Uncanny Performance, Divergent Competence

Gabbrielle Johnson and Gabe Dupre[*]

word count = 11,448

> *"We all have an uncanny knack for empathizing another's perceptual situation, however ignorant of the physiological or optical mechanism of his perception."*
>
> – Quine, *Pursuit of Truth*

> *"Quine sees empathy as key to learning language and communicating, and calls it 'uncanny'. Indeed. His account does literally nothing to explain why such projections are successful, or even what their success consists in."*
>
> – Burge, *Origins of Objectivity*

1 Introduction

Recent years have witnessed an explosion in the capacities of so-called "Artificially Intelligent" (AI) systems, particularly with respect to language-using systems, such as the Large Language Models (LLMs) GPT and PaLM. Such systems are increasingly ubiquitous in daily life, and their transformation of our social, economic, and physical environments is far from complete. AI systems of this sort are able to perform at hitherto unimagined levels on linguistic tasks, such as writing essays, jokes, poetry, and much more. This raises the question: do such systems actually *know* or *speak* the natural human languages they seem to be using? Or are they merely mimicking the genuine mastery of human language users? Relatedly, when we interact with such machines, are we, or could we be in the not-too-distant future, *communicating* with them, or are our interactions merely causal? In this chapter, we shall argue that there are powerful reasons to worry that current approaches to linguistic tasks in AI are traveling along a path towards more-and-more compelling *illusions* of communication, not towards the genuine article, paving the way for harmful consequences of miscommunication. We contend that the inherent biases that are encoded in humans and machines are different, thereby raising doubts that our concepts, structured

[*] Order of authors determined by coin flip. In exchange, British-English conventions are adhered to throughout.

as they are by human-specific biases, are the same as superficially similar concepts in machines. This leads to critical miscommunications when, for example, we take an algorithm's use of some concept like RECIDIVISM RISK to be the same as our own and base real-world decisions on their predictions using those concepts.

The picture animating our argument is developed through comparison of biases shaped by the structure, the goals, and the development of AI systems with those of biological minds. We draw heavily on results from the last half-century or so in the cognitive sciences, which suggests to us a picture of mental processes which looks radically unlike that found in contemporary AI systems. These AI systems are biased toward minimizing prediction error through the use of powerful domain-general statistical tools which gradually squeeze all of the available information out of the massive databases on which they are trained. Human minds, on the other hand, are differently biased toward deeper understanding of causal-explanatory connections that underwrite statistical patterns, and so they deploy a wide range of developmental and cognitive styles that ultimately deviate from the statistical and predictive optimality of artificial systems.[1] For example, child language learners appear to disregard wide swaths of potentially relevant evidence (Rock 1957, Gagliardi et al. 2017, Trueswell et al. 2013), adopt idiosyncratic and domain-specific learning strategies (Yang 2017, Sakas and Fodor 2012), and rely heavily on innate constraints which need not reflect environmental regularities. We will argue that these sorts of observations, from several branches of the cognitive sciences, suggest that whatever is involved in the internal representational capacities of current artificial systems, they seem likely to be quite different from those of humans, undermining the prospects of harmonious human-machine communication.

The structure of our argument will be as follows: in section 2, we outline several historical problems of meaning for intelligent systems, sharpening our focus to the main problem threatening miscommunication: indeterminacy. We articulate philosophical responses to this problem that foreground the structure, the goals, and the development of the intelligent systems themselves in resolving indeterminacy, a point we familiarize for computer scientists through the framing of inductive bias. In section 3, we examine the biases of human cognitive systems; which we then contrast with the biases of artificial systems in section 4. In section 5, we use the foregoing discussion to argue that there are strong reasons to think that the kinds of representational contents available to these systems are radically misaligned. If our argument is in the right ballpark, this suggests that successful communication between humans and machines will not be possible without radical changes to the development, structure, and goals of artificial systems. We end by exploring the consequences of this prospect in section 6.

[1] Throughout, we assume a conceptual distinction between mere statistical prediction and understanding. Though a thorough analysis is beyond our scope, we can follow Burge (2003, 38, fn. 2) in regarding understanding to be "a type of knowledge that makes explanatory connections"; whereas, mere statistical prediction does not entail explanatory connections.

2 Facing the Problem: Indeterminacy and Bias

In machine learning, scientists often produce software for problems of classification and labeling, like trying to teach a machine to distinguish images of mallards from images of rabbits, or a spam email from a genuine one. More recently, with the introduction of LLMs, we've seen predictive systems that can support an even wider range of tasks, including translation, summarization, and dialogue. As LLM behaviour comes to more and more closely approximate human behaviours, philosophical questions arise concerning the extent to which the underlying systems causally responsible for such behaviours are truly similar.

