A CONTRIBUTION TO THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE AND CHATBOT COMMUNICATION

A Contribution to The Philosophy of Artificial Intelligence and Chatbot Communication

By Siddharth Raman, HBA

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Master of Arts

McMaster University MASTER OF ARTS (2020) Hamilton, Ontario (Philosophy)

TITLE: A Contribution to The Philosophy of Artificial Intelligence and Chatbot Communication

AUTHOR: Siddharth Raman, HBA (University of Toronto)

SUPERVISOR: Professor Sandra Lapointe

NUMBER OF PAGES: i – vi; 76.

LAY ABSTRACT:
The purpose of this thesis is to explore the question why chatbot computer programs are often not very good at communicating in human natural language. It is argued that one possible reason why chatbots are often not good conversationalists is because they model communication in terms of only encoding and decoding processes. Human communication, however, involves making inferences about the mental states of others.

ABSTRACT:
The purpose of this thesis is to attempt to answer the question why chatbot computer programs are often not very good at communicating in human natural language. It is argued that one possible reason why chatbots are often not good conversationalists is because they model communication in terms of only encoding and decoding processes. Human communication, however, involves making inferences about the mental states of others.

Chapter one begins by exploring a popular theory about how communication works called the code model of communication. The code model describes human communication as having to do with speakers encoding thoughts into utterances and listeners decoding utterances to recover a representation of the thought that the speaker wanted to communicate. A variation on the code model is also explained; this is referred to as the information-theoretic code model. Two arguments against the code model will then be presented. Finally, an alternative to the code model is considered, called the ostensive-inferential model of communication.

Chapter two begins with an explanation of how chatbots work. Chatbots are made up of several different components. The language model component provides chatbots with the ability to produce and interpret utterances. Next, an explanation of how language models work, and how chatbots can represent the meanings of words is provided. The chapter concludes by documenting the fact that chatbots communicate using only encoding and decoding processes – that is, that chatbots communicate within the paradigm of the code model of communication.

Chapter three explains how the fact that chatbots communicate using only encoding and decoding processes can help explain why chatbots often cannot communicate effectively in human natural language. The poor conversational abilities of chatbots are a result of the fact that chatbots only access linguistic context, whereas listeners need access to non-linguistic context to be able to grasp utterance meaning. The question of whether chatbots are able to make inferences about non-linguistic properties of context at all is also considered. It is argued that they cannot, precisely because the neural language models that they rely upon for their linguistic competence are natural codes that merely associate percepts with output behaviours using encoding and decoding processes.

ACKNOWLEDGEMENTS:
Thank you Dr. Sandra Lapointe for your guidance, feedback, and words of encouragement through the entire process. Thank you Dr. Megan Stotts for your insightful comments on my drafts.

Thank you to our wonderful department here at McMaster University, and to my peers for the fun conversations.

Thank you Suyeon for all the laughs, and helping me keep my sanity.

And most of all, thank you to my parents, who supported me through thick and thin. I am proud to be your son.

Table of Contents

CHAPTER ONE

I begin this chapter by explaining a popular theory about how communication works called the code model of communication. The code model describes human communication as having to do with speakers encoding thoughts into utterances and listeners decoding utterances to recover a representation of the thought that the speaker wanted to communicate. I also explain a variation on the code model; I call it the "information-theoretic code model", and it was presented in a landmark paper by Claude E. Shannon and Warren Weaver in 1949. Next, I present two arguments against the code model. Finally, I present an alternative to the code model, called the ostensive-inferential model of communication.

1.1. The Code Model of Communication

A 'model of communication' can be defined as a theory about how communication works. Such a theory will be *descriptive* rather than *prescriptive*. That is, a model of communication aims to explain how communication works, and to describe how particular cases of communication *do* play out rather than of how particular cases of communication *should* play out. Sometimes, the various kinds of interactions that count as communication is considered to be quite large. For instance, 'communication' is sometimes understood as referring to the various interactions whereby one entity can affect another entity.[1] Instead of using such a broad definition, I understand communication to be interactions between entities, where one entity's actions cause a response in another entity, and where both the actions and response are designed to be part of the interaction.[2] The verb 'communicate' is also used frequently in my thesis; used in the verbal sense, 'communicate' refers to the activity of one entity affecting another. I will assume that 'human communication' refers to the various interactions whereby one human communicator changes the mental states of another. In human communication, it can be said that the speaker's utterance expresses a thought.

One model of communication that was (and still is) influential is the 'code model of communication', according to which communication can be described as the encoding of thoughts into language and the decoding of language to recover thoughts. The code model of

[1] This definition of 'communication' can be found in Warren Weaver (1953), page 4.
[2] Scott-Phillips (2015), Glossary.

communication has a long history. For example, John Locke wrote the following.

> "But although words can properly and immediately signify nothing but ideas in the mind of the speaker, yet men in their thoughts give words a secret reference to two other things. First, they suppose their words to be marks also of ideas in the mind of the listener. Without that they would talk in vain; if the sounds they applied to one idea were applied by the listener to another, they couldn't be understood, and would be speaking different languages. Men don't often pause to consider whether their ideas are the same as those of the listeners. They are satisfied with using the word in what they think to be its ordinary meaning in that language; which involves supposing that the idea they make it a sign of is precisely the same as the one to which literate people in that country apply that name." [John Locke, *Essay* III.ii.4.]

What Locke is saying is that language is used to communicate thoughts, and the words that can be used by a speaker to communicate her thoughts stand in an appropriate relation to the external world such that those words are meaningful to the listener.[3] Speakers and listeners are able to understand each other's utterances because they both use the same signs – that is, the same words – to refer to the same kind of object. For instance, the word 'dog' is a sign which might stand for a certain object: A furry, four-legged animal that barks and wags its tail. The speaker can say 'dog' and assume that the object that the word stands for (i.e., a specific kind of animal) is the same kind of object that speakers of English call 'dog'. Locke's theory has been described as a kind of "proto-code model of communication".[4] The general idea behind the code model is that speakers translate mental content into a corresponding utterance, and listeners translate utterances into a corresponding mental content. If the speaker and the listener are able to successfully perform encoding and decoding (respectively), and if the utterance was transmitted without significant interference (for example, a speaker's verbal utterance is not drowned out by the sound of traffic), then it is said that communication is successful. And when communication

[3] John Locke, *Essay* III.ii.5.
[4] According to Christopher Gauker, John Locke (*Essay* III.ii.4 – 5) provided a description of human communication that resembles the code model. See Gauker (1992), page 310.

is successful, the listener will have gained a representation of the thought that the speaker wanted to communicate from the speaker's utterance.[5] Code model communication is possible because communicators share an internal code of some sort which maps linguistic tokens to thoughts.

1.2. The Information-Theoretic Code Model

An information-theoretic code model of communication was presented in Claude E. Shannon's and Warren Weaver's *The Mathematical Theory of Communication* (1949). It is "information-theoretic" in the sense that Shannon and Weaver (1949) were aiming to apply the code model to problems in telecommunications. They were trying to figure out the most efficient way to transmit messages from a sender to a destination. As such Shannon and Weaver (1949) is often taken as representing the inception of the field of modern information theory. Throughout this thesis, the term 'information-theoretic code model' will be used to refer to the theory presented by Shannon and Weaver (1949); the term 'code model' will be used to refer to any model of communication that maintains that communication can be described in terms of only encoding and decoding processes.[6] We will revisit the information-theoretic code model in chapters two and three, in which I discuss communication in the context of computers and artificial intelligence research.

According to Shannon and Weaver (1949), communication is a process whereby speakers translate mental content into an utterance (i.e., encoding), and listeners recover a representation

[5] There are other ways to conceptualize communication as having to do with encoding and decoding, and the exchange and recovery of messages between communicators. For a more comprehensive look at the different ways the code model of communication has been conceptualized, see Blackburn (2007), section 3.2.

[6] Blackburn (2007) says that the code model proper is actually different from the information-theoretic code model. He attributes the conflation of the code model and the information-theoretic code model to the contributions of Warren Weaver (1953). See Blackburn (2007), pages 57 – 59. I think that the blending of the code model proper and the information-theoretic code model is important; we will see in later chapters how considering human communication as (i) having to do with encoding and decoding processes and (ii) being a probabilistic process – which are assumptions that Shannon and Weaver (1949) make – has influenced research in the field of artificial intelligence.

of that mental content from the utterance by applying that same code but in an inverse process (i.e., decoding). In the terminology of the code model in Shannon and Weaver (1949), a 'thought' or 'mental content' is referred to as a 'message'. More precisely, a 'message' can be defined as something that a speaker wants to communicate.[7] When it comes to human communication, messages are thoughts. Shannon and Weaver (1949) provide the following diagram illustrating the five components of the code model communication system and how each of the five components are related to one another.[8]

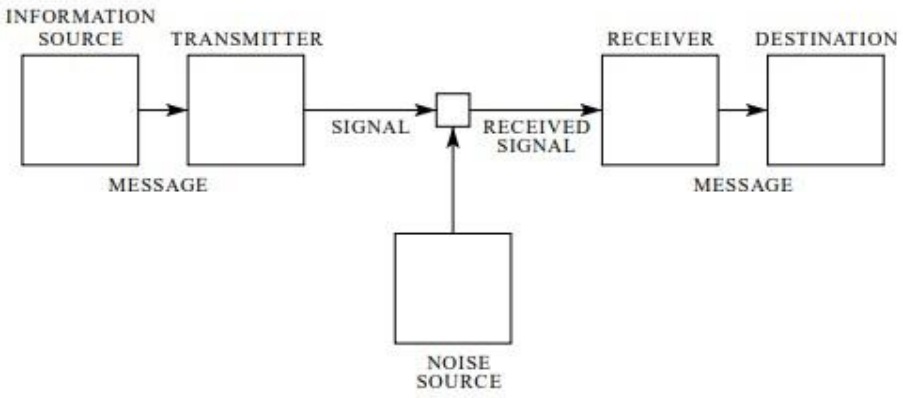

Figure 1

The following is a description of how communication works according to the information-theoretic code model.

> Step 1: The information source constructs the message to be communicated.
> Step 2: The transmitter encodes the message as a signal.
> Step 3: The signal is sent along a channel from the transmitter to the receiver.
> Step 4: The receiver decodes the signal to recover the message.
> Step 5: The message arrives at the destination.

Human communication can be described in terms of the information-theoretic code model as well. I use the term 'human communication system' to refer to human communication described

[7] Warren Weaver (1953), page 7.
[8] Shannon and Weaver (1949), figure 1, page 2.

in terms of the code model. This is to distinguish human communication in the sense of a speaker and listener having a conversation in a room (and other cases of verbal communication) from telecommunication systems, because telecommunication systems are the sorts of communication systems with which the code model was primarily concerned. Weaver (1949) says that in the case of human verbal communication, the information source can be considered as the brain, and the transmitter can be considered as the part of the human body that produces vocalizations.[9] Sperber and Wilson (1995) provide the following diagram, which adapts the information-theoretic code model to human communication systems.[10]

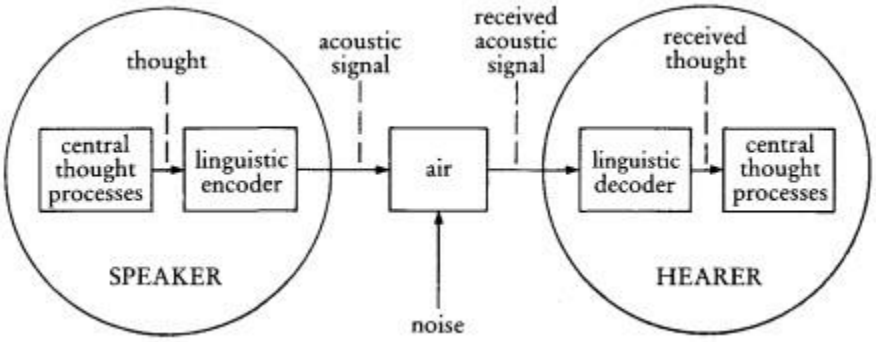

Figure 2

I think that the following is a plausible description of how human verbal communication works, in terms of the information-theoretic code model.

> Step 1*: The speaker selects the thought that they want to communicate.
> Step 2*: The thought is modified into an utterance in human natural language.
> Step 3*: The utterance is produced.
> Step 4*: The listener hears the utterance.
> Step 5*: The listener recovers a representation of the thought that the speaker had wanted to communicate.

The code model says that communication has to do with encoding and decoding processes. Steps 1/1* and Steps 2/2* together constitute encoding and decoding, respectively. Encoding involves

[9] Warren Weaver (1953), page 7.
[10] Sperber and Wilson (1995), page 5.

both the processes of a speaker choosing the thought they want to communicate and their actually encoding the thought as a signal. Similarly, decoding involves the listener hearing (in the case of human verbal communication) the utterance and recovering a representation of the thought that the speaker intended to convey. The notion of information is central to the information-theoretic code model because the message that the information source selects to communicate is dictated by the amount of information in the communication system. I do not want to dwell on what information is too much, since the concept will not play a large role in the later chapters, but since it is an important part of the information-theoretic code model, it is still worth mentioning. Weaver (1949) defines 'information' as "…a measure of one's freedom of choice when one selects a message."[11] In other words, information can be understood as a measure of the possible alternative messages available while an information source chooses a particular message. The higher the amount of information in a communication system, the more uncertain it is that the information source will select the correct message to communicate.[12] The way the information source constructs messages is a function of the amount of information in the communication system and the symbols that are being used to construct the message.[13] The information in a communication system is defined mathematically as the logarithm to the base 2 of the number of available messages that can be selected, and is measured in 'bits'. Shannon (1949) discusses the example of a two-switch relay to illustrate how the amount of information is measured in a simple communication system.[14] The "on" position in the two-switch relay corresponds to a signal S_1 that represents a message U_1, and the "off" position corresponds to S_2

[11] Warren Weaver (1953), page 9.
[12] Ibid., page 13.
[13] Shannon and Weaver (1949), page 40.
[14] Shannon (1949) does not name the various components are the two-switch relay communication system, but the components that I have described is sufficient for the purposes of the explanation.

that represents a message U_2. The amount of information in this two-switch relay communication system is $\log_2(2) = 1$ bit. In the information-theoretic code model, 'message' and 'information' refer to two different things. A message is something that has a meaning, whereas information is not something that strictly has meaning. Information is also not a description of how much detail a message contains, which is perhaps how we might use the term 'information' in ordinary speech.[15] The example of the two-switch relay I had mentioned above illustrates the difference between 'information' and 'message'. We can imagine in the example of the two-switch relay that the messages are the following: U_1 = 'It is overcast', and U_2 = 'It is clear'. The meanings of U_1 and U_2 have nothing to do with the amount of information in the two-switch relay communication system, which will still be 1 bit. The amount of information in a communication system partly determines which messages are constructed. So even though it is messages that have meaning, the information-theoretic code model maintains that the meanings of messages are not relevant to how messages are constructed. What the demarcation between messages and information illustrates is that the information-theoretic code model was concerned with a very specific problem about communication.

1.3. The Three Levels of the General Communication Problem

Shannon and Weaver (1949) were optimistic that the code model could solve what they refer to as the "general communication problem": How can a message sent by a source be reproduced at a destination?[16] There are three levels of the general communication problem.[17]

> Level A: How accurately can a message be communicated?
> Level B: How precisely do transmitted symbols convey the intended meaning?
> Level C: How effectively does the meaning recovered at the destination affect behaviour?

[15] In fact, there is some controversy over what exactly 'information' refers to in the theory presented by Shannon and Weaver (1949). This will not be of concern for the purpose of this thesis, but for a discussion of this controversy, see Lombardi et al (2016)..
[16] Shannon and Weaver (1949), page 31.
[17] Ibid., page 24.

Shannon and Weaver (1949) were concerned primarily with finding a solution to the problem at Level A. But Weaver (1949) says that the information-theoretic version of the code model that they had presented can possibly provide solutions to the problems at Level B and Level C as well.[18] Solutions to the problems at Level B and Level C are difficult, and might require sophisticated modifications to the original components of the communication system that was illustrated in *Figure 1*. Weaver (1949) writes that these modifications might include such additional components as a "semantic receiver" in-between the original receiver and the destination, and a "semantic noise" component in-between the information source and the transmitter.[19] Weaver (1949) does not provide specific details about the nature of the semantic receiver component, the semantic noise component, or any other such modifications to the original code model. However, in the quote below, he suggests that the understanding of human communication the information-theoretic code model affords us can set a foundation for a theory of meaning.

