety, or nutritious versus toxic—compared with those capable of responding only to raw sensory signals (such as photon impacts). Given the enormous fitness advantage that abstractions offer, it seems quite likely that means for creating at least some forms of abstraction evolved not long after the emergence of means for sensing the environment. Since each sensing instance is likely to differ in some way from every other sensing instance, what use is the ability to sense unless one can form abstractions from signals and then act on the basis of the abstraction?

The ability to use names and symbols to represent cognitive entities—the symbol grounding problem—may be a primitive cognitive building block (Harnad, 1990).

Symbols as Causes: Autonomous Causality

This section examines causality more carefully. Following are two of the most widely used philosophical definitions:

Physical causality (Dowe, 2000). A causal interaction involves the transfer of a quantity of some conserved property from one entity to another. The entity from which the conserved property is transferred is considered the cause. The entity to which the quantity is transferred reflects the effect of the transfer. A standard example is a billiard ball. When a moving billiard ball hits a stationary billiard ball, momentum is transferred from the former to the latter. Although physicists might describe such an interaction in more symmetric terms, human intuition is that the impact of the moving billiard ball *causes* the originally stationary billiard ball to move.

This approach to causation, while useful and convincing, is limited to direct physical interactions.

Interventionist causality (Pearl, 2000; Woodward, 2003). X has a causal relationship to Y if and only if there is a possible intervention directed toward X that, in changing the probability distribution of X, results in a change to the probability distribution of Y. Interventionist causality attempts to capture the intuition that if wiggling X results in Y wiggling, then X has a causal relationship to Y. Another way of putting it is that X has a causal relationship to Y if X can serve as something like a remote control for Y. Interventionist causality characterizes most people's intuitive understanding of causality.

Although interventionist causality requires that an empirical link be established between the cause (wiggling X) and the effect (Y wiggling), it does not require that one be able to explain the physical mechanism through which the cause produces the effect.

Symbols as Causes

Intuitive as it is, interventionist causality implies what may seem to be an unexpected result: Symbols can serve as causes.

1. A traffic light changing color causes traffic to stop or start. When a driver sees a traffic light, she converts the photons, that is, the physical elements received from the world, into a symbol such as *RedLight* or *GreenLight*. She understands that symbol and responds with some physical action—pressing the brake or accelerator—which in turn produces a physical effect on the vehicle.

2. Lowering or raising the price of an item causes more or fewer of those items to be sold. A more complex series of steps can be laid out for this interaction.

Causal interactions of this sort are so commonplace that we hardly think about them. Yet pinning down the causal mechanism produces some surprising results.

Autonomous Causality

Symbols have two defining properties.

1. **They are abstract and cannot produce physical effects**. As Rosen (2014) points out, abstract entities are by definition causally inefficacious.

2. **They have no distinctive individual properties.** Symbols are arbitrary and interchangeable. It sounds paradoxical, but (a) there is nothing distinctive about a symbol, other, perhaps, than an arbitrary label; yet (b) each symbol is distinguishable from all other symbols. (This does not hold for names, which are often designed to have aesthetic and semantic characteristics.)

A propertyless abstract entity such as a symbol would not seem to be a promising candidate to serve as a cause. A symbol's very abstractness excludes it from causing a physical effect. Even if one could get around this problem, since symbols have no individual properties, there would seem to be no way for specific symbols to produce specific effects. Yet we know that symbols serve as causes and that different symbols—*RedLight* versus *GreenLight*—produce different effects. Similarly, each of our note symbols, 𝒞, 𝒞#, . . . , produces a different sound.

The issue of symbols as causes, sometimes known as mental causation (Robb & Heil, 2014), has a long and vexing philosophical history. Yet at a common-sense level, there is no mystery. Symbols serve as causes when they are interpreted by agents. It is because of the way an agent interprets a *RedLight* that it means "Stop" to that agent, and it is because of the way a player piano interprets note symbols that they mean particular sounds.