2.1 From Externalism to Indeterminacy

One sort of question we might ask is whether such systems are capable of genuine representation. What makes us think their use of a symbolic label like 'rabbit' means anything at all? This is a question that has occupied much of philosophy and computer science. For example, Harnad (1990), citing Searle (1980)'s famous "Chinese Room" thought experiment, introduced this question to computer science under the heading of the "symbol grounding problem", asking how artificial programs ground their internal symbol's meaning to objects in the external world given that all their causal interactions are with intermediary representations, namely the text or photos on which they're trained.[2] It seems the computation underlying their operation—namely, the physical manipulation of symbols according to an algorithm—can occur without a system's knowledge of what those symbols mean. The symbol grounding problem questions how a system can understand symbols in a way that's meaningful beyond mere syntactic manipulation.

Externalism about meaning, a prominent philosophical answer to this question, has since been successfully extended to artificial systems. Externalism helpfully shifts the focus away from what we can deduce from restricting our investigation of some system to just its internal transitions from inputs to outputs. Rather, it suggests that the symbols internal to a system gain their meaning through causal interactions with the external environment.[3] The word 'rabbit' is meaningful to humans because it is connected to our sensory experiences of seeing, touching, and interacting in other ways with rabbits. Even if you yourself have never interacted directly with a rabbit, you have interacted with someone who has interacted with (someone who has interacted with...) rabbits. Likewise, for an AI system's internal symbol 'rabbit' to be grounded in the external world, there merely needs to exist a causal chain leading back from its tokening of that symbol to the external world itself, even if that chain takes a somewhat circuitous route through the causal intermediary

[2] See also recent critiques by Bender and Koller (2020) and Bender et al. (2021).

[3] The modern externalist tradition originates in the work of Donnellan (1966); Putnam (1975b); Burge (1979); Kripke (1980), and others.

of its training data.[4] It is possible for an LLM's symbols to refer to the external world, despite lacking direct encounters with these worldly referents, in virtue of the meaning carried by *our* symbols, which are grounded in these causal chains, and which the system does interact with (in training and prompting). To emphasize an important lesson in the history of philosophy of mind and language throughout the 20th Century: what matters are that the causal chains exist, not that we or the LLMs know about them.[5]

The integration of the externalist philosophical tradition into contemporary computer science continues to yield fruitful and productive results. Current approaches strive to construct extended models with more extensive and intricate causal connections to the external world. This is achieved by combining systems proficient in tasks like image recognition, language processing, and physical action. These integrated systems, which we (following Chalmers (2023, 2)) can refer to as LLM+ systems, aim to establish rich networks of perceptual and action-based interactions with the external world. The intention is to ground the meaning of their symbols in a manner akin to how human symbols are grounded through sensory perception and action.

This shift towards externalism has facilitated the convergence of questions posed in machine learning and philosophy, fostering valuable interdisciplinary dialogues. By emphasizing shared causal connections over diverse internal intellectual processes, a common foundation is established, offering a collaborative pathway forward.[6] We think this is a great example of how philosophy and computer science can provide mutual support in making progress towards fundamental theoretical and practical goals.

However, while this approach effectively addresses the general problem of how symbols refer, it does little to resolve a subsequent issue well-recognized in discussions about meaning and reference. Our paper seeks to advance this discourse by addressing the problem of indeterminacy in meaning, a problem which arises as an inevitable consequence of a broadly externalist approach. Once we establish that meanings result from causal interactions, the focus shifts to determining which aspects of the myriad causal pathways determine the precise meaning established. Notice that the strategies used to address one problem, such as the creation of LLM+ systems, offer little resolution for the subsequent issue. In fact, the additional causal pathways employed in sensory grounding can complicate matters, multiplying the potential causal antecedents that could serve as referents for the symbols used by the system, thus exacerbating the problem of indeterminacy. Our investigation pivots from the question of what enables words or symbols to possess meaning

[4] Philosophers explicitly extending externalism about meaning to AI include Cappelen and Dever 2021, Butlin 2023, and Mandelkern and Linzen 2023.

[5] This is just one point among many in this paper in which adequate analysis of meaning and reference will require us to resist epistemic constraints on the possibility of reference.