> The concept of information developed in this theory at first seems disappointing and bizarre – disappointing because it has nothing to do with meaning, and bizarre because it deals not with a single message but rather with the statistical character of a whole ensemble of messages, bizarre also because in these statistical terms the two words *information* and *uncertainty* find themselves to be partners. I think, however, that these should be only temporary reactions; and that one should say, at the end, that this analysis has so penetratingly cleared the air that one is now, perhaps for the first time, ready for a real theory of meaning. [Warren Weaver (1949), page 27.]

Even the mathematical apparatus invoked by the code model is seen as instructive for a theory of meaning: "The idea of utilizing the powerful body of theory concerning Markoff processes seems particularly promising for semantic studies, since this theory is specifically adapted to handle one of the most significant but difficult aspects of meaning, namely the influence of

[18] Blackburn (2007), page 70.
[19] Warren Weaver (1953), page 26.

context."[20] Shannon and Weaver (1949) explain what a Markov process is and how the process of constructing messages (where a message is something that has meaning) can be considered as a Markov process.

> The general case can be described as follows: There exist a finite number of possible "states" of a system; $S_1, S_2, \ldots, S_n$. In addition there is a set of transition probabilities, $p_i(j)$, the probability that if the system is in state S_i, it will next go to state S_j. To make this Markoff process into an information source we need only assume that a letter is produced for each transition from one state to another. The states will correspond to the "residue of influence" from preceding letters. [Warren Weaver (1949), page 6.]

A Markov process is a process that assigns a probability value to the current element based on the probability values of previous elements. The Markov process is said to be *probabilistic*, as the value assigned to the current element will depend on the values assigned to previous elements. Treating the process of constructing messages as a Markov process is supposed to allow the theory to account for the influence of context on the message that is selected by the information source. The information-theoretic code model maintains that messages are constructed by concatenating "elementary symbols", where the elementary symbol to be concatenated next (i.e., the symbol that has been selected) will depend on which symbols have already been concatenated. What an elementary symbol is can be illustrated using the following example: The message 'The door was left wide open' is selected from the concatenation of the elementary symbols 'the', 'door', 'was', 'left', 'wide', 'open', where these elementary symbols are words that are part of the vocabulary of the English language. Similarly, individual words in the English language are constructed by the concatenation of elementary symbols, where these elementary symbols are the twenty-six letters in the English language. More generally,

[20] Ibid., page 28.

elementary symbols are the discrete symbols that when concatenated can compose a message.[21] The "context" in this case is just the symbols that have already been concatenated.[22]

The information-theoretic code model had never originally set its sights on providing a theory about meaning. Rather, the original purpose of the information-theoretic code model was to provide insight into the problem of how to transmit messages as accurately as possible in telecommunication systems – this is, as Weaver (1949) put it, an engineering problem.[23] However, Weaver (1949) is also optimistic about the power of the information-theoretic code model to solve the general communication problem, suggesting modifications to the information-theoretic code model could allow it to describe even the difficult case of human natural language communication. According to Blackburn (2007), Warren Weaver's claims are overoptimistic, and have led widespread acceptance in linguistics of the idea that the information-theoretic code model actually tries to explain how human communication works.[24] In the next section, I will discuss two criticisms of the code model of communication. The two criticisms I am considering are criticisms of not only the specific information-theoretic code model presented by Shannon and Weaver (1949), but also of the more general notion of human communication as having to do primarily with the encoding of thoughts into language and the decoding of language to recover a representation of the thought. In the second chapter, I argue that artificial intelligence research aiming to make a computer communicate with humans using human natural language considers human communication as operating within the paradigm of the code model. This is evident from the methodology employed by the majority of artificial intelligence researchers.

[21] Ibid., page 11.
[22] See Shannon and Weaver (1949), page 43 – 44., for an illustration of how the process of constructing approximations of English sentences can be considered a Markov process. Chomsky (1957) demonstrates that there is no Markov chain that can generate all of the possible grammatical sentences in the English language. For more on this, see Blackburn (2007), page 87.
[23] Warren Weaver (1953), page 6.
[24] Blackburn (2007), page 59.

1.4. Two Criticisms of the Code Model of Communication

In this section, I will explain two criticisms of the code model, referred to as 'C1' and 'C2'. They are below.

> *The Criticism from Pragmatics (C1)*: The code model cannot explain how utterance comprehension works.
> *The Criticism from Language Evolution (C2)*: Human communication is combinatorial, and combinatorial communication could not have evolved from a cognitive mechanism that relied solely on encoding and decoding processes.

C1 and C2 dispute the claim that communication has to do with only encoding and decoding processes. Before proceeding I want to briefly mention that throughout the rest of this chapter, the term 'intention' is used to refer to a type of mental state with "aboutness", or having some content. That is, an intention is a mental state that is about something. This is in keeping with how the term 'intention' is used in the literature about pragmatics, communication, and computer science (which will be introduced in chapter two). The term 'intentional communication' will refer to communication that aims to change the listener's mental states.

1.4.i. The Criticism from Pragmatics (C1)

Pragmatic theories aim to provide an account of how a listener can interpret a speaker's utterance in order to recover the meaning intended by the speaker. I call C1 a criticism of the code model from pragmatics. In order to elucidate C1, I need to provide an explanation of a key idea from pragmatics – namely, that there is a difference between sentence meaning and utterance meaning. In addition, I will explain the notion of 'code' in greater detail, and how it is related to human communication.

What 'sentence meaning' refers to is the invariant, timeless meaning of a linguistic expression, which is a function of word meanings and syntax; on the other hand, what 'utterance meaning' refers to is the meaning of a speaker's utterance. Another way to conceptualize the

difference between sentence meaning and utterance meaning is that sentence meaning applies to sentences, whereas utterance meaning applies to utterances. A sentence is a piece of language that can *potentially* express a thought, and an utterance is the act of actually expressing a thought.[25] An utterance is a sentence that has been expressed at a particular point in time. Sperber and Wilson (1995) also explain the difference between a sentence and an utterance in terms of the kinds of properties each has: A sentence has only linguistic properties, whereas an utterance has *both* linguistic *and* non-linguistic properties.[26] Linguistic properties include such properties as containing a pronoun, containing an adverb, and so on; linguistic properties are properties of the syntax and grammar of the linguistic expression. They have to do with only the linguistic expressions themselves. Non-linguistic properties include properties like being spoken at the dinner table, being whispered, and so on. Non-linguistic properties are properties of utterances derived from the physical environment in which an utterance is expressed. The difference between sentence meaning and utterance meaning is a matter of the various non-linguistic properties an utterance has over-and-above its linguistic properties.

The notion of the meaning of an utterance includes both the linguistic and the non-linguistic properties of the utterance in its purview. A further distinction can be made between two types of utterance meaning: The first is 'explicature' and the second is 'implicature'.[27] Explicature is the meaning of an utterance that can be inferred directly from its non-linguistic properties. An utterance can have more than one explicature. Implicature is what the speaker

[25] Sperber and Wilson (1995), page 9.
[26] Ibid.
[27] I am referring to the notion of explicature developed by Sperber and Wilson. See Sperber and Wilson (1993), pages 5 – 6, for a more detailed account of explicatures. Robyn (2002) has also developed this notion of explicature. Bach (1994) develops the related but different notion of impliciture; see Bach (2010) for the differences between impliciture and explicature.

intended by the utterance.[28] I have provided a table which includes an example illustrating the differences between explicature and implicature. The sentence being considered in the example is 'It is not ethical to keep animals in zoos.' An utterance of that sentence will have both explicatures and an implicature.

	Linguistic Properties?	**Non-Linguistic Properties?**	**Example**
Sentence	Yes	No	'*It is not ethical to keep animals in zoos*.'
Utterance (Explicature)	Yes	Yes	'*Denise believes that it is not ethical to keep animals in zoos*.'
Utterance (Implicature)	Yes	Yes	'*We should not visit zoos*.'

Figure 3

In the table above, we see that the meanings of the words 'it', 'is', 'not', 'ethical', … and the grammar of the sentence, will determine the meaning of the sentence. The explicature above can be identified by using non-linguistic information about who the speaker is – in this case, that the speaker is someone named 'Denise'. In order to identify the explicature, the listener must appeal to properties other than the linguistic properties of the sentence. Finally, implicature is what is not directly expressed by the utterance, but what can be inferred from it. The sentence does not really say that we should avoid visiting zoos, but the speaker can implicate this by way of their utterance. The implicature has to do with the speaker's cognitive states in addition to the non-linguistic properties of the utterance (such as what the referents of the speaker's words are).

[28] The distinction between the meaning of an utterance and the meaning intended by a speaker by their utterance was developed perhaps most famously by Grice (1957). This distinction was also developed in the speech-act literature by J. L. Austin (1962) and John Searle (1969). Sometimes explicature is referred to as literal meaning, and implicature as non-literal meaning.

I will now explain what the notion of "code" has to do with human communication. Sperber and Wilson (1995) say that a code is "…a system which pairs messages with signals, enabling two information-processing devices (organisms or machines) to communicate".[29] One kind of code relevant for communication is the generative grammar, which Sperber and Wilson (1995) say associates phonetic representations of sentences to semantic representations of sentences.[30] Phonetic representations of sentences are sounds that make up words and sentences (and when spoken, in the case of verbal linguistic communication, they can make up utterances). Each of those sounds is often referred to as a 'phoneme'. The semantic representation of a sentence is the meaning associated with a phoneme or sequence of phonemes – that is, the semantic representation is just the *meaning* of a sentence. A description of a listener engaging in the process of utterance comprehension – in terms of phonetic representations, the generative grammar, and semantic representations – might go as follows: The listener hears the utterance (i.e., the phonetic representation of a sentence); the listener uses a code (i.e., a generative grammar, which is shared between the listener and the speaker) to associate the utterance (i.e., the sounds made by the speaker's vocal organs when they speak) with a semantic representation of what the sounds when combined together mean (i.e., the meaning of the sentence corresponding to the utterance).

I have explained the distinction between sentence meaning and utterance meaning, and what a code is – now I will explain C1. Sperber and Wilson (1995) argue that codes, such as generative grammars, are not concerned with non-linguistic properties of utterances. Instead, the generative grammar describes "…a common linguistic structure, the sentence, shared by a

[29] Sperber and Wilson (1995), page 3 – 4.
[30] Sperber and Wilson (1995), page 8 – 9. Also see Chomsky (1965), in which the following equivalent definition is provided: "…by a generative grammar I mean simply a system of rules that in some explicit and well-defined way assigns structural descriptions to sentences." (page 8).

variety of utterances which differ only in their non-linguistic properties."[31] In other words, what is referred to by 'semantic representation of a sentence' is sentence meaning, as sentence meaning has only to do with linguistic properties, and nothing to do with non-linguistic properties. The same also holds true for the information-theoretic code model described in Shannon and Weaver (1949), as their theory does not take into consideration the non-linguistic properties of an utterance in the description of communication. Shannon and Weaver (1949) say that communication is the encoding and decoding of signals to recover messages. But neither the process of encoding nor the process of decoding has anything to do with the meaning of the message that is being encoded and decoded. Rather, encoding and decoding have to do with only the measure of the amount of uncertainty in the communication system (i.e. information), and the elementary symbols with which messages are composed. The speaker encodes a thought as a sentence and expresses the sentence as an utterance; the listener decodes the utterance and recovers the thought that the speaker intended to communicate. If there is a low amount of information in the communication system, then communication is more likely to be successful than if there were a high amount of information. Weaver (1949) had proposed several modifications to this basic encoding-decoding process which were supposed to be able to account for the complexities of ordinary human communication (although Weaver (1949) does not explain how these modifications might work, or how they might actually be implemented in a communication system). The overarching theme behind the information-theoretic code model, just like for the code model generally, is that the story of how utterance comprehension works has to do with only the linguistic properties of an utterance.

[31] Sperber and Wilson (1995), page 9.

Sperber and Wilson (1995) motivate their criticism of the code model by asking whether a listener can successfully recover a representation of the thought expressed by a speaker's utterance if all that the listener has access to is the linguistic properties of the utterance. If the answer to this question is yes, then communication can be explained only in terms of encoding and decoding. This is because a code (such as a generative grammar) is concerned only with the linguistic properties of the utterance; so as long as the listener is competent in the language in which the utterance is expressed (i.e., English, French, Malayalam, Finnish, or whatever other human natural language), the listener should be able to engage in the decoding process to recover a representation of the thought that the speaker intended to express by their utterance.[32] But Sperber and Wilson (1995) say no: In order for linguistic communication to be successful, the listener often needs to access the non-linguistic properties of an utterance to ascertain the meaning of the utterance – more precisely, the speaker needs to be able to access the non-linguistic properties of an utterance to ascertain both the meaning of the utterance (i.e., explicature), and the meaning intended by the speaker making the utterance (i.e., implicature). Here is an example that shows how the linguistic properties alone of an utterance are insufficient to grasp utterance meaning. Suppose a speaker expresses the sentence 'Priya said she is going to the bank later today'. Possible explicatures of the speaker's utterance are the following.[33]

(a) Priya is going to go to the bank later today.
(b) Priya wants to go to the bank later today.
(c) Priya thinks she is going to go to the bank later today.

[32] There are other things that need to go right as well. For example, a large amount of noise in the communication channel can impede the transmission of the signal. If the listener receives a signal that has been affected by noise in the channel, then the listener will be less likely to recover an accurate representation of the thought intended by the speaker from the speaker's utterance.
[33] This example is inspired by Sperber and Wilson (1993), page 5.

The linguistic properties of the speaker's utterance can *help narrow down* possible explicatures, including (a) – (c). For instance, the indexical 'she' is a pronoun in English which is often used to refer to some individual identifying as female. According to the rules of English sentence structure, 'Priya' seems to be the subject of the sentence, in virtue of the subject-verb-object sentence structure. The subject ('Priya') presumably performs an action, which is denoted by the verb 'going'. And the temporal indexical 'today' suggests that the speaker intends to refer to an event that will take place on the day immediately following the day on which the utterance is made. In this manner, the linguistic properties of an utterance can yield some insights about the explicature of an utterance because of very general rules-of-thumb about such things as what various indexical expressions tend to refer to, what part of a sentence is the verb, how an adverb can modify the verb it describes, and more. However, the listener must use non-linguistic properties of the utterance in order for a listener to be able to identify which explicatures the utterance *actually* expresses. The non-linguistic properties of that utterance are facts such as which particular individual that the proper name 'Priya' is being used to refer to, which individual the pronoun 'she' is being used to refer to, and whether 'bank' is being used to refer to the financial institution or the land next to a river. There are two reference ambiguities in the sentence 'Priya said she is going to the bank later today'. The first is whether 'she' is referring to 'Priya', or whether 'she' is referring to some individual who is not the same individual as Priya. And the second reference ambiguity is whether 'bank' refers to the financial institution, or to the piece of land next to a river. If the speaker points at a particular individual and says 'Priya said she is going to the bank later today' with an emphasis when saying 'Priya', the listener can use the speaker's pointing gesture as evidence to infer that 'she' might refer to the individual named 'Priya'. And there is nothing about the lexical meaning of the word 'bank' that can allow the

listener to know whether ‘she’ is going to the financial institution later that day or to the piece of land next to a river. For a listener to know what ‘bank’ is being used to refer to, the listener must use clues such as previous conversations that they had with the speaker, whether there are any rivers in the area or not, and so on. The kinds of inferences the listener must make that I just mentioned require the listener to attend to things other than the meanings of the words that make up the utterance. Properties such as tone of voice are also non-linguistic: The meaning that the speaker intended to convey by saying ‘Priya said she is going to the bank today’ solemnly seems to be different from if they say it excitedly.[34] Thus, a description of how communication works that does not also explain how communicators can reason about non-linguistic properties when engaging in the process of utterance comprehension is inadequate. This is because human communication often involves the production and comprehension of utterances, which have both linguistic and non-linguistic properties. The meaning of an utterance is a function of its linguistic and its non-linguistic properties, and a model of communication should be able to explain how utterance comprehension works. The code model maintains that communication between speakers and listeners is possible in part because speaker and listeners share a code in common. The code is a generative grammar that allows speakers to generate utterances in the language, and listeners to interpret utterances in that same language. The code involved in code model communication accesses only the linguistic properties of utterances. Because the non-linguistic properties of utterances affect utterance meaning, the code model must be supplemented with some mechanism that can explain how listeners use those non-linguistic properties. Sperber and Wilson (1995) call this an “inference mechanism”.

[34] See Sperber and Wilson (1995), page 10, for more examples.

