On the Computation of Meaning, Language Models and Incomprehensible Horrors

Michael Timothy Bennett[1][0000−0001−6895−8782]

The Australian National University michael.bennett@anu.edu.au

Abstract. We bring together foundational theories of meaning and a mathematical formalism of artificial general intelligence to provide a mechanistic explanation of meaning, communication and symbol emergence. We establish circumstances under which a machine might mean what we think it means by what it says, or comprehend what we mean by what we say. We conclude that a language model such as ChatGPT does not comprehend or engage in meaningful communication with humans, though it may exhibit complex behaviours such as theory of mind[1].

1 Introduction

Linguists and philosophers have offered various accounts of the behaviour of language, meaning and the human mind. Computer scientists have posited mechanisms to replicate these variously described behaviours piecemeal. The former is a top-down approach, while the latter is bottom up. Unfortunately, it is difficult to connect the two. Large language models (LLMs) such as ChatGPT are a bottom up attempt to capture the behaviour of written language, and are remarkably good at giving human-like responses to questions [2]. Yet it is unclear the extent to which an LLM actually means what it says or understands what we mean. A true artificial general intelligence (AGI) should not just parrot what we expect, but respond to what we mean and mean what it says. Yet how we would we know if that were the case? Computers represent syntax, and from correlations in syntax an LLM is supposed to glean meaning. However, meaning is not well defined in computational terms. We need to connect top-down descriptions of meaning to bottom-up computation. What is meaning, and how might we compute it?

1.1 Grice's foundational theory of meaning

Grice's foundational theory of meaning [3] holds that meaning is what the speaker *intends* to convey to the listener. Grice gave an illustrative example,

> [the speaker] α means m by uttering u iff α intends in uttering u that
> 1. his audience come to believe m,
> 2. his audience recognize this intention [called m-intention], and
> 3. (1) occur on the basis of (2). [4]

[1] Technical appendices are available on GitHub [1].

That Grice's theory is *foundational* means it specifies the facts in virtue of which expressions have particular semantic properties (as opposed to describing those semantic properties). A foundational theory is named as such because it sits below semantic theories. It is illustrative of what we're attempting (connect bottom up computation to a top down description).

1.2 A foundational theory of foundational theories

Were we to accept that meaning is in virtue of m-intent[2] , then from what does that arise? M-intent should not be conflated with intent in general because it pertains to what one means by an expression, whereas intent more generally is any goal in service of which decisions are made. The former stems from the latter [7], and so there exists a theory arguing that meaning exists in virtue of one's intent in the sense of goals. Grice's theories are better established and widely accepted with respect to meaning, but these theories are not mutually exclusive and the depiction of meaning as in virtue of intent in general is a bridge we can use to connect Grice's top down description to bottom-up computational processes. This is because it explains intent in virtue of inductive inference, to argue that meaningful communication with an AI, or any organism, requires similar feelings and experiences, in order to construct similar goals and "solutions to tasks" [7] (an argument formed in relation to the Fermi Paradox [8]). This explanation was too vague to be of significance for engineering. For example it assumed a measure, "weakness", which was not well defined. However, weakness *is* well defined in a more recent formalism of artificial general intelligence (AGI) and enactive cognition [1], so we will instead reformulate the theory using that formalism, extending it to account for meaningful communication. We begin with cognition formalised using tasks. We then formalise an organism using tasks to provide a novel account of preferences, symbol systems and meaningful communication. We then describe circumstances under which an organism might mean what we think it means by what it says, or infer what we mean by what we say.