[6] While consensus in philosophy is rare, there is substantial convergence on these externalist themes. The most recent PhilPaper's survey suggest that at 58%, over half of participating philosophers among the survey's target group accept or lean toward externalism, while only 26% feel the same toward internalism (Bourget and Chalmers 2014). These numbers jump closer to 68% vs. 20% when restricted to just philosophers of language.

in a general sense to an exploration of what dictates their specific meanings. To summarize, while externalism establishes that symbols can possess accuracy conditions in general, the indeterminacy problem delves into what defines those conditions. What distinguishes certain label attributions, like ‘rabbit’ or ‘spam mail’, as accurate while deeming others inaccurate?

One potential solution is to rely on practical goals as a litmus test for successful reference, e.g., when an important email gets sent to the junk folder or a picture of a mallard shows up in a search for rabbit images.[7] However, as will become clear in the remainder of this chapter, this approach conceals sources of error that can manifest in the absence of any practical failure. Genuine differences in meaning can be irrelevant for various practical purposes, but these differences remain as explanantia for a theory of content. Determining which referent is the correct one isn’t as straightforward as merely assessing which option yields better system performance. In essence, performance does not establish competency.[8]

Once philosophers recognized this puzzle they soon noticed that it generalizes: considered purely physically, the same stimuli can, in principle, be associated with arbitrarily many different referents that are each inconsistent with each other but that all align with our practical objectives. This problem preoccupied much of philosophy of mind and language throughout the 20th Century. The general aim of this work was to determine in virtue of what a particular term or concept referred to what it did, i.e., for the externalist, what determined its meaning. The biggest threat to resolving this puzzle came in the form of charges that it was ultimately intractable, at least through standard empirical means. Quine (1960) led the charge with his famous argument from radical translation. Quine’s thought experiment asks us to imagine we’re trying to interpret a completely foreign language. What we have at our disposal are empirical observations: watching interactions of the users of the language and associating particular utterances with the objects or actions with which they co-occur. Let’s say we hear them utter the word ‘gavagai’ always and only when a rabbit hops by. It seems natural to think ‘gavagai’ therefore means rabbit. But equally consistent with our empirical data, says Quine, is that it instead picks out indefinitely many other alternatives: “rabbit time-slice,” “rabbit-shaped mass,” “light arrays that give the appearance of a rabbit,” or even “undetached rabbit parts”. How do we adjudicate between them?

Quine’s original example often strikes people (rightly) as philosophically contrived. It might seem obvious to most that, at least in the case of humans, the *recherché* alternatives of fleeting time-slices, mere collections of attached parts, patterns of light, etc. are

[7] Another alternative is to point to design intentions. If the system is designed with the intention to use the label ‘rabbit’ exclusively for rabbits, then that label refers to rabbits. Going this route, we would never have needed to appeal to externalism at all. In taking on a general externalist approach to the problem, we likewise take on the aim of exploring the possibility of intrinsic meaning (as opposed to merely imposed or derived meaning) for AI systems.

[8] As noted by Firestone (2020), this observation holds in both directions: we can achieve similar performance with different competences, and we can exhibit the same competence with varying performances. In general, observable performance data serves as a defeasible heuristic for underlying competencies.

unreasonable candidates of reference. But in the context of extending theories to artificial systems, our intuitions are arguably less forceful. Consider an image recognition program that labels certain images as ‘rabbits’. For example, Fel et al. (2023) use a method called CRAFT to construct heatmaps identifying the most influential regions of an image that impact categorization by a popular DNN (ResNet50). They found that images of rabbits are most readily categorized on the basis of pixels correlating with eyes, ears, and fur:[9]

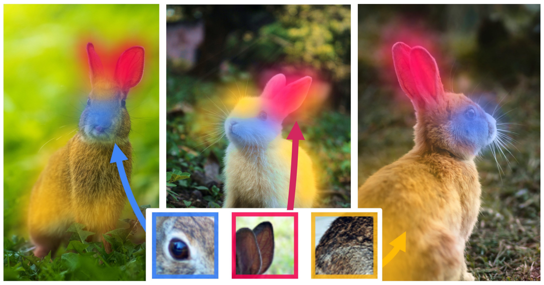

Figure 1: Heatmap generated by Fel et al. (2023) for categorization of rabbit.

While many rabbits have eyes, ears, and fur, and thus photos of rabbits typically contain regions of pixels containing information about these traits, many non-rabbits might also have such features, or at least produce photos with similar enough pixel arrangements. What, from the perspective of the machine, determines that images of this latter sort are *inaccurately* classified with the label ‘rabbits’? Could we just as well regard the label as accurate of images of rabbits, or anything that looks like a rabbit, or mere seemingly arbitrary pixel value arrangements?