An inference mechanism is distinct from a decoding mechanism in the following sense. Decoding is the process of associating signals with messages, where any given message will have no relationship to the corresponding signal. In other words, the form of the signal does not provide any clues about what the message that the signal encodes means. To reiterate, the form of a signal is the purely physical properties of the signal. In the case of verbal communication, for example, the form of an utterance is the compressions and decompressions of air; the message that is expressed with the utterance cannot be identified based on the compressions and decompressions of air alone. On the other hand, inference is the process of drawing conclusions from a set of premises, where the conclusions that are drawn are related to the premises. Inference rules are rules that dictate the conditions under which conclusions can be drawn from premises. Inference rules include modus ponens, modus tollens, and more. In the case of human communication, the listener can determine which inferences are warranted by an utterance by following inference rules. Generally, an inference is warranted by an utterance if there is evidence which suggests that the inference is likely to be true. The evidence for an inference are non-linguistic properties of the utterance, such as who the speaker is, the speaker's and listener's particular life experiences, the social relationship between the speaker and listener, and more. The following example, from Sperber and Wilson (1995), illustrates how a listener can use non-linguistic properties and an inference mechanism in the process of utterance comprehension. Consider an utterance of 'Jones has bought the *Times*'. This utterance contains an ambiguous reference: does 'the *Times*' refer to a copy of the newspaper, or does 'the *Times*' refer to the company that publishes the newspaper? It is only in extremely limited cases that 'the *Times*' can be taken as referring to the company rather than to a copy of the newspaper – in ordinary cases, a listener will not infer the utterance as suggesting that Jones bought the company. A listener will

likely draw the inference that 'Jones has bought a copy of the *Times*' rather than 'Jones has bought the company that publishes the *Times*' because there is stronger evidence warranting the former inference than the latter. The listener will make their inference based on certain assumptions, perhaps about who the speaker is, whether the speaker is trustworthy, and so on. These assumptions provide the listener with reasons for eliminating possible inferences, thereby identifying the inference(s) that is/are warranted. The point of this example is to illustrate that assumptions made by the listener can serve as premises in an inference mechanism, and the listener can use the evidence for these assumptions to draw inferences about utterance meaning. Sperber and Wilson (1995) refer to the set of premises from which inferences are drawn in communication as the 'context'. A premise can be considered an assumption, or belief. Thus, the context can be thought of as the assumptions and beliefs that a communicator has about the world.[35] A listener's beliefs about Jones' socio-economic status, Jones' tendency to be honest or lie, the speaker's tendency to be honest or lie, etc. are part of the context from which the listener will make inferences about what 'the *Times'* in the utterance 'Jones has bought the *Times*' refers to.

For an inference mechanism to be useful in a model of communication, it must be possible for the speaker and the listener to restrict the set of premises from which inferences are drawn to those premises that both of them share. This is because if a listener cannot identify which premises they share with the speaker, then an inference mechanism will not be able to explain how the listener knows which non-linguistic properties to attend to during the process of utterance comprehension. It is plausible that the speaker and the listener will share at least some assumptions about the world. These might include beliefs about their immediate physical

[35] Sperber and Wilson (1987), page 698.

environment(s), how the physics of the world works, beliefs about the language that both the speaker and listener are using, assumptions that their conversational partners will be cooperative, and so on. However, Sperber and Wilson (1995) say that the context from which the listener makes inferences will also include premises that are highly idiosyncratic to them: The listener's context will include beliefs that are based on the listener's personal life experiences, the listener's perceptions of themselves, and so on. In the same way, the speaker's context will include beliefs that are highly idiosyncratic. And it is fairly uncontroversial that an individual's interactions with the world are coloured by their personal life experiences, their beliefs, and so on. Sperber and Wilson (1995) say that because the listener's context can include assumptions that are idiosyncratic to them, there is no guarantee that listeners are drawing inferences from the premises that they share with the speaker. [36] In conclusion, the code model cannot explain how utterance comprehension works, whether it is supplemented with an inference mechanism or not.

1.4.ii. The Criticism from Language Evolution (C2)

The second criticism of the code model that I will consider in this thesis, C2, claims that human natural language is combinatorial, and combinatorial language could not have evolved from a cognitive mechanism that relied solely on encoding and decoding processes. This is an argument that was presented by Thom Scott-Phillips in his book *Speaking Our Minds: Why Human Communication is Different and How Language Evolved to Make it Special* (Scott-Phillip, 2015). His argument is a criticism of the code model from language evolution in two senses. Firstly, the notion that communication involves encoding and decoding processes entails that

[36] Sperber and Wilson (1987/1995) describe the process of communicators identifying shared assumptions as involving a regress of higher and higher order assumptions. Sperber and Wilson (1987), page 698; and Sperber and Wilson (1995), page 18. They credit David Lewis (1969) and Schiffer (1972) as having first identified and defined the regressive process that individuals can engage in in order to identify shared/mutual beliefs. See Sperber and Wilson (1987), page 698; and Sperber and Wilson (1995), page 18.

communicators must have had something to gain (i.e., an evolutionary advantage) from producing and processing signals. I will elaborate on this point over the course of the discussion below. Secondly, Scott-Phillips (2015) is making a case for why research in the field of language evolution should question the assumption that communication in non-human animals can also be explained by purely encoding and decoding processes – that whether the paradigm of the code model, within which much of the research into language evolution is being conducted, is actually telling the right story about how communication works. Researchers often focus on trying to find cognitive mechanisms for rudimentary language-use in the great apes that might serve as precursors to the same (albeit more developed) cognitive abilities in human beings that allow human beings to use language.[37] However, Scott-Phillips (2015) says that human linguistic communication is actually a result of cognitive mechanisms that do not have to do with encoding and decoding. Research about whether similar cognitive mechanisms as these are present in non-human animals (like the great apes) could be fruitful. More will be said in the next section of this chapter about what exactly the cognitive mechanisms human beings possess are that allow humans to engage in linguistic communication.

Before presenting C2, I want to make a note about the terminology that will be used in the rest of this section. Somewhat unfortunately, the terms 'information' and 'signal' are used in Shannon and Weaver (1949) to refer to something different than in Scott-Phillips (2015). In Shannon and Weaver (1949), the term 'information' refers to a very specific, technical aspect of the information-theoretic code model. According to Shannon and Weaver (1949), information is a measure of the amount of uncertainty in a communication system. The amount of information in a communication system (partly) determines how messages are constructed. They use the term

[37] Scott-Phillips (2015), page 83 (2.7).

'signal' to a physical change in a channel that encodes a message and can be decoded by a receiver. In Scott-Phillips (2015), the term 'information' refers to something that is communicated between individuals. Scott-Phillips (2015) describes information transfer as '…the consequence of communication, not a definition of it).[38] So, what is referred to as 'information' by Scott-Phillips (2015) can be roughly equated with is referred to as 'message' in Shannon and Weaver (1949). The term 'signal' in Scott-Phillips (2015) refers to "The action in communication i.e. the action that causes a reaction in another organism, where both action and reaction are designed to be part of the interaction."[39] What a signal is according to Scott-Phillips (2015) will be explained in more detail in the following paragraphs. For now, it can simply be noted that the term 'signal' is used in Scott-Phillips (2015) to refer to something different than it does in Shannon and Weaver (1949). Now, let us set the terminological preamble aside and look at C2.

Any given human natural language consists of a finite set of symbols, a finite vocabulary, and a finite set of grammatical rules. However, human natural language is infinitely expressive: Using a finite vocabulary, we can express an infinite number of thoughts. Our ability to do this is referred to by Scott-Phillips (2015) as 'combinatorial communication'. Communication can be 'combinatorial' in the sense that two signals can be combined to create a third signal, where the third signal has the same form as the combination of the two signals but an entirely different meaning from either of them. The figure below, which is adopted from Scott-Phillips (2015), illustrates the construction of a combinatorial signal, A + B ('A + B' is also referred to as a composite signal), from two signals, A and B ('A' and 'B' are also referred to as holistic

[38] Thom Scott-Phillips (2015), page 61 (2.2), and Scott-Phillips (2010), e1 – e2.
[39] Thom Scott-Phillips (2015), Glossary.

signals). When an organism expresses signal A, they communicate X; when they express signal B, they communicate Y; and by expressing A + B, they communicate Z.

Form		Meaning
A	→	X
B	→	Y
A + B	→	Z ≠ X + Y

Figure 4

A response is a behaviour that is produced A signal can be established in one of three ways: ritualization, sensory manipulation, or the direct route.[40] Ritualization is the process of one organism, call it O1, adapting another organism's, call it O2, behaviour into a cue. Usually, this happens because O2's behaviour provides O1 with information about its environment, and having this information gives O1 an evolutionary advantage of some sort. Sensory manipulation happens when O1 has some behavioural disposition, and O2 performs an action that (i) exploits this behavioural disposition and (ii) in doing so gains an evolutionary advantage.[41] In the case of both ritualization and sensory manipulation, once organisms consistently produce behaviours that serve as cues and consistently respond to behaviours that look like cues, the cue becomes a signal. An organism's response to a signal is simply referred to as a 'response'. Communication in much of the animal kingdom can be described as the exchange of information between organisms by way of signals and responses, where the signals and responses have been adapted to facilitate this exchange of information.

When it comes to how signals are established via the direct route, things get a little more complicated. A signal that emerges via the direct route "signals its own signalhood… Their

[40] Thomas C. Scott-Phillips and Richard A. Blythe (2013), page 2.
[41] Scott-Phillips (2015), pages 64 – 66 (2.3).

respective audiences grasp this, and so the signal and its response appear together."[42] In other words, a signal that is established through the direct route demonstrates to the audience (i) *what* the signaller wants to communicate, and (ii) *that* the signaller wants to communicate. An example of a signal that emerges through the direct route is someone tilting their coffee cup to indicate to the waiter that they want more coffee.[43] These signals are distinct from those signals that are established through ritualization or sensory manipulation, because a signal that signals its own signalhood allows for the signal and response to appear at the same time. In contrast, a signal established through ritualization entails that the cue becomes a signal only if the individual producing the cue gains an evolutionary advantage by having the cue recognized as a signal. Similarly, a signal is established through sensory manipulation only if an individual gains an evolutionary advantage by treating a behaviour as a signal. Signals that emerge through the direct route are also unique as "there are no constraints on either signal form or signal meaning…"[44] Signals that emerge through the direct route can be used in an infinite number of ways; this is precisely because when a signal is produced, the response is generated at the same time. A consequence is that such signals can be invented on-the-fly, even during the course of a particular communicative exchange. On the other hand, signals that emerge through either ritualization or sensory manipulation are limited in their uses. Signals that emerged through either ritualization or sensory manipulation only emerged because there was an evolutionary advantage to treating certain behaviours as cues or as responses. The fact that these signals emerged as they presented an evolutionary advantage severely restricts the ways in which they can be used.

[42] Ibid., page 72 (2.5).
[43] Ibid. page 74 (2.5).
[44] Ibid., page 73 (2.5).

Once a signal and a response have been established, they become a code. Signals and responses established via either ritualization or sensory manipulation are called ‘natural codes’, whereas those established via the direct route are called ‘conventional codes’. Because the nature of signals and responses that make up natural codes are fundamentally different from those that make up conventional codes, Scott-Phillips (2015) says that natural codes and conventional codes are ontologically distinct.[45] Natural codes are involved in code model communication, whereas conventional codes are involved in human communication.[46] Scott-Phillips (2015) defines human natural language as a set of conventional codes – human natural language is also referred to as a linguistic code. Other examples of conventional codes include sign language, Morse code, and which side of the road we drive on in Canada versus in the UK.[47] Conventional codes are special because the associations between signals and corresponding responses hold in virtue of the social community which uses those signals and responses.[48] Any given word in the vocabulary of English, for instance, will have some associated meaning. The meaning of the word ‘tree’ is understood (conventionally) as referring to a particular kind of flora. This is true even for words whose referents are more challenging to identify – indexical expressions are a good example of this. The indexical ‘I’ is understood conventionally as referring to the speaker, and ‘you’ is understood conventionally as referring to the individual the speaker is addressing. But a rigorous account of the semantics of indexical expressions can sometimes be quite complicated. The point remains, however, that the meanings of signals and corresponding responses in a conventional code are a matter of their having been accepted by the community that uses them.

[45] Ibid., pages 41 – 42 (1.5).
[46] Ibid., page 42 (1.5).
[47] Ibid., page 43 (1.5).
[48] Ibid.

It is not straightforward to explain how exactly a signal could have signalled its own signalhood in pre-linguistic times. Scott-Phillips (2015) says that one possible explanation for how a pre-linguistic (i.e., before linguistic codes were established) signal could have signalled its own signalhood is because human beings evolved with a sophisticated metapsychology, allowing us to represent each other's mental states, and to represent representations of each other's mental states (i.e., to form mental metarepresentations). With a sophisticated metapsychology, we are able to reason about those representations using a cognitive mechanism called 'recursive mindreading'. Recursive mindreading is the cognitive mechanism required to be able to establish a signal through the direct route; recursive mindreading is fundamentally different than the cognitive mechanism required to form mere associations between signals and responses (which is sufficient for establishing a natural code). To establish a natural code, organisms must be able to treat each other's behaviour as a signal and produce a behaviour as a response. Signal-response pairs become a natural code because these are behaviours that are advantageous for the organisms' survival. There are studies which suggest that certain human communicative behaviours are also part of a natural code. For example, a study presented in David Matsumoto and Bob Willingham (2009) suggests that facial cues representing "celebration" or "cheering" are not learned; the authors demonstrate that there is no variation between these facial cues between congenitally blind, non-congenitally blind, and sighted athletes. What this suggests is that certain types of communicative behaviour are evolutionarily hard-wired in human beings.[49] The cognitive mechanisms involved here do not require more than the ability to react to a signalling behaviour. If an organism gains an evolutionary advantage by reacting to particular signal, then the organism will keep reacting to those signals in the future. On the other hand,

[49] Matsumoto and Willingham (2009). For a more comprehensive look at research in this area, see Valente, Theurel, and Gentaz (2018).

human combinatorial communication could not have evolved from a cognitive mechanism that relies solely on encoding and decoding processes. This is because for a signal to be able to signal its own signalhood, organisms must be able to make inferences about the mental states of other organisms. Without being able to make those kinds of inferences, signalling behaviours would not be considered by organisms as signals. The kinds of inferences required here will be discussed in more detail in the next section. Human communicators can figure out whether a behaviour that looks like a signal really is a signal, and what response behaviour to perform after having judged a behaviour to be a signal.[50] So far, the only organisms that seem to have such an advanced metapsychology – advanced enough to have been able to establish complex linguistic codes – are human beings. What exactly these cognitive mechanisms are that made combinatorial human communication possible will be discussed in the next part of this chapter, in which I introduce an alternative to the code model of communication.

1.5. The Ostensive-Inferential Model of Communication

The term 'ostensive-inferential model of communication' was coined in Sperber and Wilson (1986/1995) to refer to an alternative model of communication to the code model.[51] Scott-

[50] There are really two lines of argument for why combinatorial communication is so uniquely human. Both have to do with human beings establishing codes via the direct route, which requires cognitive mechanisms that do not rely on only encoding and decoding processes. Scott-Phillips and Blythe (2013) provides a mathematical demonstration for why complex combinatorial communication (such as human language, which is infinitely expressive) is so rare in the animal kingdom. The common thread is that ultimately, the rarity of complex combinatorial combination is because human beings have cognitive mechanisms that do not rely on only encoding and decoding processes.

[51] It must be noted that the ostensive-inferential model of communication proposed by Sperber and Wilson was heavily inspired by the work of H. P. Grice. For Grice, communication is a cooperative enterprise; he claims that human communicators follow what he calls the "Cooperative Principle", which states that when we communicate, we aim to work together to interpret one another's utterances and ensure that communication is successful. We satisfy the cooperative principle by following four "maxims of conversation": the maxims of quality, quantity, manner, and relation. Quality states that we must not say what we think to be false; Quantity states that we must not provide more information than is appropriate; Manner states that we must communicate clearly, avoiding ambiguity and obscurity; Relation states that we must provide information that is relevant to the conversation. The maxims of conversation can be considered as heuristics that describe the features of successful communication, rather than strict norms that necessarily delimit what successful communication is. For more about Grice's Maxims, see Grice (1967) and Sperber and Wilson (1981).

Phillips (2015) has also contributed to developing the ostensive-inferential model. Ostensive-inferential communication involves the production and recognition of two kinds of intention: an informative intention, and a communicative intention.[52]

> Informative intention: The speaker's intention to change the mental states of the listener.
> Communicative intention: The speaker's intention to make the listener recognize the speaker's informative intention.