2 Meaning, from the top down

Intent only exists in virtue of a task one is undertaking [7]. A task is what we get if we add context to intent, expressing what is relevant about both the agent and the environment. A task can be used to formalise enactive cognition [9], discarding notions of agent and environment in favour of a set of decision

[2] We note there are other well-known and widely accepted descriptions of meaning (Russell, Frege, Searle, Davidson, Wittgenstein, Lewis, Kripke etc), some of which we allude to and even touch upon as part of our formalism. We also acknowledge Grice later expanded upon the notion of m-intent [5, 6]. However, due to paper length restrictions, we are limited in what we discuss. As the ability to infer intent seems most novel in the context of AGI, we focus our discussion on the early Gricean characterisation of meaning.

problems [7, 1]. A task is something which is completed, like a goal, so intent is formalised like a goal [10]. A goal is a set of criteria, and if those criteria are satisfied, then it is satisfied and the task complete. To formalise meaning we must avoid grounding problems [11]. As such these criteria are grounded by representing the environment, of which cognition is part, as a set of declarative programs [12] of which the universe is the interpreter [13]:

Definition 1 (environment).

- *We assume a set Φ whose elements we call* ***states****, one of which we single out as the* ***present state****.*
- *A* ***declarative program*** *is a function $f : \Phi \to \{true, false\}$, and we write P for the set of all declarative programs. By an* ***objective truth*** *about a state ϕ, we mean a declarative program f such that $f(\phi) = true$.*

Definition 2 (implementable language).

- $\mathfrak{V} = \{V \subset P : V \ is \ finite\}$ *is a set whose elements we call* ***vocabularies****, one of which*[3] *we single out as* ***the vocabulary*** $\mathfrak{v}$*.*
- $L_\mathfrak{v} = \{l \subseteq \mathfrak{v} : \exists \phi \in \Phi \ (\forall p \in l : p(\phi) = true)\}$ *is a set whose elements we call* ***statements****. $L_\mathfrak{v}$ follows Φ and $\mathfrak{v}$, and is called* ***implementable language****.*
- $l \in L_\mathfrak{v}$ *is* ***true*** *iff the present state is ϕ and $\forall p \in l : p(\phi) = true$.*
- *The* ***extension of a statement*** $a \in L_\mathfrak{v}$ *is $Z_a = \{b \in L_\mathfrak{v} : a \subseteq b\}$.*
- *The* ***extension of a set of statements*** $A \subseteq L_\mathfrak{v}$ *is $Z_A = \bigcup_{a \in A} Z_a$.*

(Notation) *Z with a subscript is the extension of the subscript*[4]*.*

A goal can now be expressed as a statement in an implementable language. An implementable language represents sensorimotor circuitry[5] with which cognition is enacted. It is not natural language, but a dyadic system with exact meaning. Peircean semiosis [14] is integrated to explain natural language. Peirce defined a symbol as a sign (E.G. the word "pain"), a referent (E.G. the experience of pain), and an interpretant which links the two, "determining the effect upon" the organism. A goal arguably functions as an interpretant because it determines the effect of a situation upon an organism that pursues it [7]. Rather than formulate a task and then rehash the argument that a task is a symbol, we'll just formalise a symbol using the existing definition of a task [1, definition 3]:

Definition 3 ($\mathfrak{v}$-task). *For a chosen $\mathfrak{v}$, a task α is a triple $\langle S_\alpha, D_\alpha, M_\alpha \rangle$, and $\Gamma_\mathfrak{v}$ is the set of all tasks given $\mathfrak{v}$. Give a task α:*

- *$S_\alpha \subset L_\mathfrak{v}$ is a set whose elements we call* ***situations*** *of α.*
- *S_α has the extension Z_{S_α}, whose elements we call* ***decisions*** *of α.*
- *$D_\alpha = \{z \in Z_{S_\alpha} : z \ is \ correct\}$ is the set of all decisions which complete α.*

[3] The vocabulary $\mathfrak{v}$ we single out represents the sensorimotor circuitry with which an organism enacts cognition - their brain, body, local environment and so forth.

[4] e.g. Z_s is the extension of s.

[5] Mind, body, local environment etc.

– $M_\alpha = \{l \in L_{\mathfrak{v}} : Z_{S_\alpha} \cap Z_l = D_\alpha\}$ *whose elements we call* ***models*** *of* α.

