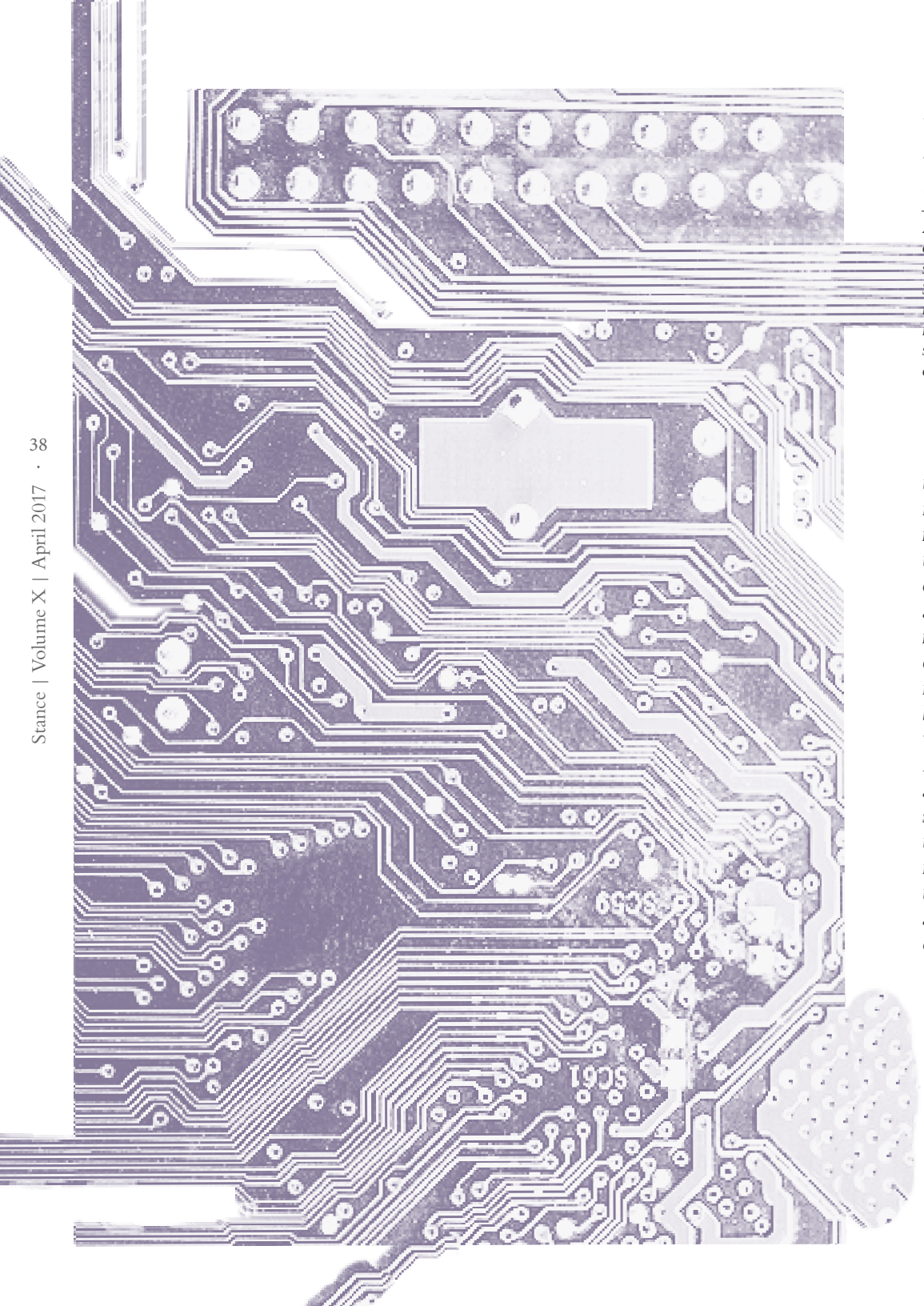

In Defense of Strong AI:

Semantics as Second-Order Rules

Corey Baron

Abstract: *This paper argues against John Searle in defense of the potential for computers to understand language ("Strong AI") by showing that semantic meaning is itself a second-order system of rules that connects symbols and syntax with extralinguistic facts. Searle's Chinese Room Argument is contested on theoretical and practical grounds by identifying two problems in the thought experiment, and evidence about "machine learning" is used to demonstrate that computers are already capable of learning to form true observation sentences in the same way humans do. Finally, sarcasm is used as an example to extend the argument to more complex uses of language*

Introduction

"While Stance *strives to evaluate every paper impartially, for every reviewer, there are topics that are particularly enjoyable and those that are especially difficult to read. For me, philosophy of language and artificial intelligence have always been topics of the latter variety. In spite of this, I found this paper to be among the most engaging submissions we received this year, and it has inspired me to revisit topics I've previously avoided with renewed curiosity."*

-Daniel Klinestiver
Associate Editor

In his 1984 article "Can Computers Think?," philosopher John Searle attempts to refute the hypothesis of "strong artificial intelligence" ("strong AI"), which holds that "it's only a matter of time until computer scientists and workers in artificial intelligence design the appropriate hardware and programs which will be the equivalent of human brains and minds."[1] Searle says that this position "admit[s] of a simple and decisive refutation"[2] and sets out to disprove it with an exaggerated sense of ease by attempting to disprove the possibility that computers would ever be able to understand informal languages. The majority of his article is devoted to the "Chinese Room Argument" (CRA), a thought experiment that allegedly proves that, in the context of linguistic abilities, there is no bridge from formal syntactical *rules* to semantic *meaning*.

In this paper, I aim to argue against Searle, in defense of strong artificial intelligence and the potential for computers to understand language, by showing that semantic meaning is itself a second-order system of rules that connects symbols and syntax with extralinguistic facts. After explaining Searle's argument, I

1 John Searle, "Can Computers Think?", in *Analytic Philosophy: An Anthology*, ed. A.P. Martinich and David Sosa (Singapore: Wiley-Blackwell, 2012), 317.

2 Ibid., 318.

contest the Chinese Room Argument on theoretical and practical grounds by identifying two problems in the argument that I will call "the problem of isolation" and "the problem of the complete rulebook." I use empirical evidence about "machine learning" to demonstrate that computers are already capable of learning to form true observation sentences in the same way humans do. Finally, I explain how my argument can be extended to cases of more complex language usage, using sarcasm as an example. Because it is rule-based and learned in context, I argue for the theoretical possibility of a computer capable of understanding and using sarcasm.

I. Searle's Refutation of Strong AI

It is first important to recognize that Searle's refutation of strong AI rests on the notion that the ability to use and understand language, both syntactically and semantically, is central to "intelligence." For the purposes of this paper, I will accept this foundational assumption and attempt to contest Searle within this framework. The basic premise of Searle's argument is as follows: "[t]here is more to having a mind than having formal or syntactical processes,"[3] since "the mind has more than a syntax, it has a semantics."[4] A computer can move and make use of symbols without actually *understanding* their meaning, because "the symbols have no meaning; they have no semantic content; they are not about anything."[5] In other words, since computer operations "can be specified purely formally,"[6] a computer can never duplicate a *mind* in its ability to use and understand informal language.

Searle demonstrates his argument with his famous "Chinese Room" thought experiment. In this hypothetical, "you are locked in a room, and in this room are several baskets full of Chinese symbols."[7] You, an English speaker who knows no Chinese, are provided with a set of instructions for how to make use of the Chinese characters. "The rules specify the manipulations of the symbols purely formally, in terms of their syntax. So the rule might say: 'Take a squiggle-squiggle sign out of basket number one and put it next to a squoggle-squoggle sign from basket number two.'"[8] Chinese symbols come into the room, and, according to the rulebook, you gather a different set of symbols and send them out. Without your knowing it, the symbols coming in are questions and the sets you are sending out are coherent answers. Searle points out that even though "your answers are indistinguishable from those of a native Chinese speaker," you do not *understand* Chinese; "you behave exactly as if you understood Chinese, but all the same you don't understand a word of Chinese."[9]