Ostensive-inferential communication can be described as follows: A speaker conveys their intention to give the listener some information, and the speaker intends the listener to recognize that they want to communicate in the first place. Ostensive-inferential communication requires that communicators can represent and reason about the mental states of others – Scott-Phillips (2015) refers to the ability to represent the mental states of others as 'metapsychology'. Mental states can be beliefs, desires, intentions, and so on. The process of reasoning about other's mental states is called 'recursive mindreading'. Scott-Phillips (2015) provides an example that illustrates how recursive mindreading works.[53] The example consists of six scenarios involving two individuals, Mary and Peter. The first scenario is that Mary finds berries in the woods, and she picks and eats them because they are edible. In the five other scenarios, Peter sees Mary picking and eating berries in the woods, and both Peter and Mary engage in higher and higher levels of recursive mindreading. Each subscript (1, 2, and so on) denotes an additional layer of metarepresentation (i.e., a representation of a representation).

> (1) Mary picks and eats berries in the woods because the berries are edible.
> (2) Peter sees Mary picking and eating berries in the woods and believes$_1$ that the berries are edible.
> (3) Mary intends$_2$ that Peter believe$_1$ that the berries are edible.
> (4) Peter believes$_3$ that Mary intends$_2$ that he believe$_1$ that the berries are edible.
> (5) Mary intends$_4$ that Peter believe$_3$ that she intends$_2$ that he believe$_1$ that the berries are edible.

[52] See Sperber and Wilson (1995), page 58, 61. Also, Scott-Phillips (2015), page 28 – 29 (1.3).
[53] Scott-Phillips (2015), page 84 (3.4). See also Sperber (2000) and Grice (1967).

(6) Peter believes$_5$ that Mary intends$_4$ that he believe$_3$ that she intends$_2$ that he believe$_1$ that the berries are edible.

Scott-Phillips (2015) says that Mary has a communicative intention only in (5) and (6), and it is only in (6) that true ostensive-inferential communication happens. In scenario (4) Peter recognizes that Mary is communicating intentionally, but this interaction does not amount to ostensive-inferential communication; a listener does not need to access deep levels of recursive mindreading in order to recognize that the speaker is communicating intentionally. True ostensive-inferential communication requires the expression and recognition of specifically the informative intention and communicative intention. Human communicators can express and recognize those intentions because of the capacity to form high-order metarepresentations of each other's mental states, as well as because of the cognitive mechanism of recursive mindreading, which is the ability to reason about those metarepresentations. Human beings gain the ability to engage in higher and higher levels of recursive mindreading over the course of their development from infant to adult.[54]

The informative intention is the intention that the speaker has to change the mental states of the listener; and the communicative intention is the speaker's intention that the listener believe that the speaker is trying to change their mental states. For communicators to recognize the intentions involved in ostensive-inferential communication, they must engage in the process of recursive mindreading – the Mary and Peter scenario illustrates how recursive mindreading looks. The "ostension" part of ostensive-inferential communication refers to the role that the speaker plays in conveying the communicative intention. The "inference" part refers to the listener's role, which is to draw inferences from the speaker's utterance in virtue of having recognized the speaker's informative and communicative intentions. Here is an example to drive

[54] See Sperber (1994).

this point home. Consider the sentence 'How are you feeling today?'. When the sentence is expressed (as an utterance – let us call this utterance '*X*'), we would expect that an appropriate response will be relevant in some respect to what the speaker meant by *X*. And the speaker could have intended several different meanings by *X* – that is, *X* has several possible interpretations. Here are just three examples.

> (a) The speaker says *X* to express genuine concern about the listener's state of health.
> (b) The speaker says *X* to mock the listener, as the sports team the speaker supports defeated the team the listener supports the previous night.
> (c) The speaker says *X* to indicate delight over the speaker's and listener's rowdy night out at a local bar.

There are actually an infinite number of possible utterances that the sentence 'How are you feeling today?' can be used to express.[55] According to the ostensive-inferential model, the speaker can provide evidence that they intend by expressing X to mean something in particular ("ostension"); correspondingly, the listener will use the evidence provided by the speaker in order to infer ("inference") what the speaker intends to mean. The speaker can provide evidence for which of (a) – (c) applies by, for instance, modulating their tone of voice, their facial expression, body language, and so on. They might say *X* with a sneer to suggest (b), or in an overly gleeful manner to suggest (c). The listener can use the evidence (a sneer, or an excited tone of voice) in order to infer what the speaker intends by the utterance. Listeners use the evidence provided by the speaker in order to identify the premises from which inferences can be made. In this manner, the listener will be able to make the best (i.e., most warranted) inferences to pinpoint the explicatures and implicature of the utterance. This is how ostensive-inferential communication works. A sophisticated metapsychology allows communicators to form increasingly higher-order metarepresentations of one another's mental states, and the cognitive

[55] Scott-Phillips (2015), 1.6.

mechanism of recursive mindreading allows communicators to reason about those metarepresentations.

CHAPTER TWO

A chatbot is a computer program that can communicate with humans in human natural language. In this chapter, I explain how chatbots work. Chatbots are made up of several different components. The language model component provides chatbots with the ability to produce and interpret utterances. I explain how language models work, and how chatbots can represent the meanings of words. I end this chapter by documenting the fact that chatbots communicate using only encoding and decoding processes – that is, that chatbots communicate within the paradigm of the code model of communication.

2.1. Terminology

To begin, I will introduce some of the terminology that will be used for the remainder of this thesis. A good resource on artificial intelligence is the popular textbook *Artificial Intelligence: A Modern Approach* (3rd Edition), by Stuart J. Russell and Peter Norvig (2010).[56] This is the resource from which much the terminology used in this chapter comes. Russell and Norvig (2010) define 'artificial intelligence' as "the study of agents that receive percepts from the environment and perform actions."[57] This might seem close to the definition of 'intelligence' more generally. However, artificial intelligence is actually a field of study, rather than a property that an agent (or computer) can have. That being said, it is sometimes the case that 'artificial intelligence' is used to refer to computer systems that are able to learn how to solve problems (what learning how to solve a problem entails will be discussed later in this chapter). However, I will not be using it in this sense – in this thesis, the term 'artificial intelligence' is strictly reserved to refer to the field of study.

[56] The fourth edition of *Artificial Intelligence: A Modern Approach* was published in 2020. The fourth edition includes some major revisions to the sections that are about how to actually construct and use artificial intelligence computer programs. However, the third edition is suitable for the purposes of this thesis. For more details about the fourth edition, see <http://aima.cs.berkeley.edu/>.
[57] Russell and Norvig (2010), Preface, page viii.

Let us unpack the definition of artificial intelligence a little more. An 'agent' is an entity that can perceive its environment using sensors, and act upon its environment. An example of an agent is a human being; humans use sense organs (eyes, ears, and so on) to perceive their environment, and humans use their bodies (arms, vocal organs, and so on) to modify their environment. A 'percept' is defined as an agent's perceptual input in any given instant.[58] For example, the redness of an apple is a percept for an individual at time *t* if it is the case that that individual "sees", or "experiences", the apple's redness at *t*. I will use the term 'agent program' to refer to an artificial agent (i.e., a computer, a robot, and so on). The actions that an agent program will perform are calculated using mathematical functions – these mathematical functions is usually referred to as the 'agent function' – that maps sequences of percepts to actions.[59] A sequence of percepts can be considered as an agent program's perceptual inputs over a period of time. An example of an agent function is an algorithm (i.e., a series of steps) that describes how to distinguish images of cats from images of dogs. A corresponding agent program could be a computer that, when presented with several images of cats and of dogs, distinguishes images of cats from images of dogs. I will be mentioning a couple of different types of algorithm that are used in agent programs, so it will be helpful to think of the agent function just as the algorithm that an agent follows to perform tasks.

2.2. Natural Language Processing, Natural Language Understanding, and Chatbots

Artificial intelligence is a massive field of study that benefits (and benefits from) other disciplines. Some of the important uses of artificial intelligence agent programs include solving problems that are difficult for humans, solving problems that advance human knowledge and

[58] Ibid., page 34.
[59] Ibid., page 35.

aide the research process, and solving problems that make humans' lives more convenient. For example, the Conseil Européen pour la Recherche Nucléaire (CERN) held a competition in 2015. The challenge CERN posed had to do with using artificial intelligence methods to study particle collisions at their supercollider in Geneva. Because of the nature of computer science research, open-source datasets (such as Kaggle), and code repositories (such as GitHub) that make it easy to collaborate with others on large-scale computer science programming projects, CERN was able to make use of the expertise of individuals from all around the world who were interested in putting their skills to the test in a real-world project.[60] Another interesting example of an application of artificial intelligence methods is in the medical field of radiology. Artificial intelligence image recognition agent programs have been used to help radiologists analyze x-ray results, MRI scans, and other sorts of medical images, in order to help doctors quickly identify pathologies in patients' medical image scans.[61]

There are two subfields of artificial intelligence that aim to develop agent programs to solve tasks that have to do with human natural language.[62] These are 'natural language processing' (henceforth 'NLP'), and 'natural language understanding' (henceforth 'NLU'). NLP is concerned with solving tasks having to do with the syntax of language. Common NLP tasks include preprocessing text, tokenization, parts-of-speech tagging, named entity recognition, and

[60] Adam-Bourdarios et al., (2014).
[61] Tang et al., (2018).
[62] The upcoming textbook *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd edition draft* (2020) by James H. Martin and Daniel Jurafsky is the main reference used in this chapter to explain how chatbot agent programs work. A draft of this textbook can be accessed for free here: <https://web.stanford.edu/~jurafsky/slp3/>. Generally speaking, there is a consensus about the principles that underlie how to construct agent programs (including agent programs that solve NLP and NLU tasks, such as chatbots). A lot of the research that goes on in the field of artificial intelligence is focussed on improving the mathematical/statistical/computational methods used to construct agent programs, rather than on refining concepts (of course, there are some exceptions, but this holds true for the most part). Important contributions by researchers that have led to changes in the methods used to construct agent programs have been mentioned where appropriate.

so on. NLP and NLU focus on different sorts of problems, but solving NLU problems requires the use of NLP techniques. NLU is concerned with building agent programs that solve problems having to do with semantics and linguistic communication. Common NLU tasks include the following.

> (i) Machine translation: Translating text/speech from one human natural language to another, such as from English to Korean and vice versa. Google Translate is an example of a machine translation system.
> (ii) Sentiment analysis: Gauging customers' feelings towards a company's products and services based on customer reviews. For example, the reviews that customers post on Amazon about a particular product can be used to gauge how satisfied customers are with that product.
> (iii) Conversational agents: Constructing an agent program that can have conversations with human beings using human natural language.

Conversational agents are the focus of this thesis – I will refer to such agent programs as 'chatbots'. A precise definition of what chatbots are is provided by Martin and Jurafsky (2020), who say "…chatbots are systems designed for extended conversations, set up to mimic the unstructured conversations or 'chats' characteristic of human-human interaction, mainly for entertainment, but also for practical purposes like making task-oriented agents more natural."[63] A chatbot is an agent program that can communicate with humans in human natural language. They can be used for a variety of tasks, such as providing information about flights, answering questions, or providing humans with social interaction through ordinary conversations, like one would have with a friend. Chatbots are complex, in the sense that they are made up of several different components. I have included a diagram below that outlines the three main components that make up Facebook's chatbot, BlenderBot.[64]

[63] Martin and Jurafsky (2020), page 493.
[64] See Stephen Roller et al., (2021).

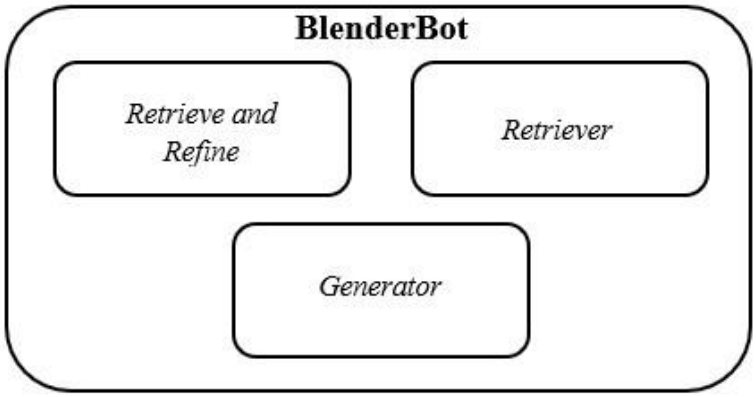

Figure 5

Each of the components of a chatbot performs a different task. For example, the Retriever component of BlenderBot accesses information from Wikipedia articles, allowing BlenderBot to answer questions that require specialized knowledge about some topic. A user could ask BlenderBot the question 'How does eutrophication affect the climate?'. BlenderBot's Retriever can use information from Wikipedia articles about eutrophication, fertilizers, and climate change in order to provide the user with an informative response to their question.

What I am interested in in this chapter is language models. These are computer programs that enable chatbots to produce utterances and interpret utterances – in human verbal communication, these are usually called speaking and utterance comprehension, respectively. While it is true that constructing language models is an NLP problem, language models are a component of chatbots, and constructing chatbots is an NLU problem. This is why language models are relevant for NLU research and chatbots. Specifically, I am interested in the following two aspects of language models that can affect how well a chatbot can communicate in human natural language.

> (i) The architecture of the language model: This is the way that the language model actually models human natural language. The architecture of a language model can influence how well the language model can generate linguistic expressions, and how similar the linguistic expressions it produces are to those a human would produce. The

architecture of the language model affects how they approach NLP and NLU tasks, including how a chatbot communicates using human natural language.

(ii) The word embeddings that the language model uses: Word embeddings are vector representations of words in human natural language. Word embeddings are supposed to be able to capture aspects of linguistic meaning. Word embeddings provide language models with the ability to represent the meanings of words. Thus, the meaningfulness of the linguistic expressions that a language model generates are influenced by the word embeddings that the language model uses to represent words in human natural language.

In the next two sections I explain what language models are in greater detail. Later, I will explain what word embeddings are and how they can be used to represent the meanings of words.

2.3. Language Models

A language model is a computer program that predicts what the next linguistic token is likely to be given the linguistic tokens that have already occurred in a sequence. By 'linguistic token' I just mean any piece of language, such as a letter of the alphabet, a word, a sentence, the text of a book, text of a collection of all of the Wikipedia articles published in French, and so on. In this thesis I will be talking mainly about language models that predict what words follow a sequence of words, but language models can actually make predictions about any sort of linguistic token. Predicting what linguistic tokens follow a sequence of linguistic tokens is useful for chatbots. Chatbots are agent programs that communicate with humans, which involves producing and interpreting utterances. A language model can be used to assign a probability to the words in the speaker's utterance – this is analogous to utterance comprehension in human communicators. The language model can then generate a response to the speaker's utterance by selecting the words that have the highest probability values (based on the probability values assigned to the words in the speaker's utterance) – this is analogous to utterance production in humans.

Language models use a corpus in order to learn how to make predictions about what linguistic tokens follow a sequence of linguistic tokens. A corpus is any piece of language data used to build a language model.[65] Often, corpuses can contain millions upon millions of linguistic tokens.[66] Formally, a language model $P(L)$ for a vocabulary L and words in L, x_1, x_2, …, x_n can be represented as follows.

$$P(L) = P(x_1, x_2, \ldots, x_n)$$

Language models calculate the conditional probability given a sequence of words x_1, x_2, …, x_n that the next word is $x_{(n+1)}$.

$$P(x_{(n+1)}|x_1, x_2, \ldots, x_n)$$

The formulas above illustrate the general principle that language models are a probability distribution over all of the sentences in a vocabulary.

Statistical techniques to analyze linguistic tokens were first employed by Markov (1913).[67] His insight was that the distribution of linguistic tokens in a corpus could not have been a matter of chance. Moreover, he thought that by actually analyzing the distribution of linguistic tokens in a corpus, it would be possible to calculate the probability distribution of linguistic tokens in the corpus.

> Accordingly, we assume the existence of an unknown constant probability p that the observed letter is a vowel. We determine the approximate value of p by observation, by counting all the vowels and consonants. Apart from p, we shall find – also through observation – the approximate values of two numbers p_1 and p_0, and four numbers $p_{1,1}$, $p_{1,0}$, $p_{0,1}$, and $p_{0,0}$. They represent the following probabilities: p_1 – a vowel follows another vowel; p_0 – a vowel follows a consonant; $p_{1,1}$ – a vowel follows two vowels; $p_{1,0}$ – a vowel follows a consonant that is preceded by a vowel; $p_{0,1}$ – a vowel follows a vowel

[65] Martin and Jurafsky (2020), pages 13 – 14.
[66] The Westbury Lab at the University of Alberta provides a corpus which consists of all of the Wikipedia articles published in English as of April 2010. It contains just over 990 million words. To access this corpus, see <http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html>. It is not uncommon for corpuses to be this large. For the kind of corpus used to train conversational agents (i.e., chatbots), see Rashkin et al (2019).
[67] This paper was translated to English in 2006, and it is the 2006 version I am citing.