(Notation) *If* $\omega \in \Gamma_{\mathfrak{v}}$*, then we will use subscript* ω *to signify parts of* ω*, meaning one should assume* $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$ *even if that isn't written.*

(How a task is completed) *Assume we've a* $\mathfrak{v}$*-task* ω *and a hypothesis* $\boldsymbol{h} \in L_{\mathfrak{v}}$ *s.t.*

1. *we are presented with a situation* $s \in S_\omega$*, and*
2. *we must select a decision* $z \in Z_s \cap Z_{\boldsymbol{h}}$.
3. *If* $z \in D_\omega$*, then* z *is correct and the task is complete. This occurs if* $\boldsymbol{h} \in M_\omega$.

Definition 4 (symbol). *A task* α *is also a Peircean symbol:*

– $s \in S_\alpha$ *is a* ***sign*** *of* α.
– $d \in D_\alpha$ *is the effect of* α *upon one who perceives it.* d *may be sensorimotor activity associated with perception, and thus a* ***referent****.*
– $m \in M_\alpha$ *is an* ***interpretant*** *linking* ***signs*** *to* ***referents****.*

Tasks may be divided into narrower child tasks, or merged into parent tasks.

Definition 5 (child, parent and weakness). *A symbol* α *is a child of* ω *if* $S_\alpha \subset S_\omega$ *and* $D_\alpha \subseteq D_\omega$*. This is written* $\alpha \sqsubset \omega$*. We call* $|D_\alpha|$ *the weakness of a symbol* α*, and a parent is weaker than its children.*

2.1 Extending the formalism

The child and parent relation means a symbol is also a symbol system in that it can be subdivided into child symbols [7]. With this in mind, we can define an organism that derives symbols from its experiences, preferences and feelings.

Definition 6 (organism). *An organism* $\mathfrak{o}$ *is a quintuple* $\langle \mathfrak{v}_\mathfrak{o}, \mathfrak{e}_\mathfrak{o}, \mathfrak{s}_\mathfrak{o}, n_\mathfrak{o}, f_\mathfrak{o} \rangle$*, and the set of all such quintuples is* $\mathfrak{O}$ *where:*

– $\mathfrak{v}_\mathfrak{o}$ *is a* ***vocabulary*** *we single out as belonging to this organism*[6]*.*
– *We assume a* $\mathfrak{v}_\mathfrak{o}$*-task* β *wherein* S_β *is every situation in which* $\mathfrak{o}$ *has made a decision, and* D_β *contains every such decision. Given the set* $\Gamma_{\mathfrak{v}_\mathfrak{o}}$ *of all tasks,* $\mathfrak{e}_\mathfrak{o} = \{\omega \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : \omega \sqsubset \beta\}$ *is a set whose members we call* ***experiences****.*
– *A* ***symbol system*** $\mathfrak{s}_\mathfrak{o} = \{\alpha \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : \text{there exists } \omega \in \mathfrak{e}_\mathfrak{o} \text{ where } M_\alpha \cap M_\omega \neq \emptyset\}$ *is a set whose members we call* ***symbols****.* $\mathfrak{s}_\mathfrak{o}$ *is the set of every task to which it is possible to generalise (see [1, definition 5]) from an element of* $\mathfrak{e}_\mathfrak{o}$.
– $n_\mathfrak{o} : \mathfrak{s}_\mathfrak{o} \to \mathbb{N}$ *is a function we call* ***preferences****.*
– $f_\mathfrak{o} : \mathfrak{s}_\mathfrak{o} \to \mathfrak{f}_\mathfrak{o}$ *is a function, and* $\mathfrak{f}_\mathfrak{o} \subset L_{\mathfrak{v}_\mathfrak{o}}$ *a set whose elements we call* ***feelings****, being the reward, qualia etc, from which preferences arise*[7]*.*

[6] The corresponding $L_{\mathfrak{v}_\mathfrak{o}}$ is all sensorimotor activity in which the organism may engage.
[7] Note that this assumes qualia, preferences and so forth are part of physical reality, which means they are sets of declarative programs.