Before going on I would like to look more carefully at the preceding sentence. I wrote that a red light means "Stop" because of the way an agent interprets it. I didn't write that a red light means "Stop" because that is its generally accepted, that is, cultural, meaning. To understand the distinction consider the following example.

Consider a company that builds an autonomous vehicle. The vehicle's software determines how it behaves on the road. Nothing prevents the company from writing software that interprets a red light to mean "Go." Of course, it would be foolish for the company to write software that interprets a red light to mean "Go." Doing so would likely ruin the company's sales; it would open the company to liability claims; and it might even result in criminal charges. But it would not be impossible to write such software. Such software would not change the culturally accepted meaning of a red light, but it would determine the meaning of red lights for the vehicles in which it was installed.

On the other hand, even a symbol's culturally accepted meaning can change. Suppose all companies wrote their software to interpret a red light to mean "Go." Suppose the laws were changed to make going through an intersection with a red light legal and going through an intersection with a green light illegal. Suppose all drivers changed their interpretations of red and green lights accordingly. Presumably someone new to that society would decide that the culturally accepted meaning of a red light is "Go." (Something similar actually occurred. In 1967, Sweden switched from driving on the left side of the road to driving on the right. Because people were so cautious, there were fewer accidents than usual. Accidents returned to their normal rate after two years; Flock, 2012).

Since the agent that interprets a symbol determines what the symbol's effect will be, I call this autonomous causality. Autonomous causality is a very strange notion, paradoxical even. As indicated earlier, a billiard ball that is struck has no choice about how it is affected. Momentum is transferred to it, like it or not. Yet when a symbol "strikes" an agent—that is, when an agent encounters a symbol—how the agent responds depends on the agent itself.

When speaking of the physics of causation. Laplace (1814/1951) famously said (emphasis added), “We may regard the present state of the universe as the *effect* of its past and the *cause* of its future.” But with autonomous causality it is not quite that simple. Instead of the laws of physics pushing the world around, agents themselves determine how they will respond to symbolic causes. The usual caveat applies: No laws of physics may be violated.

In short, an agent has autonomy when it uses internal rules to determine how to respond to symbols.

Not all interpreters are as simple as our example player piano. A jazz player piano may interpret a symbol to mean a certain note but play a different but related note. Similarly, a human being, and perhaps even an autonomous vehicle, may interpret a red light to mean “Stop” but then make the further decision to go through the intersection anyway. (How would you program an autonomous vehicle to behave if (a) it were headed to a hospital with a passenger with an emergency medical condition and (b) it came upon a red light at an intersection with no cross traffic? I would write the software to run the light after ensuring that there was no cross traffic.)

How Difficult Is It to Forecast Agent Decisions?

If having autonomy means that an agent uses internal decision-making processes to determine its behavior, how much can an outside observer predict about those processes? Often not very much.

Many agents are quite complex. Their internal decision-making processes may be so complex that even if one knew all the relevant details one might still not understand how they produce the results they do.

Consider AlphaGo, the computer system that recently beat the world Go champion (see Moyer, 2016 for a popular description). AlphaGo’s design combines two multilayer artificial neural nets with tree search (Silver et al., 2016). (Readers need not understand these technologies to understand the implications of this example.) An artificial neural net consists of a network in which the edges are given weights that determine the extent to which the node at one end effects the node at the other. The weights are determined through machine learning techniques.

In the course of its development, AlphaGo examined millions of Go games and played more millions of games against itself (Koch, 2016). In general, it is not possible to explain in any meaningful sense what it means for a weight to have a particular value. One can only conclude that the values that were arrived at work better than others.

Neural nets are particularly opaque in this way. But even traditional software is often so logically complex (or contorted or poorly documented) that one may not be able to explain how it works. As a simple example, imagine software that uses the Pythagorean theorem to compute the third side of a right triangle and then uses that result to estimate the time required for a trip. If the software includes the relevant arithmetic operations but with no explanation for why those operations are being performed, one may have difficulty explaining why the software estimates trip times successfully. All an observer would see is a sequence of computer instructions that perform arithmetic but might have no idea what the intention of the programmer was when writing those instructions.