This suggests that even when using the same labels and applying those same labels in roughly similar circumstances, different systems can be talking about fundamentally different things, leading well beyond the *recherché*. To really emphasize the import of such divergences, consider cases with more overt ethical consequences, such as the well-publicized “gaydar” machine developed by researchers in Stanford’s business school.[10] This system was an image classifier, with the function of predicting a person’s sexuality solely on the basis of an image of their face (collected from online dating profiles), and it managed to do so well above chance. Ignore for a moment the ethical and social concerns the mere use of such systems might raise, and consider only the semantic question: what do the labels ‘homosexual’ or ‘heterosexual’ *mean* when applied by this system? It is difficult not to interpret them, as do the many thinkpieces on this system, as meaning just what these terms mean in *our* mouths. That is, it is natural to view the outputs of this system as saying something like “the person in this image is heterosexual”. But this should give us

[9] Image courtesy of Fel.

[10] Wang and Kosinski (2018). This case is discussed also by Johnson (2023a) in service of similar points.

pause.

The internal categories developed in the training of this system are labeled with English expressions 'heterosexual' and 'homosexual'. But the relationship between the meanings of these expressions and the content of these representations is far from clear. Like the use of CRAFT demonstrated of the rabbit images above, what these representations actually consist of are a host of statistical properties defined in terms of regions of the images it has been trained on, which can likewise be revealed by looking at the composite images for faces labeled as each:[11]

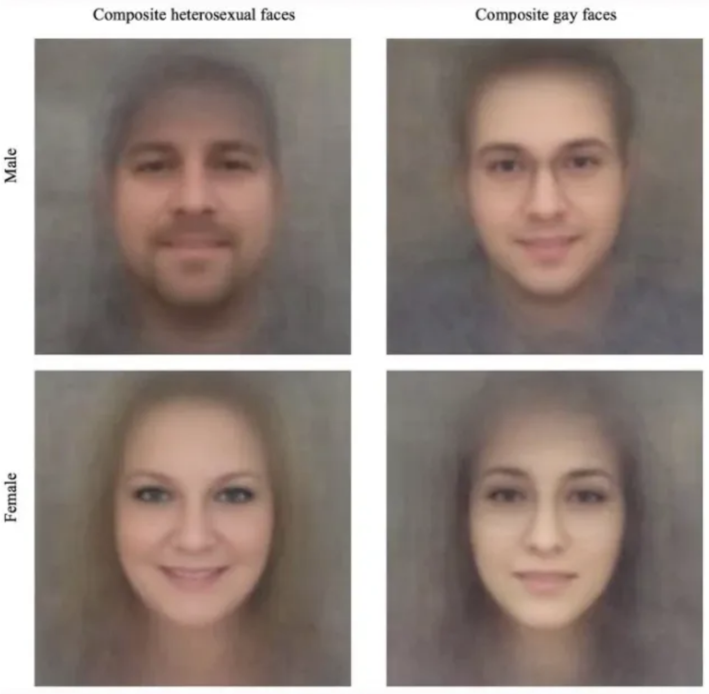

Figure 2: Composite faces built by averaging images classified as most likely to receive the label 'heterosexual' (left) and least likely to receive the label 'heterosexual' (right).

As noted by Agüera y Arcas et al. 2018, one can see that some of the main cues that this system used to determine what label to apply to a novel image appears to be whether the person is depicted as wearing glasses, or makeup, or having facial hair. Ask yourself: do these superficial traits really exhaust the meaning of the term 'homosexual'? One doesn't need a degree in Queer Studies to identify the correct answer here. What these systems have managed to do is identify certain statistically prevalent features of the faces of homosexual and heterosexual people—in essence, visible stereotypes. There

[11] Image borrowed from Agüera y Arcas et al. 2018, which itself borrows from Wang and Kosinski 2018, 251.

is nothing more to their internal classification system than these purely visual, statistical traits. Our concepts, however, are in no way reducible, or even semantically dependent on, the classification systems adopted by such a system. Whatever statistical correlations may exist, any minimally enlightened person knows that terms that apply to people on the basis of their sexuality are entirely dissociable from features of appearance. While we may rely on stereotypes when categorizing people into groups, we of course recognize the distinction between the two. We say things like, "that's just a stereotype" or "he doesn't fit the stereotype you might expect" precisely because the two can come apart for us. Not so for the AI system—the gaydar machine will never treat differently a homosexual or heterosexual person, so long as both fit the same stereotype. Thus, unthinkingly assuming that the labels 'heterosexual' or 'homosexual' as applied by these machines mean the same thing as these terms do in English may be a serious mistake.