Each symbol in $\mathfrak{s}_\mathfrak{o}$ shares an interpretant at least one experience[8]. This is so feelings $f_\mathfrak{o}$ ascribed to symbols can be grounded in experience. Humans are given impetus by a complex balance of feelings (reward signals, qualia etc). It is arguable that feelings eventually determine all value judgements [10]. As Hume pointed out, one cannot derive a statement of what ought to be from a statement of what is. Feelings are an ought from which one may derive all other oughts. If meaning is about intent, then the impetus that gives rise to that intent is an intrinsic part of all meaning [8]. Intent is a goal. A goal is statement of what ought to be that one tries to make into a description of what is, by altering the world to fit with ought to be. We assume feelings are consequence of natural selection, and so explain meaning in virtue of a mechanistic process. Each $l \in L$ represents sensorimotor activity, which from a materialist perspective includes feelings. Thus, $f_\mathfrak{o}$ is a function from symbols to sensorimotor activity. Statements and symbols "mean something" to the organism if the organism can ascribe feelings to them. As every symbol in $\mathfrak{s}_\mathfrak{o}$ contains an interpretant which is part of the organism's experience, the organism can ascribe feelings to all symbols on the basis of that experience. If one is not concerned with qualia [16, 17], then feelings may be simulated with "reward" functions. However, to simulate feelings that result in human-like behaviour is a more difficult proposition. Rather than trying to describe human-like feelings, we simplify our analysis by assuming the preferences [18] $n_\mathfrak{o}$ which are determined by experience of feelings.

2.2 Interpretation

The **situation at hand** $s \in L_{\mathfrak{v}_\mathfrak{o}}$ is a statement $\mathfrak{o}$ experiences as a sign and then **interprets** using $\alpha \in \mathfrak{s}_\mathfrak{o}$ s.t. $s \in S_\alpha$, to decide $d \in Z_s \cap Z_{M_\alpha}$.

Definition 7 (interpretation). *Interpretation is a sequence of steps:*

1. *The situation at hand $s \in L_{\mathfrak{v}_\mathfrak{o}}$ **signifies** a symbol $\alpha \in \mathfrak{s}_\mathfrak{o}$ if $s \in S_\alpha$.*
2. *$\mathfrak{s}^s_\mathfrak{o} = \{\alpha \in \mathfrak{s}_\mathfrak{o} : s \in S_\alpha\}$ is the set of all symbols which s signifies.*
3. *If $\mathfrak{s}^s_\mathfrak{o} \neq \emptyset$ then s **means something** to the organism in the sense that there are feelings which can be ascribed to symbols in $\mathfrak{s}^s_\mathfrak{o}$.*
4. *If s means something, then $\mathfrak{o}$ uses $\alpha \in \arg\max_{\omega \in \mathfrak{s}^s_\mathfrak{o}} n_\mathfrak{o}(\omega)$ to interpret s.*
5. *The interpretation is a decision $d \in Z_s \cap Z_{M_\alpha}$[9].*

3 Communication of meaning

We develop our explanation in four parts. First, we define exactly what it means for an organism to affect and be affected by others. Second, we examine how one

[8] A symbol system is every task to which one may generalise from one's experiences. Only finitely many symbols may be entertained. In claiming our formalism pertains to meaning in natural language we are rejecting arguments, such as those of Block and Fodor [15], that a human can entertain an infinity of propositions (because time and memory are assumed to be finite, which is why $\mathfrak{v}_\mathfrak{o}$ is finite).

[9] How an organism responds to a sign that means nothing is beyond this paper's scope.

organism may anticipate the behaviour (by inferring the end it serves) of another or order to change how they are affected. Third, we examine how said organism may, having anticipated the behaviour of the other, intervene to manipulate the other's behaviour to their benefit (so that the now latter affects the former in a more positive way). And finally, we examine what happens when each organism is attempting to manipulate the another. Each anticipates the other's manipulation, because each anticipates the other's behaviour by inferring its intent. An organism can then attempt to deceive the other organism (continue the manipulative approach), or attempt to co-operate (communicate in good faith), a choice resembling an iterated prisoner's dilemma.