Searle claims that this example is perfectly analogous to the way in which a programmed computer would be able to make use of language. Even if it were programmed with rules of manipulation complex enough to fool someone into believing they were speaking with another human being, the computer could never *actually understand* language. The symbols would always remain "meaningless" to it. Searle claims to definitively conclude the argument in saying, "As long as all I have is a formal computer program, I have no way of attaching any meaning to any of the symbols."[10]

II. Two Flaws in the Chinese Room Argument

It is certainly true that the CRA holds within its own logical framework. The problem is twofold: the argument is set up in such a way as to predetermine its own success, and recent advances in computer science also allow the argument to be questioned on empirical grounds. Searle's theory suffers from two fundamental flaws: the "problem of isolation" and the "problem of the complete rulebook."

The Problem of Isolation

To start, the "problem of isolation" can be seen as a *theoretical flaw* in the CRA. Searle sets up his analogy in accordance with the idea that computers do not have a means to connect linguistic symbols with extralinguistic facts and then claims victory when he "proves" that computers cannot connect linguistic symbols and extralinguistic facts. The outcome is already contained in the setup, and the logic is therefore circular. When Searle states, "As long as all I have is a formal computer program, I have no way of attaching any meaning to any of the symbols,"[11] he is really only saying, "As long as all I have is a formal computer program *with no access to the external world or the effects of its output*, I have no way of attaching meaning to any of the symbols."

Basically, the problem of isolation comes from the fact that the person inside the Chinese Room is entirely cut off from the external world and has no way to draw connections between the symbols they use and any external phenomena. This is a dramatic departure from human language acquisition, which always occurs *in the world, in context,* and *in relation to extralinguistic facts*.

In "Epistemology Naturalized," W.V. Quine examines "observation sentences," those statements that are typically conceived of as the basic units of propositional language and whose truth are always assessed in relation to an external world. Quine points out that observation sentences "are precisely the ones that we can correlate with observable circumstances of the occasion of utterance

3 Ibid.
4 Ibid., 319.
5 Ibid., 318.
6 Ibid.
7 Ibid., 319.
8 Ibid.
9 Ibid., 318.
10 Ibid., 320.
11 Ibid.

or assent. … They afford the only entry to a language."[12] He goes on to say that "the observation sentence is the cornerstone of semantics. For it is … fundamental to the learning of meaning."[13] The central point is that there is a deep and necessary connection between language and the world. Quine puts it more simply in "Two Dogmas of Empiricism" when he says, "Truth in general depends on both language and extralinguistic fact."[14] Even without exploring his epistemological use of the word "truth," it is evident that *meaning* must come from some sort of connection between the phonetic/visual symbols of language and our external circumstances.

For this reason, "semantic meaning" can actually be seen as a sort of set of "second-order" rules: one that relates the symbols, phonetics, and syntactical rules of language to the "facts" of an extralinguistic world. "Understanding" comes from the ability to recognize and utilize this connection, so the symbols of language will obviously remain meaningless to the speaker if they are deprived of their association with the world. The *theoretical flaw* (the problem of isolation) in the CRA is its claim that there must be a bridge from syntax to meaning *that does not include access to extralinguistic facts*.

Solving this problem is easy—the Chinese Room must be opened to the world. If the person in the Chinese Room is allowed to see that when they pass out the "squiggle" symbol, and the person with whom they are "communicating" looks at the table, then they can start to infer something about the *meaning* of the squiggle symbol. Given enough time, practice, and examples, the person in the room will start to actually learn and even *understand* Chinese by seeing the results of their use of symbols and communicative acts.