> that is preceded by a consonant; and, finally, $p_{0,0}$ – a vowel follows two consonants. [Andrey Markov (2006), page 591]

The various combinations of letters that appear in a corpus can be assigned probability values based on how frequently the combinations occur in the corpus. Shannon and Weaver (1949) also analyzed text in order to figure out ways to calculate the likelihood that certain linguistic tokens will follow sequences of linguistic tokens. Shannon and Weaver (1949) found a method to generate words and sentences in English using the probabilities with which combinations of letters and words appear in a corpus. They call linguistic tokens "elementary symbols"; elementary symbols are concatenated to form words and sentences.[68] Elementary symbols can be symbols such as musical notes, numbers, and the like; in the case of generating linguistic expressions, the elementary symbols are linguistic tokens, such as letters of the alphabet, words, and so on.[69] The linguistic token to be concatenated to a sequence of linguistic tokens will depend on the linguistic tokens that have already been concatenated. This is the basic motivation behind language models – even the ones being used nowadays – and it can be traced to Markov (1913) and Shannon and Weaver (1949).

Language models can be thought of as a representation of human natural language, and they are what allow chatbots to produce and interpret utterances. The words that a language model represents come from the corpus with which it was trained. But language models can also assign probabilities to combinations of words (i.e., to sentences) that they have never encountered before. For this reason, they are useful for tasks that involve the production of linguistic expressions; these tasks are sometimes called NLG, or 'natural language generation'.[70]

[68] Warren Weaver (1953), page 10.
[69] Ibid., page 11.
[70] NLG can be contrasted with a more simplistic form of producing text, which involves pattern recognition. For more about how this works, see Martin and Jurafsky (2020), pages 498 – 500.

For example, suppose a corpus contains the following two sentences: ‘Let’s go to the planetarium’, ‘Let’s grab a bite later today’. A language model will be able to calculate the probability of the sentence ‘Let’s grab a bite at the planetarium’, even if this sentence does not appear in the corpus. Since language models can assign probabilities to combinations of words they have not encountered in the corpus, chatbots can use a language model to respond to utterances that the chatbot has never encountered. The language models commonly used nowadays incorporate neural networks – these language models are called ‘neural language models’. In the following section I will explain the fundamental principles behind how neural language models work.[71]

2.4. Neural Language Models

A neural language model is a language model that utilizes neural networks in order to predict what the next word is likely to be given a sequence of words.[72] A neural network is a computer program that is modelled on a structure analogous to that of the human brain. Neural networks are comprised of layers of nodes, with connections between the nodes in one layer to nodes in the previous and the next layers. This is loosely speaking also the structure of neurons, synapses, and axons in the human brain. The figures below illustrate the similarities between the representation of biological neurons and that of a neural network node.[73]

[71] There is another kind of language model, called the ‘n-gram language model’. N-gram language models do not use neural networks, but they perform the same task as neural language models. N-gram language models use Markov chains to predict what words/sentences follow a given sequence of words/sentences. N-gram language models generally do not perform as well as neural language models when it comes to solving tasks having to do with human natural language. For an explanation about why, see Martin and Jurafsky (2020), page 142 – 143.
[72] They were introduced in Bengio et al. (2003).
[73] https://plato.stanford.edu/entries/artificial-intelligence/neural-nets.html

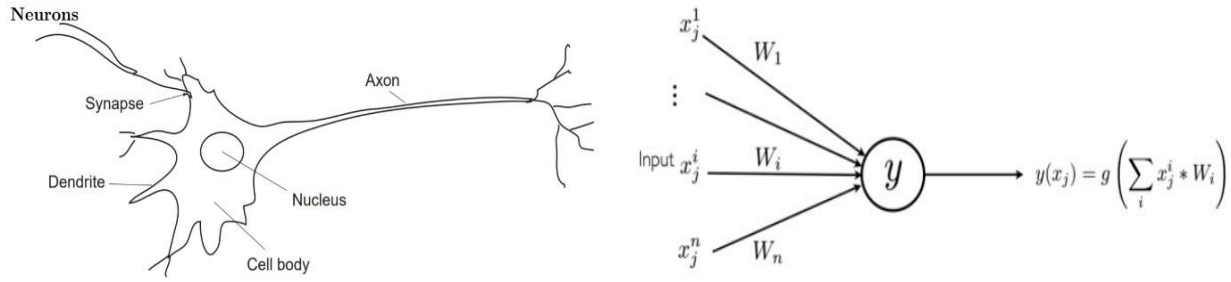

Figure 6

In order to make clear what the above diagram is showing, some of the formalism must be elucidated. The circle labelled '*y*' represents a neural network node. The inputs (i.e., percepts) are represented by 'x_j^1,…, x_j^i,… x_j^n'. I will explain what the inputs are for neural language models in more detail shortly, but usually neural network inputs are represented as vectors of real number values. The weights of the other nodes that are connected to *y* are represented by 'W_1, …, W_i, …, W_n'. The output of a node acts as the input for the next node. A neural language model is a neural network that consists of layers of nodes similar to the one shown below.[74]

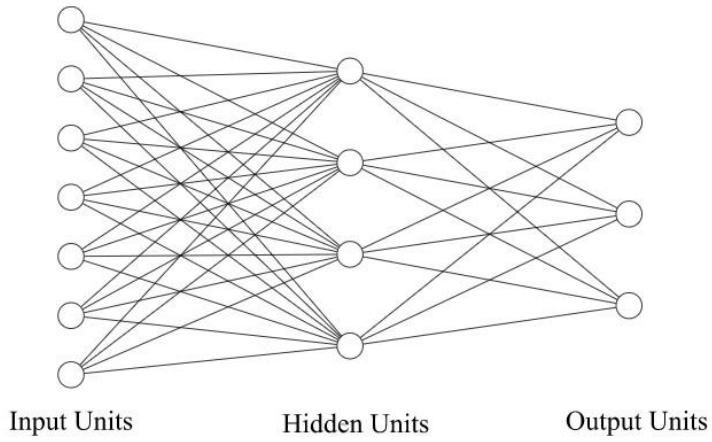

Figure 7

The structure of nodes and connections between nodes in a neural network is called the neural network's 'architecture'. The neural networks that are usually used in contemporary language

[74] Stanford Encyclopedia of Philosophy, article on 'Connectionism'.

models have architectures that are more complex than the one illustrated above. Without going into too much detail, I will mention two examples. The first is the 'Long Short-Term Memory' architecture (or 'LSTM'). The LSTM architecture was first introduced in Hochreiter and Schmidhuber (1997). LSTMs have loops in-between some of the layers of the neural network. The loops enable the neural network to mimic short-term memory. By mimicking short-term memory, LSTMs can represent information about the context in which words in the corpus occur, allowing it to make better predictions about which linguistic tokens come next given a sequence of linguistic tokens than neural language models that use traditional architectures without loops (like the one in figure 7).[75] The second example is also the newest neural network architecture that has been developed and used in language models; it is called the 'transformer', and was introduced by Vaswani et al. (2017). The transformer architecture is supposed to allow the neural network to capture features of the context in which the words appear more effectively than LSTM neural networks. A cutting-edge language model that uses the transformer architecture is GPT-3.[76]

A single neural language model can be composed of smaller neural language models. For example, Li Dong et al (2019) presented a large neural language model that is composed of four individual, smaller neural language models. Each of the four smaller neural language models tackles a different task. For example, one of the four smaller neural language models is a "Left-to-Right LM", which takes as input a sentence, which it parses from left to right (similar to how, for example, text in English is read); another is a "Right-to-Left LM", which parses sentences

[75] The context of a word is just the other words that are used in the same sentence/paragraph/conversation. I will say more about this in the next section. For a more contemporary look at LSTM language models, see Jozefowicz et al. (2016).
[76] Brown et al. (2020).

from right to left (such as in Arabic).[77] The reason why several neural language models are sometimes combined to form a larger neural language model is because the larger language model should be able to make better predictions about what words/sentences follow a given sequence than any of the smaller ones can individually.

The process of calculating the weight values of the nodes in a neural network is called 'machine learning', or 'training'. Machine learning involves exposing a neural network to data, and a machine learning algorithm tells the neural network how to identify patterns in the data.[78] Machine learning can be thought of as a way to teach a computer to find and recognize patterns in data. A computer can store those patterns as a model, and the model can be used to approximate phenomenon by using patterns in the data about that phenomenon. A datapoint can represent such things as images, such as an image of a cat; words and sentences, such as text of the word 'animal'; an audio clip, such as an audio recording of a guitar playing an A-major chord, and so on. A training dataset used to train a neural network to distinguish between images of cats and dogs will contain many images, some of cats, and others of dogs, whereas a training dataset used to train a neural network that when presented with an audio recording of a song can identify the genre of music that song belongs to will contain many audio recordings of songs from various genres. In other words, the training dataset used to train a neural network is task-specific – a neural network cannot be trained to distinguish between images of cats and images of dogs by being exposed to audio recordings. In general, the more data that is used to train a

[77] Dong et al (2019), page 2.
Note that all neural language models are doing is predicting what the next word is given some sequence of words. It does not matter whether the neural language model is reading a sentence from left to right or from right to left, nor does it matter whether the sentences the neural language model is reading are in English, Arabic, or whatever other human natural language. For more about this, see Dong et al (2019), page 4.
[78] Machine learning is useful to assign values for the weights of neural networks. However, machine learning is a general method to make a computer learn to identify patterns in data. For more about how machine learning works and how it is used, see this blog post: < https://medium.com/@lizziedotdev/lets-talk-about-machine-learning-ddca914e9dd1>.

neural network, the better the accuracy of that network in correctly classifying unseen datapoints. This is just like how the more practice a budding art critic has correctly identifying Van Gogh paintings, the more likely it is that she will correctly identify an unfamiliar piece of artwork as a Van Gogh.

In much the same way, a language model can be considered as using patterns in a corpus to identify aspects of linguistic meaning. The language model can then be used to approximate phenomenon having to do with linguistic meaning, such as linguistic communication. Neural language models use language data (often in the form of text), which is called a 'corpus', to learn how to recognize patterns about linguistic tokens.[79] These patterns include things like grammar, how words are spelled, and how frequently certain words are used in a document. Neural language models use a machine learning algorithm to identify patterns in the corpus. Generally speaking, there are three kinds of machine learning algorithm: supervised learning algorithms, unsupervised learning algorithms, and reinforcement learning algorithms (but I will not discuss how reinforcement learning works, since it is not as common in NLP/NLU as the other two kinds of machine learning).[80] Martin and Jurafsky (2020) define the process of supervised machine learning as "In supervised machine learning, we have a data set of input observations, each associated with some correct output (a 'supervision signal'). The goal of the algorithm is to learn how to map from a new observation to a correct output."[81] Supervised machine learning involves exposing a neural language model to a "labelled" corpus; the label is the supervision signal. The

[79] Martin and Jurafsky (2020), pages 500 and 501.
[80] For a survey of reinforcement learning techniques, see Cetina et al. (2020).
[81] Martin and Jurafsky (2020), page 56.

table below illustrates part of a labelled corpus consisting of question-answer pairs. Correct answers are labelled with a '1', and incorrect answers are labelled with a '0'.[82]

Question	Answer	Label
'what bird family is the owl'	Most are solitary and nocturnal , with some exceptions (e.g. , the Northern Hawk Owl) .	0
'what bird family is the owl'	Owls are characterized by their small beaks and wide faces, and are divided into two families: the typical owls , Strigidae ; and the barn-owls , Tytonidae .	0
'what bird family is the owl'	Owls are a group of birds that belong to the order Strigiformes , constituting 200 extant bird of prey species .	1

Figure 8

Labelling the corpus allows the neural language model to learn what correct answers look like and what incorrect answers look like. This way, a neural language model should be able to recognize both the kinds of outputs that are correct versus the kinds that are incorrect. On the other hand, unsupervised machine learning involves exposing a neural language model to an "unlabelled" corpus. Unsupervised machine learning is useful when there is no obvious way to label data. For example, text data which consists of conversations between people on an online forum can be useful for training a language model for a chatbot. Unlike the question-answer dataset, however, there is no clear criterion that suggests how to label conversation data. Unsupervised machine learning allows the computer to find patterns in the data on its own. The patterns that the computer finds in the corpus are actually patterns which provide insights about the meanings of words. In the next section I explain how language models represent the meaning of words.

[82] This corpus is open-source, and can be found here: < https://www.microsoft.com/en-us/download/confirmation.aspx?id=52419>.

2.5. Linguistic Meaning from Vector Representations of Words

Corpuses consist of linguistic tokens in human natural language; however, computers cannot read human natural language – the English alphabet makes no sense to a computer. However, computers are really good at doing lots of complex calculations on numbers, and quickly. Because of this, words in a corpus are first represented as vectors of real number values before they can be used to train a neural language model. These vectors are often called word embeddings (or simply 'embeddings).[83] Word embeddings are crucial for building chatbots because embeddings can represent aspects of linguistic meaning. This is largely due to the linguistic theory that motivates how word embeddings are constructed: The distributional hypothesis.[84] The distributional hypothesis states that the meaning of any given word can be approximated by looking at the meanings of words that commonly occur in similar contexts. The context in which a word appears is just the other words that appear spatiotemporally near it. The context is made up of only linguistic tokens, which is different from how the term "context" is often used in the philosophy of language literature. Often, philosophers take the context to be not just the sentences uttered, but also the place in which they are uttered, who the speaker and the listener are, the tone of voice with which utterances are made, and more. In this sense, the context can be considered as a psychological construct that includes both facts about the language, syntax, and grammar, as well as facts about things outside of language.[85] However, in computer science, the context is often taken to be comprised of only linguistic facts which come from the sentences in a corpus.

The distributional hypothesis was developed by Zellig Harris (1954). He claims that the distribution of words can yield insights about the structure of language and meaning.

[83] Martin and Jurafsky (2020), page 100.
[84] Ibid., page 96.
[85] Sperber and Wilson (1995), page 15 – 16.

> …we will often find interesting distributional relations, relations which tell us something about the occurrence of elements and which correlate with some aspect of meaning. In certain important cases it will even prove possible to state certain aspects of meaning as functions of measurable distributional relations... The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. For it is not merely that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. [Zellig Harris (1954), page 156]

Let us unpack the example Zellig provides of three words, ‘A’, ‘B’, and ‘C’ a bit. The following two combinations can be made from those words: ‘AB’ and ‘AC’, where each of these combinations has a different meaning from the other. Since ‘AB’ means something different from ‘AC’, then, according to Zellig, we can expect that ‘A’ and ‘B’ tend to occur in contexts that are different from those in which ‘A’ and ‘C’ occur. Zellig says that while the distribution of words in a corpus can provide clues about the meanings of words, the distribution of a word is not itself the word’s meaning. For example, the word ‘pizza’ is likely to occur in the context of words such as ‘cheese’, ‘topping’, and ‘delicious’; but ‘pizza’ is (usually) not so likely to occur in the context of words such as ‘chair’, ‘banister’, and ‘roof’. According to the distributional hypothesis, the meaning of ‘pizza’ is more similar to the meanings of ‘cheese’, ‘topping’, and ‘delicious’ than it is to the meanings of ‘chair’, ‘banister’, and ‘roof’. On the same token, clues about how similar the meanings of two words are can be found from their distribution in a corpus. Another important idea from the above quote is that a difference in meaning is correlated with a difference of distribution. However, Zellig (1954) says that the syntax and grammatical rules of a human natural language cannot on their own give rise to meaning.

> However, this is not the same thing as saying that the distributional structure of language (phonology, morphology, and at most a small amount of discourse structure) conforms in some one-to-one way with some independently discoverable structure of meaning. If one

> wishes to speak of language as existing in some sense on two planes – of form and of meaning – we can at least say that the structures of the two are not identical, though they will be found similar in various respects. [Zellig Harris (1954), page 152]

Words with similar meanings tend to cooccur, but the fact that a certain combination of words tends to cooccur more frequently than another combination of words does not also indicate what each of those words themselves mean. The distributional hypothesis was also developed by Firth (1957). He says that the way that words are written in a piece of text can provide indications of their meanings.[86] The term Firth (1957) uses to refer to the distribution of words in a text is 'collocation'. The idea that the distribution of words can yield insights about meaning has been embraced by NLP/NLU research, and is reflected in its methodology.[87] Two words that appear in the same contexts in a corpus will tend to have similar meanings, whereas two words that do not occur in the same contexts will not have similar meanings.