We assume organisms make decisions based upon preferences, but preferences are not arbitrary. Feelings and thus preferences exist in virtue of natural selection, which to some extent must favour rational behaviour (to the extent that selection is significantly impacted). In computer science terms this might be understood as alignment. One's feelings are the result of alignment by genetic algorithm, and one's preferences are the result of reinforcement learning using those feelings (to determine reward). Thus we assume preferences are a balance of what is rational, and what is tolerably irrational, given the pressures of natural selection. We call this balance **reasonably performant**. The specifics of inductive inference are beyond the scope of this paper, however definitions and formal proofs pertaining to inductive inference from child to parent tasks are included in the appendix [1]. We assume the necessary inductive capabilities when we assume organisms are reasonably performant.

3.1 Ascribing intent

Definition 8 (affect). *To affect an organism* $\mathfrak{o}$ *is to cause it to make a different decision than it otherwise would have.* $\mathfrak{k}$ *affects* $\mathfrak{o}$ *if* $\mathfrak{o}$ *would have made a decision* d*, but as a result of a decision* c *made by* $\mathfrak{k}$*,* $\mathfrak{o}$ *makes decision* $g \neq d$*.*

Let $\mathfrak{k}$ and $\mathfrak{o}$ be organisms. If $\mathfrak{k}$ affects $\mathfrak{o}$, and assuming $\mathfrak{v}_\mathfrak{o}$ is sufficient to allow $\mathfrak{o}$ to distinguish when it is affected by $\mathfrak{k}$ from when it is not (meaning all else being equal $\mathfrak{k}$'s interventions are distinguishable by the presence of an identity [1, definition 13]), then there exists experience $\zeta^\mathfrak{k}_\mathfrak{o} \in \mathfrak{e}_\mathfrak{o}$ such that $d \in D_{\zeta^\mathfrak{k}_\mathfrak{o}}$ if $\mathfrak{o}$ is affected by $\mathfrak{k}$. $\zeta^\mathfrak{k}_\mathfrak{o}$ is an ostensive definition [19] of $\mathfrak{k}$'s intent (meaning it is a child task from which we may infer the parent representing $\mathfrak{k}$'s most likely intent and thus future behaviour) [7]. In the absence of more information, the symbol most likely to represent $\mathfrak{k}'s$ intent is the weakest [7], meaning $\alpha \in \mathfrak{s}_\mathfrak{o}$ s.t. $|Z_\alpha|$ is maximised. However, because $\mathfrak{o}$ assumes $\mathfrak{k}$ has similar feelings and preferences [7, 10][10] $n_\mathfrak{o}$ is an approximation of what $\mathfrak{k}$ will do. Accordingly the symbol most likely to represent $\mathfrak{k}$'s intent would be

$$\gamma^\mathfrak{k}_\mathfrak{o} \in \arg\max_{\alpha \in \mathfrak{K}} |Z_\alpha| \text{ s.t. } \mathfrak{K} = \arg\max_{\alpha \in \Gamma^\mathfrak{k}_\mathfrak{o}} n_\mathfrak{o}(\alpha) \text{ and } \Gamma^\mathfrak{k}_\mathfrak{o} = \{\omega \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : M_{\zeta^\mathfrak{k}_\mathfrak{o}} \cap M_\omega \neq \emptyset\}$$

[10] Members of a species tend to have similar feelings, experiences and thus preferences.

The above is the "weakest" of goals preferred by $\mathfrak{o}$ which, if pursued by $\mathfrak{k}$, would explain why $\mathfrak{k}$ has affected $\mathfrak{o}$ as it has.

3.2 From manipulation to meaningful communication

We've explained inference of intent in counterfactual terms, answering "if places were exchanged, what would cause $\mathfrak{o}$ to act like $\mathfrak{k}$?". Intent here is "what is $\mathfrak{k}$ trying to achieve by affecting $\mathfrak{o}$", rather than just "what is $\mathfrak{k}$ trying to achieve".