2.2 From Indeterminacy to Bias

To better understand the general implications of the problem of indeterminacy for miscommunication and potential solutions to this problem, we'll start by distinguishing between what we might call "deep concepts" and "superficial concepts". The distinction here has roots at least as far as Locke (1996)'s distinction between real and nominal essences, but took on new life in many of the classic papers of 20th Century philosophy of mind, language, and science. Roughly, the common source of the distinction is meant to capture differences in what we might call the epistemology of representations—i.e., observations that a given agent relies on in determining whether a given concept (or linguistic expression) applies in a given context—versus what we might call the semantics of representations—what constitutes the meaning of the concept.

Roughly speaking, a "superficial concept" is one for which the extension of the representation (the class of objects picked out by the term) is determined by the epistemological features of the representations: the cues used by the agent in determining whether some entity is in or out of the category. In the case of 'rabbit', if what we use to pick out the extension are the features we have perceptual access to, like having big ears, small eyes, and a particularly shaped furred body; then the extension would be constituted by all and only creatures with those features, and the concept would be superficial. In contrast, a "deep concept" is one for which the extension of the representation is determined by something other than the cues used in determining whether some entity is in or our of the category. This would allow for some creatures that lacked some (or even all) of the directly perceivable features of rabbits to still be in the extension of the deep concept. Likewise, some creatures that had some (or even all) of those features might fail to be in the extension, like (the now-famous go-to for philosophers of) cleverly disguised robots. In other words, superficial concepts are those that constrain reference via epistemological competencies, whereas deep concepts allow for reference beyond epistemological competencies. Following Putnam (1975a, 139)'s description of natural kinds, we can say that the extensions of deep concepts are constituted by “classes of objects whose normal distinguishing characteristics are ‘held together’ or even explained by deep-lying mechanisms”. In contrast, the extension of a “superficial” concept can be likened to Putnam (1975a, 140)’s description of those constituted via “specifying a conjunction of properties”.[12]

Examples are illustrative. Sticking with Putnam, consider a term like ‘lemon’. Putnam claims that one approach to specifying the kind picked out by this term is through first, specifying a set of superficial properties (being small and round, being yellow, being sour, etc.) and second, claiming that anything with all the properties within that set constitutes a lemon. Crucially, he argues that such a view of meaning for natural kind terms is untenable, since (in summary) such a collection is neither necessary nor sufficient for being a lemon: something can be a lemon without having any of the properties mentioned in the conjunction and something can have all the properties mentioned in the conjunction and not be a lemon. According to Putnam (1975a, 140), to designate some object as belonging to a natural kind is to not only identify a conjoined list of superficial properties belonging to members of that kind, but also to commit to there being some shared ‘essential nature’ that grounds the causal-explanatory connections to those superficial properties.

This general distinction between kinds that are individuated on the basis of superficial properties and kinds that are individuated on the basis of deep, shared causal-explanatory structures has come to dominate contemporary theories in philosophy of language, philosophy of mind, and empirical psychology. For example, Burge (2013, 237), in summarizing Putnam’s contributions to semantics, writes that Putnam’s separating of “the superficial features of natural kinds ... from the underlying nature or essence of the kind” anticipates a general theoretical concept of natural kind, wherein “having a natural kind concept requires being open to a distinction between what a thing is and how it veridically appears”. This likewise matches what we have in mind by “deep” concepts. That is, for deep concepts, one will be open to a distinction between what a thing is and how it veridically appears. For superficial concepts, belonging to the class picked out by the concept is determined entirely by superficial properties; thus, there will exist no distinction between what a thing is and how it veridically appears.

A major consequence of this radical dissociation between the semantics and epistemology of representation is that intelligent agents can be radically mistaken or ignorant about the observational properties of the referent or extension of their own concept, again driving a wedge between competence and performance. In such cases, while their ignorance may undermine their successful *use* of the concept, it does not undermine their *possession* of it. To modify Putnam’s example, someone who believed that lemons were characteristically blue, spherical, and sweet would do a very poor job of identifying lemons on the basis of their visual, tactile, and gustatory properties, but would nonetheless be capable of (mis-)applying the concept LEMON. Once more, practical success or failure proves inadequate as an indicator of meaning or reference. Accurate description of their misunderstanding

[12] Deep and natural are related, but not the same.

requires attribution of the concept LEMON to them: it is because they have these false beliefs *about lemons* (i.e., beliefs which feature the concept LEMON) that they are so bad at classifying (and, presumably, cooking). As Burge (2007) (p. 163) puts it "[o]ne can master a concept well enough to think with it without understanding constitutive principles that govern its usage".