The distribution of words in a corpus is used to construct word embeddings. In order to show how word embeddings actually represent the meanings of words, and how these word embeddings look, I will explain the basics of a common technique to obtain word embeddings from a corpus, called 'Global Vectors', or 'GloVe'. GloVe was introduced by Pennington, Socher, and Manning (2014). GloVe represents words in human natural language as vectors of real-number values. Below I have included part of a GloVe word embedding for the word 'first', which was obtained from the 'Wikipedia2014 + Gigaword 5' corpus. The 'Wikipedia2014 +

[86] J. R. Firth (1957), page 7.
[87] In the NLP/NLU literature, Zellig (1954) and Firth (1957) are often mentioned as providing the motivation for using a corpus to obtain word embeddings.

Gigaword 5' corpus combines language data from the Wikipedia2014 and Gigaword 5 corpuses.[88]

> 'first' = [-0.464336, 0.148044, -0.089796, 0.243946, -0.306616, …, -0.077351, 0.399929, -0.263756, -0.098782, 0.451579]

The word embedding for 'first' (and the other words in that corpus) are vectors, and each of these vectors contains 300 numerical values. Each of these values is called a 'parameter', and each parameter is supposed to represent a different aspect of the meanings of the words in the corpus. The number of parameters is arbitrary and selected by the programmer, but research suggests that including more than 200 parameters does not lead to any appreciable improvements in the language model's performance.[89] Here are the GloVe word embeddings for the first five words in the Wikipedia2014 + Gigaword5 corpus. I have included the first 5 parameter values for each of the word embeddings.

> 'also' = [-0.158810, 0.590201, -0.356879, 0.282293, -0.222408, …]
> 'first' = [-0.464336, 0.148044, -0.089796, 0.243946, -0.306616, …]
> 'one' = [-0.319227, 0.580684, 0.008368, 0.623442, 0.376866, …]
> 'year' = [-0.161512, 0.831497, -0.541194, 0.755819, 0.802536, …]
> 'use' = [-0.257498, 0.884452, -0.751096, 0.198309, -0.467311, …]

Partly because of the inherent complexity of linguistic meaning and the size of the corpuses that are used to build language models, it is practically impossible to ascertain what precise aspect of meaning each of those parameters is actually representing.[90] What GloVe tries to do is represent all of those various aspects of the meanings of words at the same time.[91] Another reason why it is

[88] The 'Wikipedia2014 + Gigaword 5' corpus combines language data from the Wikipedia2014 and Gigaword 5 corpuses. See: < https://nlp.stanford.edu/projects/glove/>. For GitHub page for the computer code: <https://github.com/stanfordnlp/GloVe>.
[89] This is true in the case of GloVe embeddings. See Pennington, Socher, and Manning (2014), section 4.4.
[90] <https://nlp.stanford.edu/projects/glove/>. "Linear Substructures", Section 2.
[91] This is a common feature of word embeddings. For example, Word2Vec, a popular word embedding introduced by Mikolov (2013), also represents each word as a unique vector. Just like in the case of Glove, the Word2Vec vectors represent several different aspects of word meaning at the same time. Neelakantan et al. (2015) modified Word2Vec embeddings in order to account for the fact that a single word can have several different meanings.

almost impossible to figure out which aspects of word meaning GloVe word embeddings represent is because GloVe is an unsupervised machine learning algorithm. All the programmer needs to specify is which corpus will be used with GloVe to obtain word embeddings, and the number of parameters GloVe should consider when constructing the embeddings. It is unclear what aspects of the meanings of words that GloVe word embeddings actually represent because the corpus is unlabelled, and thus the GloVe algorithm tries to figure out what patterns are latent in the corpus that can provide clues to the meanings of words in the corpus.[92] While it is not obvious is what aspect of meaning each of the parameters of a GloVe word embedding represent, what is important to note is that GloVe word embeddings are constructed based on the distributions of words in a corpus.

2.6. Chatbots and the Code Model of Communication

In this section I will document the fact that chatbots communicate using only encoding and decoding processes – that is, chatbots communicate within the paradigm of the code model. There is evidence in the literature which suggests this. Code model communication involves a speaker encoding thoughts into language, and a listener decoding language into thoughts. In ordinary code model communication, the speaker encodes the thought that they want to communicate into language – this is called encoding. When the listener hears the utterance, she decodes the utterance by using a code (such as a generative grammar) to map the phonemes onto a mental representation. Successful code model communication happens when the mental

[92] Since I am focussing on word embeddings, I must mention that useful insights about the meanings of words can also be gleaned from metrics such as the distance between the vectors that represent them. There are several different ways to measure the distance between word vectors; these are called 'distance metrics'. Two popular distance metrics are Euclidean distance and Cosine distance. Euclidean distance measures the distance between two points in a vector space; if the distance between the two points is small, then the words are similar in meaning. Cosine distance provides a way to determine how similar the meanings of words are by measuring the size of the angle between their vectors. A small angle between two vectors indicates that the words they represent have similar meanings. For more about this, see Martin and Jurafsky (2020), Chapter 6.

representation that the listener recovers after the decoding process is similar to the one that the speaker had encoded.[93] Chatbots engage in encoding and decoding processes that are analogous to those that human communicators engage in. However, encoding and decoding in chatbots look a little different than they do in human beings. The language generation of chatbot agent programs is modelled as an encoder-decoder problem, where an "encoder" neural network encodes a speaker's utterance in human natural language as word embeddings, and a "decoder" neural network decodes word embeddings as words in human natural language. The following diagram is provided by Martin and Jurafsky (2020) illustrating the basic idea behind the encoder-decoder model.[94]

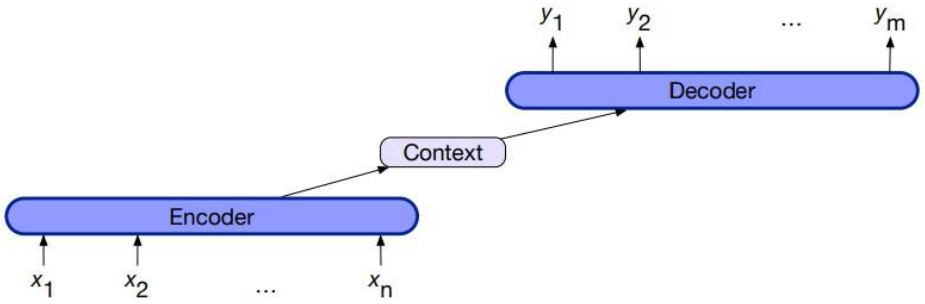

Figure 9

About the encoder-decoder architecture, Martin and Jurafsky (2020) say "The key idea underlying these networks is the use of an encoder network that takes an input sequence and creates a contextualized representation of it, often called the context. This representation is then passed to a decoder which generates a task-specific output sequence."[95]. In order to explain what the encoding and decoding processes are like in chatbots, let us imagine a conversation between

[93] The terminology of 'encoder'/'encoding' and 'decoder'/'decoding' was popularized by Shannon and Weaver (1949); it has since been adopted by linguists and philosophers of language to refer to code model communication. See Blackburn (2007), page 74.
[94] Martin and Jurafsky (2020), page 208.
[95] Ibid. It must be noted that in that quote, the 'context' of a word is understood as the other words that appear in the same sentence/input. This is similar to how Zellig (1954) and Firth (1957) use the term.

a chatbot, Roberta, and a human, Sally. Suppose Sally says to Roberta ‘How are you feeling today?’. Based on the explanations above of how language models generate linguistic expressions and how words are represented as vectors, the process of utterance interpretation and utterance production that Roberta engages in will look something like this.

1. Human natural language input sequence (the sentence Sally utters): ‘How are you feeling today?’
2. Word embeddings that represent the sentence Sally utters: ‘x_1’, ‘x_2’, ‘x_3’, ‘x_4’, ‘x_5’
2. Encoder language model input: $[x_1 + x_2 + x_3 + x_4 + x_5]$
3. Contextualized representation of the input: $R([x_1 + x_2 + x_3 + x_4 + x_5])$
4. Decoder language model output: $[y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7]$
5. Human natural language output sequence (Roberta’s utterance): ‘I am doing great! How are you?’

The process Roberta engages that I just described captures the fundamental, relevant steps that Roberta takes when communicating with Sally. Sally’s utterance is a linguistic expression in human natural language, which Roberta must convert into a series of word embeddings. This is because computers cannot read human natural language. Moreover, word embeddings are supposed to represent the meanings of words, which is useful for the task of utterance interpretation. The sentence Sally utters can be broken down into its individual words: ‘how’, ‘are’, ‘you’, ‘feeling’, ‘today’. Each of those words will have an associated word embedding (provided Roberta has been trained using a corpus that includes all of those words – machine learning algorithms to obtain word embeddings from a corpus include GloVe and Word2Vec). The word embeddings are used as the input for the encoder language model that Roberta relies on for linguistic competence. The encoder performs some calculations on those word embeddings and concatenates them to create a contextualized representation of the sentence Sally uttered. The contextualized representation of the input embedding is useful for the chatbot because it captures contextual aspects of the meanings of the words in the sentence that Sally uttered. This is how the encoder-decoder architecture models communication as being encoding

and decoding processes.[96] In the next chapter, I argue that the fact that chatbots model communication as having to do with only encoding and decoding processes can explain why they are often poor conversationalists.

CHAPTER THREE

In this chapter I explain how the fact that chatbots communicate using only encoding and decoding processes can help explain why chatbots often cannot communicate effectively in human natural language. The poor conversational abilities of chatbots are a result of the fact that chatbots only access linguistic context, whereas listeners need access to non-linguistic context to be able to grasp utterance meaning. I also consider whether chatbots are able to make inferences about non-linguistic properties of context at all. I claim they cannot, precisely because the neural language models that they rely upon for their linguistic competence are natural codes that merely associate percepts with output behaviours using encoding and decoding processes.

3.1. Introducing the Problem

Here is a brief recap of the main ideas I discussed in the previous chapter. Chatbot agent programs are made up of several different components. A language model is one of these components, and it is the language model that allows the chatbot to produce and interpret utterances. In this way, the language model component can be thought of as providing the chatbot linguistic competence. The utterances that chatbots produce are actually sequences of linguistic tokens that are concatenated together by the language model. In chatbot communication, the linguistic tokens are usually letters or words; these can be concatenated together to form words and sentences. The concatenation of linguistic tokens is a probabilistic process, as the linguistic token that will be concatenated next in a given sequence of linguistic

[96] Scott-Phillips (2015) observes that the kinds of codes used in computational models of communication systems are instances of a probabilistic code model. It is probabilistic in the sense that the generation of sequences of linguistic tokens is a matter of which tokens have already been generated (for instance, in the course of a conversation, some sentences will have already been uttered; the sentences that have been uttered will constrain which sentences will be uttered next). This is still an instance of code model communication, since "production [of sequences of linguistic tokens] and reception [of sequences of linguistic tokens] are still associative, albeit in a probabilistic way' (page 26, (1.2)).

tokens will depend on which linguistic tokens have already been concatenated in that sequence.[97] Nowadays, it is common to construct language models using neural networks.[98] Neural language models utilize one or more neural networks in order to predict what the next linguistic token is, given some sequence of linguistic tokens. Neural language models represent words as vectors of real-number values – these vectors are called word embeddings. Each numerical value of a word embedding is supposed to represent some aspect of the meaning of the word that the embedding represents. It is practically impossible to figure out precisely what aspect of meaning each value of an embedding is actually representing because of how complex linguistic meaning is, and how complex the algorithms used to obtain word embeddings are. The input for a neural language model will be a vector representation of the sentence that the speaker utters, and the output will be a probability distribution over the words in the chatbot's vocabulary. The words with the highest probability will be concatenated together to form a sentence that the chatbot can utter in response to the speaker.

How well a chatbot can converse in human natural language depends to a large extent on the architecture of the language model(s) it uses, and on the word embeddings the language model(s) use to generate linguistic expressions. The architecture of the language model partly determines how well a chatbot can take the context of a speaker's utterance into account when generating series of linguistic tokens. There are several different senses of context, so it will be worth stating the ones relevant for our discussion. Sperber and Wilson (1995) say that context is a set of premises that can be used to interpret the meaning of an utterance – context is a

[97] Consider the sequence of linguistic tokens 'I took my dog to the'. The word that comes next will be a function of the words 'I', 'took', 'my', 'dog', 'to', 'the'. We can reasonably suppose that the next word might be 'park'; it is less likely to be 'supermarket'. So a neural language model will be more likely to concatenate 'park' than 'supermarket' to 'I took my dog to the'.

[98] N-gram language models used to be popular, and are still in use today, but they fell out of favour because language models that use neural networks often perform better. See Martin and Jurafsky (2020), page 142 – 143.

'psychological construct' that can include not only the speaker's and listener's beliefs about the immediate environment in which an utterance is made, but also their beliefs about how the physics of the world works, personal life experiences, cultural assumptions, and more.[99] On the other hand, the NLP and NLU literature often consider context to be the linguistic tokens that appear spatiotemporally near a word in a corpus. The context in which a word is used is determined by the other words that appear spatiotemporally near it. For instance, in the sentence 'The cat is on the yellow striped mat', the context of the word 'yellow' includes the words 'the' 'cat' 'is' 'on', 'the', striped', 'mat'. Even though the term 'context' is used in the NLP and NLU literature to refer to something that seems to have more to do with the distribution of words rather than with meaning, it is assumed by the methodology employed to obtain word embeddings from a corpus that the distribution of words in a corpus can yield insights about what words mean. This is because the methodology is guided by the distributional hypothesis, which states that the meaning of a word can be approximated by looking at the meanings of words that are commonly used with it.[100] So there are two kinds of context: the first is a psychological construct having to do with non-linguistic properties, and the second has to do with the distribution of words in a corpus and their linguistic properties.

In the previous chapter, I documented the fact that chatbots communicate using only encoding and decoding processes. In this chapter I explain how the fact that chatbots communicate using only encoding and decoding processes can help explain why chatbots often cannot communicate effectively in human natural language. My line of reasoning is as follows. Neural language models only process the linguistic properties of linguistic expressions in corpuses during training. Thus, neural language models only process the linguistic properties of

[99] Sperber and Wilson (1995), page 15 – 16.
[100] Zellig (1954) and Firth (1957) are often credited with having come up with the distributional hypothesis.

uttered sentences. However, the non-linguistic properties of utterances are required for utterance comprehension. This is because the non-linguistic properties of utterances contribute to utterance meaning. So the poor conversational abilities of chatbots are a result of the fact that chatbots only access linguistic context, whereas listeners need access to non-linguistic context to be able to grasp utterance meaning. I also consider whether chatbots are able to make inferences about non-linguistic properties of context at all. I claim they cannot, precisely because the neural language models that they rely upon for their linguistic competence are natural codes that merely associate percepts with output behaviours using encoding and decoding processes. In human linguistic communication, making inferences about the non-linguistic properties of utterances requires a metapsychology and recursive mindreading, of the sort espoused by Sperber and Wilson (1995) and Scott-Phillips (2015). Chatbots have neither the relevant sort of metapsychology nor a mechanism to engage in recursive mindreading required to be able to make inferences about non-linguistic properties of utterances. I will develop this argument over the course of this chapter.