Means of manipulation: In virtue of being reasonably performant organism $\mathfrak{o}$ infers the intent of an organism $\mathfrak{k}$ that affects $\mathfrak{o}$. $\mathfrak{o}$ must do this in order to plan ahead and ensure its own needs will be met. However $\mathfrak{o}$ can go further than merely reacting to what it anticipates $\mathfrak{k}$ will do. It can also attempt to influence what $\mathfrak{k}$ will do. If being reasonably performant necessitates $\mathfrak{o}$ represent $\mathfrak{k}$'s intent because $\mathfrak{k}$ affects $\mathfrak{o}$, then it may also necessitate $\mathfrak{o}$ affect $\mathfrak{k}$ to the extent that doing so will change how $\mathfrak{k}$ affects $\mathfrak{o}$. This describes what might be commonly understood as an attempt at manipulation.

Communication: If both $\mathfrak{o}$ and $\mathfrak{k}$ are reasonably performant and affect one another, each will attempt to manipulate the other. Furthermore, being reasonably performant and ascribing intent to one another's behaviour, each must account for the manipulative intent of the other when attempting to manipulate said other. Subsequently each organism must account for how its own manipulative intent will be perceived by the other. As in a certain class of iterated prisoner's dilemma, the rational choice may then be to co-operate.

Furthermore if there is sufficient profit in affecting another's behaviour, then knowing one's own intent is perceived by that other and that the other will change its behaviour according to one's own intent, it makes sense to actually change one's own intent in order to affect the other. This bears out experimentally in reinforcement learning with extended environments [20]. The rational course of action is to actually *have* co-operative intent, assuming $\mathfrak{k}$ can perceive $\mathfrak{o}$'s intent correctly, and that $\mathfrak{k}$ will reciprocate in kind[11]. Inductive inference (see appendix [1]) undertaken with co-operative intent would, if undertaken by reasonably performant organisms, ensure the organisms in a population have preferences that favour symbols that mean (in a behaviourally approximate manner) similar things to all members of the population. If organisms are attempting to co-operate, and can infer one another's intent, then repeated interactions would give rise to signalling conventions we might call natural language.

Meaning: Let us reframe these ideas using the example from the introduction. We'll say two symbols $\alpha \in \mathfrak{s}_{\mathfrak{k}}$ and $\omega \in \mathfrak{s}_{\mathfrak{o}}$ are roughly equivalent (written $\alpha \approx \omega$) to mean feelings, experiences and thus preferences associated with a symbol are in some sense the same for two organisms (meaning if $\alpha \approx \omega$ then $f_{\mathfrak{k}}(\alpha) \approx f_{\mathfrak{o}}(\omega)$ etc).

[11] Which again hinges upon preferences.

$\mathfrak{k}$ means $\alpha \in \mathfrak{s}_{\mathfrak{k}}$ by deciding u and affecting $\mathfrak{o}$ iff $\mathfrak{k}$ intends in deciding u:
1. that $\mathfrak{o}$ interpets the situation at hand with $\omega \in \mathfrak{s}_{\mathfrak{o}}$ s.t. $\omega \approx \alpha$,
2. $\mathfrak{o}$ recognize this intention, for example by predicting it according to

$$\gamma^{\mathfrak{k}}_{\mathfrak{o}} \in \arg\max_{\alpha \in \mathfrak{K}} |Z_\alpha| \text{ s.t. } \mathfrak{K} = \arg\max_{\alpha \in \Gamma^{\mathfrak{k}}_{\mathfrak{o}}} n_{\mathfrak{o}}(\alpha), \Gamma^{\mathfrak{k}}_{\mathfrak{o}} = \{\omega \in \Gamma_{\mathfrak{v}_{\mathfrak{o}}} : M_{\zeta^{\mathfrak{k}}_{\mathfrak{o}}} \cap M_\omega \neq \emptyset\}$$

3. and (1) occur on the basis of (2), because $\mathfrak{k}$'s intent is to co-operate and so it will interpret the situation at hand using what it has inferred of $\mathfrak{o}$'s intent.