The difficulty of understanding how agents make decisions is compounded by the fact that agents may change. Many biological organisms change (i.e., learn) as they experience their environments. Children go to school to learn effective ways to respond to symbols. People study with the goal of improving their performance. Society implements and publicizes policies that are intended to deter certain behaviors. Although the effectiveness of these practices varies from individual to individual, the effect is that agents change their decision-making processes based on their experiences.

Consider AlphaGo again. As it trained, it modified the weights in its neural net. This is standard machine learning. More interestingly, expert Go players were able to extract new strategies by watching AlphaGo play.

> AlphaGo has created a unique and extremely powerful approach to the game of Go. [After] an exhaustive analysis of the five games between AlphaGo and Lee Sedol and of three games AlphaGo played against itself shortly before the match, . . . it became clear to us that AlphaGo represents not only a scientific and technological advancement, but also a milestone in human understanding of Go. Unconstrained by human biases, and free to experiment with radical new approaches, AlphaGo has demonstrated great open-mindedness and invigorated the game with creative new strategies. (Hui, 2016, p. 1)

AlphaGo invented Go strategies that no human had considered—and presumably would not have predicted.

Much simpler examples illustrate similar phenomena. Consider a database system. Presented with a query about information that is not in its database, the system will not provide a useful answer. But once that information is stored in the database, the same query will return a more useful answer.

The point is that most entities have internal states that play a role in the entity’s decision-making processes. Those internal states may change. Often those changes occur as a result of interacting with the external world. To know about them all, an observer would need complete information about the entity’s history, which generally is not available.

Autonomy consists of the ability of an agent to apply its decision-making apparatus to symbolic information and then to use the result to control how it behaves. Since most agents’ decision-making processes depend in part on the agent’s internal state, and since an agent’s internal state generally depends on its history, which is virtually impossible for an outside observer to know in detail, no outside observer can be confident that it will be able to forecast how an autonomous agent will behave.

Implications for Free Will

I will end by examining some implications for free will. Following is a paraphrased extract from physicist Sabine Hossenfelder (2016):

> According to our best present understanding, everything that happens in our universe is due to four forces: gravity, electromagnetism, and the strong and weak nuclear force. These forces conform to one of two types of laws. One type is deterministic, which means that the past entirely predicts the future. [This is the Laplacian universe mentioned earlier.] There is no free will in such a law because there is no freedom. (p. 1)

> The other type, which appears in quantum mechanics, has a random component. But the randomness cannot be influenced by anything. There is no free will in such a law either—just some randomness sprinkled over the determinism.

> These forces and the laws that characterize their physical effects have been extremely well studied; they leave no room for free will. (p. 1)

Dispiriting as this may sound, Hossenfelder goes on to say that these facts *do not preclude one from making choices or decisions*. I will explain this in terms of the framework developed above.

People Interact With the World in Two Ways: Physically and Symbolically

- We interact physically when we experience and are affected by physical properties such as gravity, momentum, temperature, and so forth
- We interact symbolically when we read, speak, listen to others, or identify things we see, hear, touch, and so forth

It is the destiny of autonomous agents always to be mediating between these two forms of interaction. Recall the automobile driver. Her interaction with the world involved

a) the (physical) photons that entered her eyes;

b) the (symbolic) interpretation of those photons as a red or green traffic light;

c) the (physical) forces to which she and her car are subjected;

d) the (symbolic) decision regarding how to respond to the traffic light; and

e) the (physical) act of pressing the brake or the accelerator.

The two levels operate with distinctly different rules. The physical level operates according to the laws of physics. The symbolic level operates according to internal information processing rules.