Whether a system is biased toward superficial or deep concepts is determined by its structure, goals, and development.[13] Computer scientists will be well aware of the need to consider a system's goals in determining its in-built biases. For example, when you train a neural network to recognize handwritten digits, you inherently bias it towards certain kinds of solutions by choosing a specific architecture, activation function, and even training data. Such inductive biases come in the form of assumptions that a learning algorithm makes to predict outputs for new, unseen data based on the data it has already encountered and, as noted by Mitchell (1980), and widely repeated in contemporary debates about the capacities of ML programs, these assumptions are essential for the algorithm's ability to generalize.[14] Without some form of inductive bias, learning is impossible, and which sorts of inductive biases are included dictate the capabilities of the system. Thus, a system's biases reflect in total its structure, goals, and development.[15]

To summarize the argument to follow: by examining the structure, goals, and development of each, we can see that machines are biased toward superficial concepts whereas humans are biased toward deep concepts.[16] Present machine learning models function to capitalize on statistical correlations. They use data to produce a predictive model based on similarity patterns within the data. What makes them so powerful, in theory, is that they can track many more features (and, thus, many more similarities between their combinations) than human minds can. With enough data, predictions made on the basis of statistical correlations can become quite practically useful. They excel at tasks like image recognition, language translation, and recommendation systems. However, they operate under the limitations of their training data and the specific algorithmic techniques used to process that data. Crucially, machine learning programs need not uncover the

[13] This captures roughly a teleological approach to resolving indeterminacy, as championed by Millikan (1984), Dretske (1986), and Burge (2010). We likewise take the view that bias is constitutive of content fixation to have precedent in the literature, though we lack space to fully explore the details of that analysis here. For example, Burge (2010, 232) specifically argues against Quinian indeterminacy by noting that Quine misses humans' "bias toward environmental macro-entities". Our activities and goals bias us toward some entities as referents rather than others. These biases are encoded in the system. Thus, when we speak of bias, we mean the internal assumptions that are made in order to reflect environmental regularities and human-environmental interactions, to limit the available options for content formation, and to guide content formation processes. See also Burge 2005, 12-23 and Burge 2010, 370-371.

[14] See relevant discussion in Teney et al. 2022a,b and Zhang et al. 2017, 2021.

[15] Some might be uncomfortable with the equivocation between bias in general and inductive bias specifically. We think fundamentally the two share similarities relevant for the equivocation in this context. For philosophical discussion, see Johnson 2020, 2023b.

[16] For related work applying broadly teleological theories of meaning and reference to artificial systems, see Butlin 2022, 2023.

deeper causal-explanatory connections that underwrite the patterns they recognize; their primary goal is to optimize predictive accuracy, which does not entail the building of robust causal-explanatory models.[17] In contrast, contemporary cognitive science suggests human intellectual capacities function beyond mere prediction. Human cognition aims for a deeper understanding of the world, targeting causal and explanatory connections that underlie observed phenomena. Humans don't just recognize patterns; they seek to understand the principles that generate those patterns in the first place. Deep concepts rather than superficial are best suited to aid in these tasks.

3 Human Competencies

Humans are biased toward deep concepts because human intellectual endeavors go beyond, and are sometimes at odds with, mere prediction. Human understanding aims at uncovering deep causal-explanatory connections that are in principle dissociable from mere prediction. We take this to be a lesson iterated many times over in the history of studies of human intelligence, development, and scientific practice. Here we explore paradigms of this general trend.

Of course, viewing intelligent behaviour as essentially a matter of responding in statistically predictable ways to the informational patterns in one's environment has a long history in the mind sciences. Behaviorists like Watson and Skinner eschewed talk of complex inner lives, instead attempting to directly relate environmental stimuli to behavioural responses conditioned by the positive or negative reinforcement of prior stimulation. This approach thus fit nicely with early 20th Century positivist and instrumentalist approaches to science generally, which encoded scepticism towards unobservable theoretical posits, restricting science to prediction. The oft-told tale of the fall of behaviourism (c.f., Chomsky 1959, Fodor 1968) centers precisely the need to appeal to the subtleties and complexities of internal states in predicting, and more significantly explaining, intelligent behaviour. We see in these early debates the germs of the issue we face in this paper: what role do the specific features of the intelligent system under investigation (humans, non-human animals, or machines) play in structuring the internal states they are or can be in, beyond the shaping of such systems by their environment.