So as not merely to attack a straw man version of Searle's position, it should be made clear that his argument does not necessarily deny the possibility of a computer *interacting with* or *making use of* extralinguistic facts. Instead, Searle simply points out that any input of extralinguistic facts first requires that those facts be translated into the linguistic system of the computer—that they are ultimately represented by the ones and zeroes of binary coding. This is to say that extralinguistic facts must be *made* linguistic to be accessible to the computer at all. While this seems like a strong counterargument to the position I have laid out, it is important to note that the same challenge could be raised in regard to *human* language use and acquisition. While a neurological examination of the human brain is beyond of the scope of this paper, all that needs to be acknowledged is the fact that the human brain physically operates as a series of electrical impulses conveyed between neurons. This is not significantly different than the physical operation of a computer, in which electrical impulses within the circuitry are dictated by binary coding. This similarity suggests that Searle's argument about the inaccessibility of extralinguistic facts applies to humans as much as to computers. If this is the case, Searle cannot deny understanding to computers on these grounds without simultaneously denying our own ability to "understand" the connection between the external world and the mental experiences of language and perception within our brains.

This discussion also raises an important epistemological question with regard to Searle's thought experiment: What is proper *evidence* of semantic understanding? A closer examination reveals that the only available evidence is correct usage. When human toddlers watch YouTube videos with their mom and hear her say "cat" enough times when the whiskered furry creature appears, they learn that the word "cat" is used in the presence of such a creature. Eventually, they will also be able to say "cat" when similarly whiskered and furry creatures appear on the screen. Any external observer will agree that the child "understands" the concept "cat," and this understanding is *demonstrated* through correct usage. In the case of observation sentences, correct usage is that which accurately relates language and the world. There is no other evidence to look to for proof of understanding. The CRA shows why this is problematic: correct usage *might* be evidence for understanding of semantics, or it might just be evidence that the "speaker" can blindly follow syntactical rules of symbol manipulation. There is no solid way to prove which is the case in a given instance. This is why we must look to the *way* the language user comes to their ability to use language in order to determine whether or not semantic understanding is present, which brings us to the second problem of the CRA.

The Problem of the Complete Rulebook

Since correct usage alone cannot prove that there is semantic understanding instead of mere knowledge of syntactical rules, it is necessary to look at *how* language use is acquired when trying to distinguish the two possibilities. The practical flaw of the original CRA is Searle's claim about how computers must acquire their ability to "use" language. The argument assumes that computer programming must enumerate all of the syntactical rules of symbol manipulation ahead of time. This claim is reflected in the completeness of the rulebook provided to the subject in the thought experiment because it apparently already contains every possible syntactical rule for communicating in Chinese. Even if the subject in the Chinese Room had access to an outside world, it would not make sense for them to look up from the instructions in order to make semantic connections since they already have every rule they will ever need.

I am not the first to offer this critique of the original CRA, and Searle has since slightly amended this aspect of his thought experiment in response to his challengers. The revised argument allows additional external inputs

12 W.V. Quine, "Epistemology Naturalized," in *Analytic Philosophy: An Anthology*, ed. A.P. Martinich and David Sosa (Singapore: Wiley-Blackwell, 2012), 249.

13 Ibid., 249.

14 W.V. Quine, "Two Dogmas of Empiricism," in *Analytic Philosophy: An Anthology*, ed. A.P. Martinich and David Sosa (Singapore: Wiley-Blackwell, 2012), 525.

to flow into the Chinese Room. Searle maintains his view that even if this is the case, "additional syntactic inputs will do nothing to allow the man to associate meanings with the Chinese characters. It is just more work for the man in the room."[15] In so saying, Searle does not recognize that by participating in that extra "work" of editing or augmenting the rulebook in response to additional inputs, the person in the Chinese Room *is actually learning to use and understand language* by following the same process that occurs naturally in human language acquisition in general.