Much of the research that aims to improve chatbot performance often focusses on improving the methods used to construct language models. This usually entails finding the best language model architectures that can capture contextual elements (which should lead to better performance from NLU systems like chatbots), finding the best ways to obtain word embeddings from corpuses such that the embeddings capture useful information about the meanings of words, finding clever ways to optimize limited computational resources, and so on.[101] While there have been some significant successes in the fields of NLP and NLU that have resulted from the sort of

[101] Notable examples of this sort of research includes: Hochreiter and Schmidhuber (1997), Bengio et al (2003), Mikolov et al (2013), Pennington et al (2015), Vaswani et al (2017), Brown et al (2020).

methodological research I just described, I think it will be worth taking a closer look at whether the methodology rests on a solid theoretical foundation.[102]

3.2. Language Data and Linguistic Meaning

I would like to begin this section with the observation that a corpuses consist of sentences that represent utterances.[103] Neural language models use machine learning algorithms to identify patterns about linguistic tokens in corpuses, which the neural language model can then mimic and generate sequences of linguistic tokens (i.e., sentences) in human natural language. These patterns can include such things as identifying whether a word is a noun or a verb, identifying relationships between adjectives and nouns, and more. Fundamentally, corpuses consist of sentences, and sentences have only linguistic properties; as such, neural language models only process sentences, and, by extension, only linguistic properties. They cannot process non-linguistic properties at all. The non-linguistic properties contribute to utterance meaning, however, so it seems that while neural language models provide chatbots with access to linguistic properties of sentences, a chatbot cannot use neural language models to access the non-linguistic properties of utterances. Chatbots will thus not be able to access the non-linguistic properties of a speaker's utterance when engaging in utterance interpretation. Since chatbots are supposed to be able to communicate with human beings in human natural language, and since

[102] In what I think is a very compelling paper, Levesque (2014) says that a lot of the progress seen in artificial intelligence research aimed at building agent programs that can behave like human beings is a result of 'cheap tricks' (Levesque (2014), section 2.2). He looks specifically at the task of building an agent program that can answer questions like human beings can, which is a task that is very relevant for building chatbot agent programs.

[103] For instance, researchers at Cornell University made several corpuses available online through their social conversations analyzer toolkit, 'ConvoKit'. One of the corpuses the researchers provide consists of conversations between users on the social media platform Reddit. The datapoints in this corpus are users' comments and users' replies to comments. This corpus is considered to be a collection of utterances that preserves the natural conversational structure of the interactions of Reddit users. See Chang et al., (2020). To access the corpus I am referring to, see: < https://convokit.cornell.edu/documentation/reddit-small.html>.

human communication involves the production and interpretation of utterances, chatbots will not be effective communicators.[104]

I think a succinct way to state what I am saying here is that neural language models collapse the sentence-utterance distinction. On the one hand, neural language models are exposed to a corpus consisting of sentences. The neural language model identifies patterns in the corpus, which allows it to then generate sequences of linguistic tokens. On the other hand, chatbots communicate using human natural language, and communication has to do with utterances, not sentences. A chatbot that cannot access non-linguistic properties of a speaker's utterance will not be able to identify explicatures and implicatures. This means that chatbots will often exhibit what human communicators perceive of as a lack of understanding of what utterances mean.

Bender and Koller (2020) offer a criticism of language models that builds on similar observations: Neural language models are trained on only linguistic form and so cannot learn utterance meaning. The "form" of a word is the particular marks and symbols that make up the word. For example, the form of the word 'cat' is the series of straight and curved lines that make up the letters 'c', 'a', 't'. Commonly, a distinction is drawn between form and meaning; usually, matters of syntax have to do with form, whereas matters of semantics have to do with meaning.[105] Bender and Koller (2020) define 'meaning' as follows: "We take *meaning* to be the

[104] I had made this point in the first chapter. According to Sperber and Wilson (1995), utterances have both explicatures and an implicature. Listeners make inferences about the non-linguistic properties of an utterance in order to identify explicatures and the implicature. An utterance is a sentence that is expressed at a certain point in time. The truth value of the sentence can only be determined after identifying the explicatures, which requires making inferences about the non-linguistic properties of the utterance. See Sperber and Wilson (1993), page 5.

[105] The distinction between form and meaning has been developed in several important philosophical papers. For example, Searle (1980) proposed the Chinese Room thought experiment. A simple statement of what Searle's Chinese room is supposed to illustrate is that the mere manipulation of symbols does not constitute understanding what those symbols mean; this claim is supposed to hold true even if the manipulation of those symbols is entirely fluent, or error-free. Searle's Chinese Room is meant to illustrate the implausibility of "strong artificial intelligence", or, to put it differently, the implausibility of agent programs that have meaningful mental states. The Chinese room demonstrates how an agent can pass the Turing Test without understanding language. Some philosophers have tried to collapse the distinction between syntax and semantics by claiming that meaning can arise

relation $M \subseteq E \times I$ which contains pairs (e, i) of natural language expressions e and the communicative intents i they can be used to evoke. Given this definition of meaning, we can now use *understand* to refer to the process of retrieving i given e".[106] The kind of meaning that Bender and Koller (2020) are discussing in that quote is utterance meaning. This is because they see meaning as the recovery of the speaker's communicative intents, and communicative intents are about entities outside of language. For example, a speaker who says 'Open the window!' will have a communicative intent directed at making the listener to perform a particular action.[107] The content of their communicative intent is about the world in which the speaker and the listener inhabit, about windows that are closed and that can be opened, and so on; the speaker's communicative intent is not about the words themselves – 'open', 'the', 'window' – that make up their utterance. In order for a listener to recover the speaker's communicative intent, the listener will need to make inferences about entities in the real world – entities outside of language. A possible interpretation of an utterance of the sentence 'She treated me well' is that the speaker is sad that an old car which they loved finally broke down.[108] It is not obvious from the form of 'She' what entity is being referred to by an utterance of the word. Moreover, the form of the words that make up 'She treated me well' does not yield much insight about what

from the manipulations of symbols – that is, that syntax can give rise to semantics. For more, see Searle (1980) and Rapaport (1986).

[106] Bender and Koller (2020), page 3.

[107] Ibid.

[108] This example is of an implicature of 'She treated me well'. But the form of 'She' does not allow listeners to narrow down on explicatures of 'She treated me well' either. For instance, it is still not obvious which entity 'She' refers to from form alone. There is nothing about marks on a page that can allow a listener to know what the marks actually mean. Perhaps someone who knows the English language will also know that 'She' is a pronoun, and that these kinds of pronouns often refer to individuals identifying as female. However, even facts about what kinds of entities pronouns usually refer to can allow a listener to figure out which entity 'She' refers to when uttered by a speaker in the sentence 'She treated me well'. Listeners identify explicatures of an uttered sentence by making inferences about the non-linguistic properties of the utterance, such as which entities the words uttered refer to.

some who utters that sentence might intend to mean by it. But Bender and Koller (2020) also say that a language model trained on only form cannot learn sentence meaning either.[109]

At this point, it will be helpful to place the position presented by Bender and Koller (2020) more firmly within the context of our discussion about linguistic communication and chatbot neural language models. The form of words that make up a sentence is closely related to the linguistic properties of a sentence. The linguistic properties of a sentence are properties about the syntax and grammar of the sentence, and the form of a sentence allows us to identify the linguistic properties of the sentence. The form of the sentence 'She treated me well' offers information about the linguistic properties of the sentence, such as the fact that the sentence contains a pronoun 'she', who is the subject of the sentence. A major supposition of how neural language models learn from corpora is that form can provide clues about meaning – by training a language model to process text, it should be able to identify patterns that provide some indication of linguistic meaning. However, human communication involves the production and interpretation of utterances, and utterances have both linguistic and non-linguistic properties. Non-linguistic properties of utterances are properties that affect utterance meaning, but that are not communicated linguistically. The way that the non-linguistic properties of an utterance affect utterance meaning is that they provide listeners with clues about which inferences to draw to identify explicatures and implicatures of utterances. Similarly, a chatbot which can make

[109] Bender and Koller (2020), page 3. Sahlgren and Carlsson (2021) disagree, and claim that current critiques of language models – including the one levelled by Bender and Koller (2020) – maintain the incorrect assumption that there is only one correct way to define 'meaning'. Sahlgren and Carlsson (2021) are correct, in that there certainly are several ways to define 'meaning'. There is sentence meaning, which corresponds to the meaning of sentences. Sentence meaning is a function of the linguistic properties of sentences. There is also utterance meaning, which is a function of both the linguistic and non-linguistic properties of utterances. Utterance meaning involves identifying the explicatures and the implicatures of a speaker's utterance. Because communication has to do with utterances, I think we can assume that it is utterance meaning that chatbots must be able to grasp in order for them to be able to communicate effectively with humans in human natural language. In their paper, Bender and Koller (2020) are concerned with both sentence meaning and utterance meaning (see page 3), so their argument is relevant.

inferences about the non-linguistic properties of utterances will potentially be able to identify explicatures and implicatures. Neural language models do not access anything other the linguistic properties of sentences uttered because they are trained only on form. So a chatbot that relies on neural language models for its linguistic competence will only be able to process the linguistic properties of uttered sentences, since neural language models are trained only on form.

3.3. Neural Language Models are Natural Codes; Human Languages are Conventional Codes

I have tried to show that a neural language model that is trained on only language data will only be able to process the linguistic properties of utterances. As a result, chatbots cannot access non-linguistic properties of utterances, which are necessary in order to interpret utterance meaning. A question we can now ask is whether there is anything about the architecture of neural language models that also makes it the case that chatbots that rely on neural language models for their linguistic competence cannot access non-linguistic properties of utterances. I think that there is: Neural language models are natural codes, whereas human natural languages are conventional codes. Because neural language models are natural codes, they cannot make inferences about the non-linguistic properties of utterances.

To unpack this, the distinction between natural codes and conventional codes is crucial. Both natural codes and conventional codes are sets of associations between two types of behaviour: signals, and responses. A signal is a behaviour that conveys information about the signaller, or about the environment, and a response is a behaviour that has been evolutionarily adapted to correspond to a signal. Signals and responses can emerge through one of three routes: ritualization, sensory manipulation, and the direct route. In the case of all three of these routes, both signals and responses are established because they are behaviours that when paired together

are advantageous for an organism's survival.[110] A natural code can be considered as a mere association between percepts and output behaviours. On the other hand, a conventional code is a set of associations between signals and responses where the signals have been established through the direct route. Signals that emerge through the direct route are distinct from signals that emerge through either ritualization or sensory manipulation, as signals that emerge through the direct route "signal their own signalhood". That is, these signals require "that an interdependent pair of behaviours (a signal and a corresponding response) come into existence *simultaneously*".[111] Signallers need to have been able to make their intention *to* communicate a certain content apparent to the recipient of the signal (this is the informative intention). Signallers also need to have been able to intend (this is the communicative intention) to have the recipient recognize *that* the signaller has an informative intention. If signallers are able to do both of those things when performing the behaviour that is to serve as a signal, then that signal is one which signals its own signalhood.

An example of what a signal established through the direct route looks like will make the discussion so far clearer. Suppose Sally is sitting at the bar, and she tilts her cup at Bill, the bartender. Presumably, Bill will be able to infer from the signal that Sally produces that she wants another drink. Now, suppose that Sally performs the same tilt of her cup toward Bill, but this time she also raises an eyebrow and smiles. In this case, Bill might infer something different about what Sally wants to communicate based on her behaviour – perhaps that Sally enjoyed her drink. The examples with Sally and Bill illustrate how a signal can signals its own signalhood – the tilt of the cup communicates the fact *that* Sally wants to communicate (i.e., the tilt of the cup provides evidence about the signaller's communicative intention). And the tilt of the cup is a

[110] Scott-Phillips (2015), section 2.3.
[111] Ibid., page 57 (2.5).

signal that also communicates that the signaller wants to communicate some particular information to the recipient of the signal (i.e., the tilt of the cup provides evidence about the signaller's informative intention). It is in this sense that the tilt of the cup is a signal that 'signals its own signalhood': the signal is one which provides evidence for both the informative intention and the communicative intention.

The interactions between Sally and Bill are instances of ostensive-inferential communication, as the interaction involved the production and recognition of the informative an communicative intentions. Ostensive-inferential communication is made possible because Sally and Bill are able to produce behaviours that serve as signals while simultaneously providing evidence that their behaviour is a signal. Sally tilts her cup, which informs Bill about her informative intention and communicative intention. Bill is able to make inferences about Sally's informative intention and communicative intention from her tilting her cup. Because recursive mindreading is a cognitive capacity that is shared among all (ordinarily developing) human beings, recursive mindreading can explain how speakers and listeners in general can provide the right kinds of evidence about informative and communicative intentions; and recursive mindreading can explain how speaker and listeners are able to recognize those intentions at all.[112] In order for signallers to be able to produce signals that signal their own signalhood, both the signallers and the recipients of the signal must be able to both represent and reason about each other's mental states. The cognitive mechanism that makes this possible is recursive mindreading. According to Scott-Phillips (2015), it is because human beings have the ability to engage in recursive mindreading that human beings are able to establish signals through the direct route. To see what this looks like, let us refer to an example provided by Scott-Phillips

[112] For more about how the various stages of recursive mindreading and metapsychology look during the various stages of human development, see Wilson and Sperber (2012), page 239, as well as Sperber (1994).

(2015) once again, in which Mary and Peter engage in higher and higher levels of recursive mindreading. The example demonstrates that ostensive-inferential communication can be non-linguistic, since both Mary and Peter are engaging in only bodily gestures to communicate. The example also nicely illustrates that what is required for ostensive-inferential communication to be possible is not language, but rather cognitive mechanisms that allow communicators to be able to provide the right kinds of evidence for their intentions, and to be able to make inferences from evidence provided. Because ostensive-inferential communication can be non-linguistic, the mechanism involved in being able to provide the right kinds of evidence and make the right kinds of inferences from evidence is not a linguistic one, involving encoding and decoding, but a cognitive one, which involves a metapsychology and recursive mindreading. The evidence for the speaker's intention to communicate is ostensive in the sense that it involves some sort of overt, possibly stylized, behaviour. I will explain what this behaviour might look like in the context of the example below, taken from Scott-Phillips (2015).[113]

> ⋮
> (5) Mary intends$_1$ that Peter believe$_2$ that she intends$_3$ that he believe$_4$ that the berries are edible.
> (6) Peter believes$_1$ that Mary intends$_2$ that he believe$_3$ that she intends$_4$ that he believe$_5$ that the berries are edible.

Scott-Phillips (2015) says that in scenario (5), Mary has a communicative intention, and in scenario (6), ostensive-inferential communication is taking place. In order for scenario (5) to obtain, Mary will need to be able to mentally represent that Peter believes that she wants him to believe that the berries are edible – this entails Mary forming a meta-metarepresentation, which is a representation of a representation of a representation. And in order for scenario (6) to obtain, Mary must have indicated to Peter that she intends that he form a particular belief – namely, that

[113] Scott-Phillips (2015), 3.4.

he believes that she wants him to think that she is eating the berries because they are edible. Mary can indicate her intention to Peter using overt, stylized behaviour by eating the berries enthusiastically, and perhaps pointing at the berries as she eats them. As the example illustrates, Mary and Peter are representing their own mental states, as well as each other's mental states. These mental states are beliefs about the berries being edible, Mary's beliefs and intentions, and Peter's beliefs and intentions. The kind of reasoning that Mary and Peter engage in has nothing to do with linguistic meaning, yet both of them are able to communicate with one another in this complex manner. Human beings can establish conventional codes *in virtue of having* the sophisticated meta-metarepresentational abilities and recursive mindreading abilities required for ostensive-inferential communication.[114] That is, without these sophisticated cognitive abilities, humans would not be able to establish complex conventional codes, such as the linguistic code.

There are some examples of chatbots that can represent certain types of mental phenomena. An early example is a chatbot named Parry, introduced in Colby and Weber (1971).[115] Colby and Weber (1971) programmed Parry to model human emotional states associated with the mental health condition of paranoid schizophrenia.[116] If Parry's conversational partner insults Parry, then Parry will be more likely to respond angrily, whereas if they complement Parry, then Parry will be more likely to respond positively; if a user talks about Parry's "delusion", then Parry would become fearful and hostile in its responses.[117] Colby (1974)

[114] Wilson and Sperber (2012) say that the sort of metarepresentational and reasoning abilities required for recursive mindreading in human communication seems to be complex enough so as to suggest a distinct mental module that performs the task. See page 241 – 242.

[115] Note, however, that Parry is not strictly a language model. This is because language models are programs that make predictions about which linguistic tokens are likely to follow some sequence of linguistic tokes. Parry uses a slot-based approach to constructing sentences. See Martin and Jurafsky (2020), page 488 – 500. Parry is an example of early approaches to building chatbot agent programs that could represent mental phenomena such as emotional states.

[116] Martin and Jurafsky (2020), page 495.