The symbolic level operates according to rules constrained in part by logic and computer science. To take a classic example, Turing (1937) proved that it is not possible to determine for an arbitrary computer program whether it will ever stop. That question, known as the halting problem, has nothing to do with the laws of physics. Yet symbol processing as performed by both human beings and computers is constrained by that result.

The broader conclusion is that laws apply at the symbolic level that to not appear at the physics level. Here is how two Nobel prize winners put it:

> [L]iving matter, while not eluding the "laws of physics" . . . is likely to involve "other laws," [which] will form just as integral a part of [its] science. – Schrödinger (1944/2012), p. 68

> [The] workings of all the animate and inanimate matter of which we have any detailed knowledge are . . . controlled by the . . . fundamental laws [of physics]. . . . [W]e must all start with reductionism, which I fully accept. [Even so, the] behavior of large and complex aggregates of elementary particles is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear. – Anderson (1972), p. 393.

For additional discussion, see Abbott (in press).

Such a two-level perspective has been noted by a number of philosophers. For example, Dennett (2003) writes (emphasis added),

> Nonhuman animals can engage in *voluntary* actions of sorts. The bird that flies wherever it *wants* is *voluntarily* wheeling this way and that, *voluntarily* moving its wings. (p. 48)

The aspects of a bird's actions that Dennett refers to as voluntary correspond to actions I would describe as being directed by a bird's decision-making processes. I would describe birds as being autonomous in that their behaviors derive from internal rules.

Neither Anderson, Dennett, Hossenfelder, nor Schrödinger would deny that the internal decision-making processes also depend on the forces of physics. After all they are physical activities and hence subject to the laws of physics. Yet both Dennett and Hossenfelder find it useful to distinguish the direct forces of physics from those that run an agent's decision-making processes. (Neither Anderson nor Schrödinger discuss this issue.)

These two manifestations of the forces of physics produce different sorts of effects. Some forces of physics effect agents directly in what I have referred to as a Laplacian manner. Some forces of physics run agents' decision-making processes and produce their effects indirectly as a result of symbolic processing. So even though both manifestations obey the same physical laws, they contribute to an agent's behavior very differently.

As a simple example consider a railroad train. It is propelled forward by energy generated by burning fuel. Yet its direction of motion may be determined by a switch that controls which track it will take. The energy required to control the switch is miniscule compared to the energy that powers the forward motion. Both forms of energy obey the laws of physics; they affect the train quite differently.

What Is Free Will and Where Does It Fit In?

Dennett does not consider what he calls the voluntary actions of birds to reflect free will. For him, autonomy appears not to be the same as free will. But Dennett does talk about free will.

> Humans differ from every other species in that we represent our reasons to ourselves and others. This is what gives us the power, and the obligation, to think ahead, to anticipate, to see the consequences of our actions, to be able to evaluate those actions in the light of what other people tell us, and to share our wisdom with each other. That's what makes us free in a way that no bird is free. . . . We have added a layer on top of the bird's (and the ape's and the dolphin's) capacity to decide what to do next. . . . Your dog can be "asked" to do a variety of voluntary things, but it cannot ask why you make that request. A male baboon can "ask" a nearby female for some grooming, but neither of them can discuss the likely outcome of compliance with this request, which might have serious consequences . . . if the male is not the alpha male of the troop. (p. 48)

For Dennett, free will consists of the ability to think ahead before acting.

> [Evolution produced] creatures capable of considering different courses of action in advance of committing to any one of them, and weighting them on the basis of some projection of the probable outcome of each. In the quest by brains to produce a useful future, this

> is a major improvement over the risky business of *blind* trial and error, since, as Karl Popper (1978) once put it, it permits some of your hypotheses to die in your stead. (p. 45)

The ability to anticipate the consequences of actions and then to make a choice based on that analysis offers significant advantages. This, Dennett says, is the only kind of free will worth having. Dennett does not say whether he would argue that computing systems such as AlphaGo, which also select actions after analyzing multiple possibilities, also have free will. Computers have played games using what are called look-ahead strategies for decades. I know of no one who has attributed free will to them.