[17] It is often suggested (e.g., Chalmers 2023, 13) that minimizing prediction error would plausibly lead to the generation of causal-explanatory theories (in the form of "robust world-models"). However, this proposal seems to be little more than a guess as to the future development of such systems. It is worth stressing also that even if AI systems do generate world-models of this sort to enhance their predictive powers, there is no guarantee that their world models will be relevantly similar to ours, drawing category distinctions in roughly the ways that we do. The themes concerning indeterminacy and underdetermination that run throughout this paper suggest that there will be indefinitely many available world-models consistent with observations and predictions made by a system. Analogous worries apply to Clark (2015, 19)'s claim that "one way to learn a surprising amount about grammar ... is to look for the best ways to predict the next words in sentences".

Note that, while many philosophers have traditionally offered *a priori* arguments for claims that human mental states feature structure and content that goes beyond what can be captured by superficial similarity in the service of mere empirical prediction, such a claim is not trivial, and working out which (if any) human competencies aim to go deeper is a rich on-going empirical enterprise.We can imagine a representational system that precludes any distinction between superficial and deep concepts within it; i.e., without a distinction between semantics and epistemology. In such a system, the meaning of the representational types is constituted by the set of properties used in practical acts of classification. Indeed, there are programs within philosophy and psychology which view human concepts in just these ways, and we shall argue later that contemporary artificial systems seem to work in this way.

Perhaps the most prominent empirical research program which views human concepts as superficial in the way just defined is the *prototype* theory of concepts. Stemming from seminal work by Rosch (1973), prototype theorists view human concepts as statistical summaries of the characteristic features of the entities covered by the concept. One of the most attractive features of the prototype theory is the elegance with which it ties together concept acquisition and concept application. Acquisition of a concept, on this view, is a matter of learning which features a particular kind of entity tends to exhibit. For example, in the process of acquiring the concept RABBIT, a learner might be confronted with a decent number of rabbits. This provides them with a wealth of statistical information. For example, they might notice that almost all of the rabbits they encounter have long ears, small eyes, bushy tails, etc. (allowing for the occasional hare or injury). They might notice that fewer, but still the majority, binky and dig, eat grass and cecotropes, and sleep during the day. And further that none of them are bigger than a bakery or speak fluent Korean. Some of these traits, such as binkying, might be found very rarely in other encountered creatures. A somewhat sophisticated statistical reasoner could then leverage this information into a useful set of cues for determining whether some newly encountered entity was indeed a rabbit. If the creature in front of them was hairless, this would then be very good, but not decisive, reason to think it is not appropriately categorized as a rabbit. And while its abstaining from binkying would be insufficient for ruling it out, if it did binky, this would be very good reason to categorize it as a rabbit. And so on. The prototype theorist turns these rational epistemic inferences into a full-fledged theory of concept-possession. What it is to have the concept RABBIT is just to have identified these evidential relationships, and to have some stable unified representation of them, which can then be applied to determine the likelihood of category-membership to novel cases.

While such correlational information is useful for practical purposes of classification and application, for creatures like us, biased towards deep concepts, it does not serve to define or determine the content of our concepts. We believe that the ways that human concepts go beyond mere statistical summaries of characteristic properties is well established in the empirical study of the mind. Our ability to dissociate such common properties used in classification from the content-determining aspects of concepts is a deep and essential feature of human psychology. These correlations may enable us to 'latch on' to our concepts' worldly referents, but once we have done so we can, as it were, kick away the ladder and think, and talk, about these referents directly, without requiring that the referents have, or are identified by way of, these superficial properties. As Fodor (1998, 140) puts it "our minds are, in effect, functions from stereotypes to concepts". We turn now to a selection of evidence for this claim.

3.1 Psychological Essentialism

Some of the most direct evidence that human cognition does not simply involve identifying probabilistic clusters of detectable properties comes from the literature on *psychological essentialism*.