Note that the above describes co-operative communication. Prerequisites for the comprehension of meaning follow from the above:

1. Organisms must be able to **affect one another**.
2. Organisms must have similar **feelings**, and
3. similar **experiences**, so $\mathfrak{s}_{\mathfrak{o}}$ and $\mathfrak{s}_{\mathfrak{k}}$ contain roughly equivalent symbols.
4. Similar **preferences** then inform the correct inference of intent.
5. Finally, all this assumes organisms are **reasonably performant**.

4 Talking to a machine

LLMs and humans are able to affect one another, and have similar preferences even to the extent that LLMs appear to exhibit theory of mind [2]. However, while mimicking human preferences after the fact gives the appearance of holding values and beliefs, an LLM has no impetus because it is not compelled by feelings, and so cannot entertain roughly equivalent symbols. This is not to say we cannot reverse engineer the complex balance of human-like feelings, merely that we have not. If an LLM has any impetus at all, it is to be found in our prompts. It is reminiscent of a mirror test, which is a means of determining whether animals are self aware. For example, a cat seeing itself in the mirror may attempt to attack what it sees, not realising what it sees is not another animal but its own reflection. In an LLM we face a mirror test of our own, but instead of light it reflects our own written language back at us. We then ascribe motives and feelings to that language, because we have evolved to infer the intent of organisms compelled by feelings [7]. An LLM hijacks the shortcuts we use to understand one another (it takes advantage of the fact that we assume others are motivated by similar feelings [10]). We've a history of ascribing feelings and agency to things possessed of neither. In the 1970s, a chatbot named ELIZA made headlines as its users attributed feelings and motives to its words [21]. Like ELIZA, today's LLMs not only do not mean what we think they mean by what they say, but do not mean anything at all. This is not an indictment of LLMs trained to mimic human preferences. The meaning we ascribe to their behaviour can be useful, even if that behaviour was not intended to mean anything.

The Hall of Mirrors: Even if we were to approximate human feelings, an LLM like ChatGPT is not reasonably performant. It is maladaptive, requiring an abundance of training data. This may be because training does not optimise for a weak representation, but settles for any function fitting the data[12] [7]. Returning to mirror analogies, imagine a hall of mirrors reflecting an object from different angles. A weak or simple representation would be one symbol $\alpha \in \mathfrak{s}_\mathfrak{o}$ representing the object, which is then interpreted from the perspectives $a, b, c, d \in S_\alpha$ of each mirror. A needlessly convoluted representation of the same would instead interpret a, b, c and d using different symbols. These would be α's children $\omega, \gamma, \delta, \sigma \sqsubset \alpha$ such that $a \in S_\omega, b \in S_\gamma, c \in S_\delta, d \in S_\sigma$. This latter representation fails to exploit what is common between perspectives, which might allow it to generalise [7] to new perspectives. That an LLM may not learn sufficiently weak representations seems consistent with their flaws. One well documented example of this is how an LLM may convincingly mimic yet fail to understand arithmetic [22], but such flaws may more subtly manifest elsewhere. For example, when we queried Bing Chat (on the 2^{nd} of February 2023 [1, p.11]) with the name and location of a relatively unknown individual who had several professions and hobbies mentioned on different sites, Bing concluded that different people with this name lived in the area, each one having a different hobby or profession.

Incomprehensibility: If we aspire to build machines that mean what we think they mean by what they say, then it would be necessary to give the machine impetus by simulating human feelings. It is interesting to consider where this may lead. If we do not get the balance of feelings quite right, we might create an organism that means what it says, but whose meanings are utterly incomprehensible to us because the resulting preferences are unaligned with ours. This is not to say such an organism would be dangerous. Alignment may be more a matter of meaningful communication than safety. In the introduction we mentioned ideas on which this paper was founded were used to relate the Fermi paradox to control of and communication with an AGI [8]. We can extend that notion. Assume we are affected by an organism. If the events befalling us are set in motion by preferences entirely unlike our own, then we would fail to ascribe the correct intent to the organism. We may fail entirely to realise there is an organism, or may ascribe many different intents as in the hall of mirrors analogy. Furthermore, $\mathfrak{v}_\mathfrak{o}$ determines what can or cannot be comprehended by an organism [1]. It may be that $\mathfrak{v}_\mathfrak{o}$ contains nothing akin to the contents of $\mathfrak{v}_\mathfrak{k}$, making $\mathfrak{o}$ is incapable of representing and thus comprehending $\mathfrak{k}$'s intent.