Baumeister (Baumeister, 2014; Baumeister & Masicampo, 2010; Baumeister & Monroe, 2014) adopts a similar approach. Free will is the ability to think things through before acting—especially about things having to do with one's social situation:

> Culture includes systems that require people to follow rules, and so a capacity for self-regulation according to rules would be very useful. Culture is based on information, so the ability to communicate and alter behavior on the basis of communicated information is important. Rational choice would be highly useful to deduce implications for action from abstract guidelines such as laws and moral principles, as well as for functioning as an economic agent in a marketplace (another vital aspect of culture). In both cases, and perhaps more generally, the ability to base actions on ideas is central to being a civilized person. Developing the capacity to control action in this way presumably comprised the key steps in the evolution of free will. (Baumeister & Monroe, 2014, p. 12)

Even If We Have Free Will, Why Bother to Exercise It?

I am not convinced that the preceding adequately characterizes free will. But assuming that free will has to do with the ability to think things through before acting, the question arises as to why one should bother. After all, one's thought processes are determined by the forces of nature. (They are performed by physical materials, which must follow the laws of nature.) Therefore the outcome of those thought processes are predetermined. Why spend the time and energy to think?

Hossenfelder's answer is that thinking is the way to answer a question for which one does not yet have an answer. Suppose that in thinking through how it should respond to some situation, an agent needs to perform some calculation. Even though the answer is predetermined, the agent must still do the calculation to determine the answer. Searle (2001) offers the following scenario:

> Imagine that you are in a restaurant and you are given a choice between veal and pork, and you have to make up your mind. You cannot refuse to exercise free will. [It makes no sense to say] "Look, I am a determinist—*que sera sera*—I'll just wait and see what I order." . . . We cannot think away our free will. (p. 494)

Long ago, Augustine of Hippo (391–395) raised a similar question. Why is one responsible for one's choices even though God has foreknowledge of one's choices? His answer: Everyone has the power to make his own decisions; that is no less true even though God knows what one will decide.

The essence of this argument is that even if one agrees that the laws of nature predetermine the outcome of any thought process, one cannot refuse to do the thinking—if the possibility of refusing to do something is even coherent in the context of determinism.

Deterrence and Free Will

The discussion so far provides a rationale for implementing deterrence against criminal behavior, whether or not agents have free will.

The traditional argument is that if agents have free will, deterrence can be effective. That argument holds even for the sort of free will espoused by Dennett and Baumeister—the ability to think ahead before acting. An agent that can anticipate negative consequences can refrain from taking actions that produce those consequences.

But, the argument also goes, agents without free will are not responsible for their behavior. For them, deterrence will be ineffective. That is not the case. As we saw, agents may be reprogrammed by their experiences: An agent that sees lawbreaking being punished may reprogram itself not to break the law. That is the case even if the act of self-reprogramming does not rely on free will.

AlphaGo illustrates this phenomenon. No one would claim that computer systems have free will—other, perhaps, than the ability to think ahead. Yet AlphaGo reprograms itself on the basis of what are effectively rewards and punishments based on its actions. The result is that it reprograms itself to act "properly," that is, to play winning moves and to win games. Since deterrence works with computer systems, there is no reason to believe it will not work with biological systems, free will or not.

Consciousness

None of the preceding mentioned consciousness. No need was found to distinguish between biological and computer-based agents. This is not to argue either (a) that computers have what most people would regard as consciousness or (b) that there is no such thing as consciousness. I strongly encourage efforts to develop a scientific understanding of subjective experience—which I take to be synonymous with consciousness. I find myself concluding, however, that meaning and autonomy do not require consciousness. Yet I do not dismiss the relevance of consciousness for meaning. Consciousness makes meaning come alive, adding to it awareness, depth, richness, and intensity, both conceptual and emotional.