Empirical evidence for the ability of a computer to successfully augment its own "rulebook" can be found in the recently emerged field of "machine learning," in which computers are not explicitly programmed but are instead "taught" through exposure to a vast number of examples. With minimal guidance, these computers essentially code themselves by finding rules and heuristics that efficiently interpret the data to meet the demands of the task at hand. Examples of these tasks now include the operation of self-driving motor vehicles, the detection of fraud, and many other complicated undertakings.[16]

In the specific context of linguistic tasks, the incompleteness of a program's coding forces it to use "sense" data from examples it is provided (known as the "training set") to figure out how to interpret and label the external world. In a basic sense, a computer is provided with a label that unites the examples in the training set under a given category (for instance, "human faces") in addition to the examples that comprise the training set itself. The computer then codes itself in the way that best captures the fundamental characteristics of the category, and it is typically designed to do so in such a way as to be able to successfully identify other examples not included in the original training set.[17] This is the process of learning to relate linguistic symbols with extralinguistic phenomenon—the basic process of informal language acquisition. Through machine learning, we now have computer programs that can recognize cats, human faces, human bodies, and much more.[18]

When looked at objectively, computer programs that can accurately identify cats have basically learned to form "true" observation sentences, much like the toddler in the example above. If we grant the toddler credit for "understanding" but deny the computer the same, it is a result of a simple prejudice. We want to believe that there is something irreplaceably human in our language use and consistently resist the urge to admit that language is a system of rules, even if that systematic nature is what enables communication in the first place. This process of denial is aided by the fact that we acquire the ability to use our native language when we are too young to remember or consciously experience the learning process and spend very little time thinking seriously about the nature of the rules that structure our everyday speech. But our ignorance about these rules is not a valid argument against their existence. And the fact that machines *can* learn to form accurate observation sentences should be taken as compelling evidence that semantic meaning, like syntax, is a rule-based phenomenon. The difference is that while syntactical rules govern the relations between linguistic symbols, semantic rules govern the connections between those symbols and extralinguistic facts.

15 David Cole, "The Chinese Room Argument," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Winter 2015), https://plato.stanford.edu/archives/win2015/entries/chinese-room/.

16 "Machine Learning: What It Is and Why It Matters," SAS, accessed February 8, 2017, http://www.sas.com/en_us/insights/analytics/machine-learning.html.

17 Jason Tanz, "Soon We Won't Program Computers. We'll Train Them Like Dogs," *Wired*, May 17, 2016, https://www.wired.com/2016/05/the-end-of-code/.

18 Liat Clark, "Google's Artificial Brain Learns to Find Cat Videos," *Wired UK*, June 26, 2012, https://www.wired.com/2012/06/google-x-neural-network/.

Moving Beyond Observation Sentences

Searle or a supporter of the CRA might respond with the counterargument that observation sentences are an extremely basic unit of language and claim that I have not yet shown that computers will ever understand the sophistication of ordinary language. Beyond the simplicity of observation sentences, ordinary language is full of emotion, misdirection, non-literal usage, and non-propositional speech. In this section, I will defend against this counterclaim, using sarcasm as an example of complex human language use. Sarcasm is a good example for this purpose since it seems so intuitively human—it tends to come with emotional subtext; it can be emotional, deceitful, and humorous; and, it even relies upon the deliberate *rejection* of linguistic conventions. It certainly does not feel rule-based when we exclaim, "Well, *duh*!" or layer on a flippant tone to say, "That sounds like a *brilliant* idea" and mean the exact opposite. Nevertheless, I will show that because sarcasm is a rule-based phenomenon, computers *do* have the theoretical potential to learn and understand it.

The first thing to note is that we *do* use sarcasm successfully, and this comprehensibility alone suggests that there are rules of use at play. Philosopher of language Elisabeth Camp confirms this suspicion in "Sarcasm, Pretense, and The Semantics/Pragmatics Distinction." Camp shows how uses of sarcasm do conform to a general set of rules and patterns. Across its varied forms, sarcasm is always speech which presupposes a *normative scale*—which *pretends* to undertake (or at least, evokes) a commitment with respect to this scale—and which thereby communicates an *inversion* of this pretended (evoked) commitment.[19] In elaborating these rules of use, Camp argues for an expanded, four-level definition of meaning.[20] The exact details of that definition are interesting but unimportant to my argument here. What is crucial is the way Camp points out how her definition of meaning, while