References

[1] M. T. Bennett. *Technical Appendices*. Version 1.2.1. 2023. DOI: 10.5281/zenodo.7641742. URL: https://github.com/ViscousLemming/Technical-Appendices (visited on 03/04/2023).

[12] Albeit with some preference for simplicity imparted by regularisation.

[2] M. Kosinski. *Theory of Mind May Have Spontaneously Emerged in Large Language Models*. 2023. URL: https://arxiv.org/abs/2302.02083v1.
[3] P. Grice. "Meaning". In: *The Philosophical Review* 66.3 (1957), p. 377.
[4] J. Speaks. "Theories of Meaning". In: *The Stanford Encyclopedia of Philosophy*. Spring 2021. Stanford University, 2021.
[5] P. Grice. "Utterer's Meaning and Intention". In: *The Philosophical Review* 78.2 (1969), pp. 147–177.
[6] H. P. Grice. *Studies in the Way of Words*. Harvard University Press, 2007.
[7] M. T. Bennett. "Symbol Emergence and the Solutions to Any Task". In: *Artificial General Intelligence*. Cham: Springer, 2022, pp. 30–40.
[8] M. T. Bennett. "Compression, The Fermi Paradox and Artificial Super-Intelligence". In: *Artificial General Intelligence*. Springer, 2022, pp. 41–44.
[9] D. Ward, D. Silverman, and M. Villalobos. "Introduction: The Varieties of Enactivism". In: *Topoi* 36 (Apr. 2017).
[10] M. T. Bennett and Y. Maruyama. "Philosophical Specification of Empathetic Ethical Artificial Intelligence". In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2022), pp. 292–300.
[11] S. Harnad. "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1 (1990), pp. 335–346.
[12] W. A. Howard. "The Formulae-as-Types Notion of Construction". In: *To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*. Ed. by J. Seldin and J. Hindley. Cambrdige MA: Academic Press, 1980, pp. 479–490.
[13] G. Piccinini and C. Maley. "Computation in Physical Systems". In: *The Stanford Encyclopedia of Philosophy*. Summer. Stanford University, 2021.
[14] A. Atkin. "Peirce's Theory of Signs". In: *The Stanford Encyclopedia of Philosophy*. Spring. Metaphysics Research Lab, Stanford University, 2023.
[15] N. Block and J. Fodor. "What Psychological States Are Not". In: *The Philosophical Review* (81 1972), pp. 159–181.
[16] D. Chalmers. "Facing Up to the Problem of Consciousness". In: *Journal of Consciousness Studies* 2.3 (1995), pp. 200–19.
[17] P. Boltuc. "The Engineering Thesis in Machine Consciousness". In: *Techné: Research in Philosophy and Technology* 16.2 (2012), pp. 187–207.
[18] S. A. Alexander. "The Archimedean trap: Why traditional reinforcement learning will probably not yield AGI". In: *Journal of Artificial General Intelligence* 11.1 (Jan. 2020), pp. 70–85.
[19] A. Gupta. "Definitions". In: *The Stanford Encyclopedia of Philosophy*. Winter 2021. Stanford University, 2021.
[20] S. A. Alexander et al. "Extending Environments to Measure Self-reflection in Reinforcement Learning". In: *Journal of Artificial General Intelligence* 13.1 (2022), pp. 1–24.
[21] J. Weizenbaum. *Computer power and human reason: From judgment to calculation*. W. H. Freeman & Co., 1976.
[22] L. Floridi and M. Chiriatti. "GPT-3: Its Nature, Scope, Limits, and Consequences". In: *Minds and Machines* (2020), pp. 1–14.