In my view, free will requires consciousness in two ways:

a) Like qualia, free will is a subjective experience.

b) It is not clear what it would mean to attribute free will to an agent without consciousness.

Searle (2010) characterizes free will as a response to experiencing three causal gaps: between reasoning and deciding, between deciding and initiating an action, and between initiating an action and (if the action spans an extended period) seeing it through. But, as Searle says, these are gaps in subjective experience, that is, consciousness. I agree that understanding free will as commonly understood will depend on understanding consciousness, which seems to me the greater mystery.

Summary

To summarize, I would emphasize the following:

- Meaning involves the transformation of symbols, that is, abstractions, into physical actions.
- Although computers and biological organisms do this regularly, such transformations are quite extraordinary. Symbols are nonmaterial, not subject to the laws of physics. Yet agents' physical behaviors depend on their symbol-processing rules.
- Because these rules are internal, I call such agents autonomous.
- Because these rules are often inaccessible to outside observers and because they may depend on the agent's history, it may not be feasible to predict an agent's behavior.

By acting physically in response to symbols, autonomous agents link the abstract to the material.

References

Abbott, R. (in press). A software-inspired constructive view of nature. In Berkich, Don (Ed.) *Computing and philosophy: Selected papers from IACAP 2016*. Preprint available at https://drive.google.com/file/d/0B-I58s-_d3o5V2UzNG5JZlczOTA/view?usp=sharing

Anderson, P. W. (1972). More is different. *Science, 177,* 393–396. http://dx.doi.org/10.1126/science.177.4047.393

Augustine of Hippo (391–395) *On the free choice of will*.

Balaguer, M. (2016). Platonism in metaphysics. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Baumeister, R. F. (2014). Constructing a scientific theory of free will. In W. Sinnott-Armstrong (Ed.), *Moral psychology (Vol. 4)*: *Free will and responsibility* (pp. 235–255). Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/9780262026680.003.0007

Baumeister, R. F., & Masicampo, E. J. (2010). Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal-culture interface. *Psychological Review, 117,* 945–971. http://dx.doi.org/10.1037/a0019393

Baumeister, R. F., & Monroe, A. E. (2014). Recent research on free will: Conceptualizations, beliefs, and processes. In M. P. Zanna & J. M. Olson (Eds.), *Advances in experimental social psychology* (Vol. 50, pp. 1–52). Waltham, MA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-800284-1.00001-1

Bricker, P. (2016). Ontological commitment. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Cheney, D. L., & Seyfarth, R. M. (1990). *How monkeys see the world*. Chicago, IL: University of Chicago Press.

Dennett, D. C. (2003). The self as a responding—and responsible—artifact. *Annals of the New York Academy of Sciences, 1001,* 39–50. http://dx.doi.org/10.1196/annals.1279.003

Ditz, H. M., & Nieder, A. (2015). Neurons selective to the number of visual items in the corvid songbird endbrain. *PNAS, 112,* 7827–7832. http://dx.doi.org/10.1073/pnas.1504245112

Dowe, P. (2000). *Physical causation*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511570650

Fernández, M. (2014). *Programming languages and operational semantics*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4471-6368-8

Flock, E. (2012, February 17). Dagen H: The day Sweden switched sides of the road. *Washington Post*.

Floridi, L. (2017). Semantic conceptions of information. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Gómez, J. C. (2005). Species comparative studies and cognitive development. *Trends in Cognitive Sciences, 9,* 118–125. http://dx.doi.org/10.1016/j.tics.2005.01.004

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42,* 335–346. http://dx.doi.org/10.1016/0167-2789(90)90087-6

Hossenfelder, S. (2016, January 10). Free will is dead, let's bury it. *BackReAction*. http://backreaction.blogspot.com/2016/01/free-will-is-dead-lets-bury-it.html

Hu, B., & Tu, Y. (2014, June). Behaviors and strategies of bacterial navigation in chemical and nonchemical gradients. *PLOS Computational Biology, 10,* e1003672. http://dx.doi.org/10.1371/journal.pcbi.1003672

Hui, F. (2016, September, 5). *AlphaGo games—English*. https://deepmind.com/research/alphago/alphago-games-english

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology, 3,* 318–356. http://dx.doi.org/10.1016/S0022-2836(61)80072-7

Keller, H. (1903). *The Story of my life*. New York, NY: Doubleday, Page & Co.