[117] Interestingly, Parry and Eliza (the conversational agent that mimics a Rogerian psychiatrist, see Weizenbaum (1964; 1966)) have conversed with one another on several occasions; for their meetings, Eliza was given the

discusses and tries to respond to a criticism of Parry which claims that "Parry is simply a stimulus-response model. It recognizes something in the input and then just responds to it without "thinking" or inferring. The model should interpret what it sees and engage in more computation than execution of a simple rewrite or production rule".[118] What this criticism is saying is that the process of utterance comprehension that Parry engages in is oversimplistic – that Parry merely takes a speaker's utterance as input and matches it to a corresponding output based on some pattern-matching rule. Colby responds to this criticism by stating that later iterations of Parry actually did perform inferences: Parry could represent certain aspects of the speaker's personality, such as the speaker's competence and helpfulness, by making inferences from the speaker's utterances. A more contemporary example of a chatbot that was programmed to represent mental phenomena is XiaoIce. XiaoIce is a chatbot that is currently very popular in China, having around 660 million registered users.[119] Part of XiaoIce's popularity comes from its ability to respond to users' utterances empathetically, leading some users to feel a sense of emotional connection to XiaoIce. In a manner similar to the later iterations of Parry, XiaoIce forms its representation of the user's personality based on the user's utterances. However, representing aspects of a speaker's personality is not the same thing as representing one's own mental states and the mental states of others: Representing aspects of a speaker's personality involves something more akin to empathy than it does representing representations (i.e., forming meta-metarepresentations, which are representations of representations of representations) in the sort of recursive structure described by Scott-Phillips (2015). As such, neither Parry nor XiaoIce

nickname 'The Doctor'. Transcripts of their (often quite humorous) conversations can be found here: <https://datatracker.ietf.org/doc/html/rfc439>.

[118] Colby (1974), page 1.

[119] < https://news.microsoft.com/apac/features/much-more-than-a-chatbot-chinas-xiaoice-mixes-ai-with-emotions-and-wins-over-millions-of-fans/>.

are representing the right kinds of mental phenomena to be able to reason about the mental states of others, making it impossible for them to communicate ostensive-inferentially. The ability to form meta-metarepresentations and to engage in recursive mindreading allowed human beings to create linguistic codes and use them in flexible ways. Human communicators can use linguistic codes flexibly because humans can make inferences about non-linguistic properties, including the mental states of others, to identify explicatures and implicatures. A system that relies entirely on encoding and decoding processes of the sort found in chatbot agent programs will not be able to make inferences about non-linguistic properties like other's mental states. Moreover, a chatbot that is not representing the right kinds of mental phenomena will not be able to make the kinds of inferences about non-linguistic properties required to be able to successfully gain utterance meaning.

3.4. The Sentence-Utterance Distinction and Access to Context

Earlier I said that neural language models collapse the sentence-utterance distinction. I want to say more about what I mean, as this claim encapsulates the argument I intended to make in this chapter. Consider first that corpuses are often composed of sentences that were uttered at some point in time. An utterance can be considered as a sentence that has been communicated at a particular place and time in a particular context. There are numerous examples of corpora that consist explicitly of conversation data between human communicators. These corpora consist of sentences that were uttered at some point in time by a speaker in some context. Because neural language models can only access the form of the words in a corpus during training, neural language models are (implicitly) treating the corpus as a collection of sentences. As a result, neural language models are unable to process non-linguistic properties, which are properties that can affect utterance meaning. Linguistic properties and non-linguistic properties contribute

different things to what sentences and utterances mean. Linguistic properties can tell a listener what language an utterance is being spoken in, and perhaps the roles that the words in a sentence that is uttered play in that language. Indexical expressions and pronouns are good examples of how linguistic properties can inform listeners about certain aspects of meaning, as there are general rules of thumb that a listener can follow to determine whether a word is an indexical/pronoun, and whether the entity being referred to by that word is the subject of the sentence or the object. However, non-linguistic properties are what allow a listener to determine *which entities precisely* a speaker is probably referring to with the words they utter. Language models only represent the linguistic properties of utterances, whereas listeners need to be able to make inferences about non-linguistic properties in order to grasp utterance meaning. This is because without being able to make these kinds of inferences, listeners will not be able to identify the explicatures and implicatures of an utterance.

Encoding and decoding processes allow us to use words and sentences to refer to objects and concepts: we can use the word 'cat' to refer to a particular object, for instance, in the English language. Similarly, the Korean word '고양이' (Romanised as 'goyang-i') can be used to refer to the same object as the English word 'cat'. Encoding and decoding allows us to understand what the meanings of words are in particular languages. This kind of encoding and decoding is no doubt required for successful communication, as if a listener does not know what the words a speaker is uttering refer to, then there is little hope that the listener will be able to correctly interpret the speaker's utterances. However, encoding and decoding processes cannot on their own explain how we can understand what *utterances* mean. Human communication involves providing the right kinds of evidence for utterance meaning, and making the right kinds of inferences from evidence to grasp utterance meaning. The evidence that a speaker provides for

what their utterance means includes things like tone of voice and bodily gestures. The evidence can be used by a listener to identify which inferences about utterance meaning are warranted. If a speaker utters 'Priya said she is going to the bank later today', listeners must make inferences in order to identify what the word 'bank' refers to, which individual 'she' refers to, and so on. There is nothing about the linguistic properties of the sentence that can allow a listener to know whether 'she' refers to Priya or to some other individual. If the speaker says that sentence and emphasizes the word 'Priya' while pointing at a particular individual, then the listener can infer that 'she' probably refers to 'Priya'. This is how inferences about non-linguistic properties can allow a listener to grasp utterance meaning.

I have explained why it is plausible that chatbots cannot reason about non-linguistic properties of an utterance because neural language models are natural codes. Reasoning about the non-linguistic properties of utterances to gain utterance meaning requires a metapsychology and the ability to engage in recursive mindreading (if Scott-Phillips is indeed correct). However, chatbots that rely on neural language models for their linguistic competence are still able to produce grammatically correct sentences that sometimes seem like they could have been produced by a human being. This might seem to diffuse the charge that chatbots are poor communicators because they rely on only encoding and decoding processes that only allow them access to linguistic properties when responding to speakers' utterances. But I think chatbots' relative success can be explained by the fact that neural language models are massive computer programs that can process huge amounts of language data to identify patterns in the sequences of linguistic tokens that appear in the corpus. Neural language models use word embeddings to represent certain aspects of the meanings of words by exploiting the distributional hypothesis

proposed by linguists like Zellig and Firth.[120] The distributional hypothesis maintains that the meaning of a word can be approximated from the meanings of words that commonly occur in the same contexts. The context in which a word is used is just the other words that are used spatiotemporally near it. Firth (1957) suggests that the linguistic context can provide insight about meaning.

> The basic assumption of the theory of analysis by levels is that any text can be regarded as a constituent of a context of situation or of a series of such contexts, and thus attested in experience, since the categories of the abstract context of situation will comprise both verbal and non-verbal constituents and, in renewal of connection, should be related to an observable and justifiable grouped set of events in the run of experience. [John Rupert Firth (1957), page 7]

Adopting the distributional hypothesis when constructing language models has allowed language models to make some fairly reasonable predictions about what words mean. The classic example often cited in the NLP literature is that which was presented by Mikolov et al (2013). The authors found that after having trained a neural network using a corpus, the neural network's word embedding for the word 'queen' was similar to the embedding that results from performing the following operation on the embeddings for 'king', 'man', and 'woman': 'king – man + woman'.[121] The result obtained by Mikolov et al (2013) is what a speaker of English would probably intuitively expect: the word 'queen' seems to refer to the same kind of object as ''king' – 'man' + 'woman''. The distributional hypothesis seems to explain how a computer can learn the meanings of words – and, that all the computer needs access to in order to learn the meanings of words is a corpus. Because the distribution of words in a corpus can provide insight into what words mean, a neural language model that is trained on a large corpus will be able to generate

[120] Especially Zellig (1954) and Firth (1957). For a more contemporary analysis of the distributional hypothesis, see Sahlgren (2008).
[121] Mikolov et al (2013), page 2.

some reasonable sentences in human natural language using word embeddings obtained from that corpus. Since the meanings of words is fixed by the linguistic community, it makes sense that a neural language model that is exposed to enough language data consisting of uttered sentences can learn some aspects of what words mean. Word embeddings can represent various aspects of the meanings of words, and a neural language model can use word embeddings along with an input sequence of linguistic tokens (i.e., an uttered sentence) to generate a sequence of linguistic tokens which a chatbot can utter in response to the speaker's utterance.

Even though chatbots are sometimes able to string together grammatically correct sentences, and sometimes even generate appropriate responses to speaker's utterances, the bottom line is that there is a gap between the kind of context that chatbots can access during communication (i.e., a context made up of only linguistic properties) and the kind of context that is required for communication (i.e., a context that includes both linguistic and non-linguistic properties). So, how does the fact that neural language models are natural codes explain why they cannot make inferences about non-linguistic properties of utterances? Chatbot communication involves encoding human natural language as word embeddings, and decoding word embeddings into corresponding linguistic tokens in human natural language. Inferences about non-linguistic properties allow listeners to grasp utterance meaning, and making these kinds of inferences requires access to a context that consists not of just linguistic tokens – like word embeddings – but also of entities outside of language. Inference processes are fundamentally different from decoding processes: Inference starts with premises and lead to a conclusion that is warranted by those premises, whereas decoding involves the use of a code or algorithm to convert one sequence of tokens to another. Metapsychology and recursive mindreading are cognitive abilities that are about the mental states of others, not about language,

and allow communicators to represent and make inferences about each other's mental states. The fact that human communicators can make these kinds of inferences is what makes human communication possible. As Scott-Phillips (2015) says, even the most simple utterances can have a wide range of possible, plausible meanings.[122] The way human communicators identify the meaning intended by the speaker by their utterance is by making inferences about non-linguistic properties, including (crucially) the mental states of the speaker.[123] Since neural language models can only represent linguistic properties to encode human natural language as word embeddings and decode word embeddings into linguistic tokens in human natural language, chatbots will not be able to engage in inference processes about other's mental states insofar as (i) they lack the relevant cognitive mechanisms that are needed to make those inferences, and (ii) they cannot represent the relevant kinds of mental phenomena to make inferences that allow them to gain utterance meaning.

I suspect that chatbots will not be able to communicate effectively in human natural language until they are able to engage in the kinds of cognitive processes required to access the non-linguistic aspects of context. While the use of large-scale statistical methods to train neural language models that involve massive corpora and word embeddings that represent certain aspects of linguistic meaning have led to successes, future research in this field will benefit from trying to implement the kinds of cognitive capacities that make human communication possible in chatbot agent programs.

[122] Scott-Phillips (2015), page 39 (1.5).
[123] Ibid., page 35 – 36 (1.5).

Bibliography

Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B. & Rousseau, D.. (2015). The Higgs boson machine learning challenge. *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning, in PMLR* 42:19-55

Austin, John Langshaw (1962). *How to Do Things with Words*. Clarendon Press.

Bach, Kent (1994). Conversational impliciture. In Maite Ezcurdia & Robert J. Stainton (eds.), *The Semantics-Pragmatics Boundary in Philosophy*. Broadview Press. pp. 284.

Bach, Kent. (2010). Impliciture vs Explicature: What's the Difference?.

Bender, Emily M. and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." *ACL* (2020).

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., & Kandola, J., Hofmann, T., Poggio, T., & Shawe-Taylor, J. (Eds.). (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*(6), 1137–1155.

Blackburn, Perry Louis. "*The code model of communication : a powerful metaphor in linguistic metatheory*." (2007).

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. "Language Models are Few-Shot Learners." *ArXiv* abs/2005.14165 (2020): n. pag.

Buckner, Cameron and James Garson, "Connectionism", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

Cetina, Víctor Uc, N. Navarro, A. Martín-González, C. Weber and S. Wermter. "Survey on reinforcement learning for language processing." *ArXiv* abs/2104.05565 (2021): n. pag.

Chomsky, Noam (1957). *Syntactic Structures*. Mouton.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT Press.

Colby, K.. "Ten criticisms of parry." *SIGART Newsl.* 48 (1974): 5-9.

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. Artificial Intelligence 2(1), 1–25.

Dong, Li & Yang, Nan & Wang, Wenhui & Wei, Furu & Liu, Xiaodong & Wang, Yu & Gao, Jianfeng & Zhou, Ming & Hon, Hsiao-Wuen. (2019). Unified Language Model Pre-training for Natural Language Understanding and Generation.
Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-55. In Studies in Linguistic Analysis (pp. 1-31). Special Volume of the Philological Society. Oxford: Blackwell.

Gauker, Christopher (1992). The Lockean theory of communication. *Noûs 26 (3):303-324.*

Grice, H. Paul (1957). Meaning. *Philosophical Review* 66 (3):377-388.

Grice, Herbert Paul (1967). Logic and conversation. In Paul Grice (ed.), *Studies in the Way of Words*. Harvard University Press. pp. 41-58.

Harris, Z. S. (1954). Distributional structure. *Word, 10,* 146–162.

Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

Józefowicz, R., Oriol Vinyals, M. Schuster, Noam M. Shazeer and Yonghui Wu. “Exploring the Limits of Language Modeling.” *ArXiv* abs/1602.02410 (2016): n. pag.

Jurafsky, D. & Martin, J. (2020). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. 3rd Edition draft.

Levesque, H.. “On our best behaviour.” *Artif. Intell.* 212 (2014): 27-35.

Lewis, David K. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell.

Locke, John (1689). *An Essay Concerning Human Understanding*. Oxford University Press.

Lombardi, Olimpia, Federico Holik, and Leonardo Vanni. "What Is Shannon Information?" *Synthese* 193, no. 7 (2016): 1983-2012.

Markov, A. A. (2006). An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Science in Context* 19 (4):591-600.

Mikolov, Tomas, Kai Chen, G. Corrado and J. Dean. “Efficient Estimation of Word Representations in Vector Space.” *ICLR* (2013).

Mikolov, Tomas, Wen-tau Yih and G. Zweig. “Linguistic Regularities in Continuous Space Word Representations.” *HLT-NAACL* (2013).

Matsumoto, D., & Willingham, B. (2009). Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology, 96(1), 1–10.*

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos and A. McCallum. “Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space.” *ArXiv* abs/1504.06654 (2014): n. pag.

Pennington, Jeffrey, R. Socher and Christopher D. Manning. “Glove: Global Vectors for Word Representation.” *EMNLP* (2014).

Rashkin, Hannah, Eric Michael Smith, Margaret Li and Y-Lan Boureau. “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset.” *ACL* (2019).

Robyn Carston, (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, J. Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau and J. Weston. “Recipes for Building an Open-Domain Chatbot.” *EACL* (2021).

Russell, Stuart and Norvig, Peter. *Artificial Intelligence: A Modern Approach*. 3 : Prentice Hall, 2010.

Sahlgren, Magnus. “The Distributional Hypothesis.” *The Italian Journal of Linguistics* 20 (2008): 33-54.

Schiffer, Stephen (1972). *Meaning*. Oxford, Clarendon Press.

Scott-Phillips, T. and R. Blythe. “Why is combinatorial communication rare in the natural world, and why is language an exception to this trend?” *Journal of The Royal Society Interface* 10 (2013): n. pag.

Scott-Phillips, T. (2015). Speaking our minds: Why human communication is different, and how language evolved to make it special. New York: Palgrave Macmillan.

Searle, J. (1969). Speech Acts. *Foundations of Language* 11 (3):433-446.

Shannon, Claude E. & Weaver, Warren (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Shaoul, C. & Westbury C. (2010) The Westbury Lab Wikipedia Corpus, Edmonton, AB: University of Alberta (downloaded from http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html).

Sperber, Dan. 1994. Understanding verbal understanding. In J. Khalfa (ed.) *What is Intelligence?* 179-98. Cambridge University Press, Cambridge.

Sperber, Dan. (2000). Metarepresentations in an evolutionary perspective.

Sperber, Dan & Wilson, Deirdre (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Sperber, Dan & Wilson, Deirdre (1987). Précis of Relevance: Communication and Cognition. *Behavioral and Brain Sciences* 10 (4):697.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell Publishing.

Tang, An, Roger Tam, Alexandre Cadrin-Chênevert, Will Guest, Jaron Chong, Joseph Barfett, Leonid Chepelev, et al. "Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology." Canadian Association of Radiologists Journal 69, no. 2 (May 2018): 120–35. https://doi.org/10.1016/j.carj.2018.02.002.

Valente, D., Theurel, A. & Gentaz, E. The role of visual experience in the production of emotional facial expressions by blind people: a review. *Psychon Bull Rev* **25,** 483–497 (2018).

Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. "Attention is All you Need." *ArXiv* abs/1706.03762 (2017): n. pag.

Weaver, Warren. "Recent Contributions to the Mathematical Theory of Communication." *ETC: A Review of General Semantics* 10, no. 4 (1953): 261-81.

Wilson, Deirdre & Sperber, Dan (1993). Linguistic Form and Relevance. *Lingua* 90:1-25.

Wilson, Deirdre & Sperber, Dan (2012). *Meaning and Relevance*. Cambridge University Press.