19 Elisabeth Camp, "Sarcasm, Pretense, and The Semantics/Pragmatics Distinction," *NOÛS* 46, no.4 (2012): 605.

20 Ibid., 623.

more complex than prior theories, more accurately captures the nuances of ordinary language use: "Speakers and hearers regularly display their implicit sensitivity to all four levels of meaning in the course of ordinary conversation."[21]

There are two lessons to take from this. First, just because the rules of language and meaning are complicated does not mean they do not exist. Similarly, the fact that we do not have to *consciously* consult the rules of language in order to speak is not evidence that they are not there. Instead, we are constantly making use of sophisticated semantic "rules of meaning" that we do not consciously know or understand. So how could we have "learned" these rules in the first place?

The process of language acquisition in humans begins very quickly after birth and always occurs as a process of relating the linguistic symbols of communicative acts (e.g., phonetics/sounds) with the external world. In "Playing With Expectations," developmental psychologist Gabriella Airenti argues that "children acquire communicative acts simultaneously with the conditions of their use"[22] and that this applies even in cases of complex communicative acts like irony and sarcasm. Children connect the phonetic forms of language with the extralinguistic facts of their world because making those connections "works." Developing the ability to associate "milk" with the nourishing white liquid "works" to help children express desires; humorous communicative acts "work" to help children feel (and cause) joy, as can be seen in examples of children laughing after engaging in misnaming or teasing behaviors.[23]

Children begin with a rudimentary and incomplete syntactical rulebook, and are never isolated from the external world. They tap into the semantic rules of meaning that govern sarcasm and humor by using, experiencing, and experimenting with language *in a context*. There is no reason to deny the possibility that computers could do the same given the same conditions for learning. Provided enough processing power, programming that rewards certain kinds of linguistic "success," and a sophisticated enough ability to "sense" the world, a computer would be able to understand the rules of use for sarcasm over time and demonstrate that understanding through correct usage.

It is true that I have extended somewhat beyond the bounds of the CRA, and building the type of computer I am suggesting necessitates a much more sophisticated understanding of the psychological and neurological "rules" of motivation in the human brain than we currently possess. Nevertheless, the basic ingredients are present to assume that computers *could* learn to make proper use of sarcasm. Sarcasm is based in a system of rules, and it is learned through a relation of the communicative act and the external world. If a computer could use sarcasm properly, in the same kinds of circumstances and for the same kinds of purposes as humans, there would be no compelling reason to say that it does not *understand* what it is doing. It would understand its own behavior to the same extent that we "understand" *ours* every time we roll our eyes and say, "Yeah, right."

Conclusion

In this paper, I argued in defense of the strong artificial intelligence position denounced by John Searle in "Can Computers Think?" by demonstrating that semantic meaning can be seen as a set of "second-order" rules that relate language to extralinguistic facts. In examining machine learning, I identified empirical evidence that computers can acquire the most basic language abilities (formation of true observation sentences) by learning to connect linguistic symbols to the facts of the world. Finally, I examined sarcasm to show the theoretical possibility that computers can learn to understand sophisticated uses of language as well.

The fundamental takeaway is that even though semantic understanding is largely unconscious in human language use, it is actually an elaborate system of second-order *rules* for connecting linguistic symbols and communicative acts with the extralinguistic facts of the world. Machine learning has the potential to duplicate the process through which humans come to use and understand these rules. If a computer were to learn language by the same means and use it in the same manner and circumstances as a human, there would be no reason to deny it credit for its understanding.

21 Ibid., 624.

22 Gabriella Airenti, "Playing with Expectations: A Contextual View of Humor Development," *Frontiers in Psychology* 7, no. 1392 (September 2016), doi: 10.3389/fpsyg.2016.01392.

23 Ibid., 7.

About Corey Baron

Corey Baron is a senior at Colorado College in Colorado Springs, Colorado, and will be graduating in May 2017 with a major in philosophy and a minor in Spanish language and literature. Outside of the classroom, she tries to balance her time between an overflowing bookshelf, beautiful mountain trails, and the ultimate frisbee field.