Koch, C. (2016, March 19). How the computer beat the Go master. *Scientific American*.

Laplace, P. S. (1951). *A Philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Trans). Mineola, NY: Dover. (Original work published 1814)

Loukola, O. J., Perry, C. J., Coscos, L., & Chittka, L. (2017). Bumblebees show cognitive flexibility by improving on an observed complex behavior. *Science, 355,* 833–836. http://dx.doi.org/10.1126/science.aag2360

Moyer, C. (2016, March 28). How Google's Alphago beat a Go world champion. *The Atlantic*.

Nelson, M. (2016). Existence. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM, 19,* 113–126. http://dx.doi.org/10.1145/360018.360022

Orilia, F., & Swoyer, C. (2016). Properties, In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Pearl, J. (2000). *Causality*. New York, NY: Cambridge University Press.

Pepperberg, I. M. (1999). *The Alex studies: Cognitive and communicative abilities of grey parrots*. Cambridge, MA: Harvard University Press.

Pepperberg, I. M. (2002). Cognitive and communicative abilities of grey parrots. *Current Directions in Psychological Science, 11,* 83–87. http://dx.doi.org/10.1111/1467-8721.00174

Popper, K. (1978). Natural selection and the emergence of mind. *Dialectica, 32,* 339–355. http://dx.doi.org/10.1111/j.1746-8361.1978.tb01321.x

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience, 13,* 587–597.

Quiroga, R. Q., Fried, I., & Koch, C. (2013). Brain cells for grandmother. *Scientific American, 308,* 30–35. http://dx.doi.org/10.1038/scientificamerican0213-30

Reicher, M. (2016). Nonexistent objects. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Robb, D., & Heil, J. (2014). Mental causation. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Roberts, W. A., & Macpherson, K. (2016). Of dogs and men. *Current Directions in Psychological Science, 25,* 313–321. http://dx.doi.org/10.1177/0963721416665007

Rosen, G. (2014). Abstract objects. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford, CA: Stanford University, Center for the Study of Language and Information.

Rumbaugh, D. M. (1977). *Language learning by a chimpanzee*. Waltham, MA: Academic Press.

Schrödinger, E. (2012). *What is life?* New York, NY: Cambridge University press. (Original work published 1944).

Searle, J. R. (2001). Free will as a problem in neurobiology. *Philosophy, 76,* 491–514.

Searle, J. R. (2010). Consciousness and the problem of free will. In R. F. Baumeister, A. R. Mele, & K. D. Vohs (Eds.), *Free will and consciousness: How might they work?* (pp. 121–134). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195389760.003.0008

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529,* 484–489. http://dx.doi.org/10.1038/nature16961

Smith, J. D., Zakrzewski, A. C., Johnson, J. M., Valleau, J. C., & Church, B. A. (2016). Categorization: The view from animal cognition. *Behavioral Sciences, 6,* 12. http://dx.doi.org/10.3390/bs6020012

Turing, A. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, s2–42,* 230–265. http://dx.doi.org/10.1112/plms/s2-42.1.230

Woodward, J. (2003). *Making things happen: A theory of causal explanation* New York, NY: Oxford University Press.

Zentall, T. R., & Pattison, K. F. (2016). Now you see it, now you don't: Object permanence in dogs. *Current Directions in Psychological Science, 25,* 357–362. http://dx.doi.org/10.1177/0963721416664861

Received December 8, 2016
Revision received June 15, 2017
Accepted July 27, 2017 ■