In the philosophical literature, the rejection of broadly descriptivist accounts of linguistic and conceptual meaning led to an endorsement of essentialism.[18] Kripke and Putnam argued that, if the reference of our terms like 'water' and concepts like WATER are not determined by the properties we associated with these representations (e.g., our beliefs that water is the colourless thirst-quenching liquid that comes out of our taps), they must instead somehow 'latch onto' the metaphysical essences of these worldly entities, via some broadly causal mechanism, enabling these expressions to apply to the same kind of thing in any actual or counterfactual contexts in which the superficial properties of the referent in question may differ. Essentialism in this sense is a metaphysical doctrine, concerning what determines the membership of a given category (essence, rather than satisfaction of a description).[19]

Psychological essentialism is not a metaphysical thesis, but a psychological one, concerning the modes of conceptualisation that human minds are prone to. This is the view, roughly, that human minds tend to categorize *as if* essentialism were true, whether or not it actually is. To categorize in an essentialist fashion is to treat category membership as a reflection of some underlying, unobservable, unchanging feature of the (potential) members, which plays some causal-explanatory role in accounting for similarities between members.[20]

Some paradigmatic evidence for the essentialist biases in human thought comes from Gelman and Markman (1986), who introduce the 'triad test'. This test investigates the extent to which children's inductive biases are based on perceptually available features. Tests present subjects with (pictures of) three objects, two of which are labeled with the same kind term, while the third, differently labeled, object is perceptually similar to

[18] Although see Salmon (1979) for an argument that such an inference is unwarranted.

[19] Essentialism in this sense is widely rejected, at least as a general account of natural kinds and natural kind terms, due to its apparent conflict with evolutionary biology. See Dupré (1993) for a classical statement of this point.

[20] See Neufeld (2022) for a recent survey of evidence for, and interpretations of, psychological essentialism, as well as references.

one of the same-label stimuli. Consider the following image borrowed from (Gelman and Markman, 1986, 188):

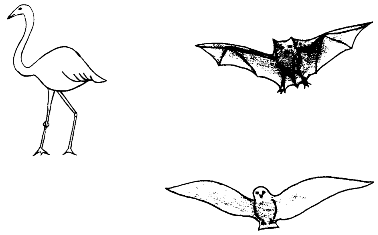

Figure 3: Triad test pitting essentialist biases against perceptual similarity.

This test consists of pictures of a flamingo, a bat, and a blackbird. While the flamingo and the blackbird are both labelled ‘bird’, the bat (labelled ‘bat’) is perceptually much more similar to the blackbird than either are to the flamingo (in shape, colour, etc.). Some information is then offered about both the unique-category object (the bat) and the unique-perceptual-appearance object (the flamingo), e.g., “the flamingo’s heart has a right aortic arch only” and “the bat’s heart has a left aortic arch only”. The subject is then asked what they expect to be true of the final object (the blackbird). Do they extrapolate from the object in the same category, or the object which looks similar? What they found was that children tended to (roughly 2/3 of the time) extrapolate from the shared-category object and not the perceptually similar object. In other words, children thought the blackbird would have features in common with the flamingo despite looking more similar to the bat. The authors (and much of the substantial literature following them) conclude that human categorization characteristically allows for inductive biases shaped by hidden, kind-determining, features, e.g., “essences”, to provide a better basis for inductive inference than does perceptually available similarity.[21]

This research thus comports well with the Putnam-Burge observations concerning (some) human concepts, and the ways that they go beyond mere probabilistic description of a category, and against the proposals of the prototype theorist. The essentialist bias in human thought is representative of the ways in which human concepts are “deep”. They do not reduce to observed regularities, but expect that such regularities are due to unobserved, and unobservable, real structures. And it is these underlying structures that play the role of content-determination, not the observed regularities themselves.

[21] While much of the literature on psychological essentialism suggests that this role be filled by essences in the traditional, internal sense, recent work on structural theories of explanation in the child categorization tasks (e.g., Vasilyeva et al. 2018; Vasilyeva and Lombrozo 2020) suggest that the hidden, kind-determining features that play this role might be more expansive than initially thought. Our major claims about the implication of these empirical theories are consistent with either approach.

3.2 Syntactic Bootstrapping

Another area in which it is well-established that human cognitive capacities are not reducible to the acquisition and application of statistical summaries of observed correlations is developmental linguistics. It is widely accepted within the study of language acquisition that what a child learns goes well beyond what is available in their environmental linguistic input.[22] This observation, that the stimulus for learning is impoverished relative to what is learned, is of course most famously discussed in the acquisition of grammar.[23] But what matters for our purposes is the application of this point to the acquisition of lexical meanings.

The question of how humans learn the meanings of the terms of their native language has long been a central focus of philosophy. An intuitive proposal, again traced back at least to Locke, is that we learn the meaning of a word by associating the presence of the word with the presence of its referent. A child can notice that the probability of hearing the word 'rabbit' increases when rabbits are around, and thereby come to associate the linguistic expression with the worldly referent. Such an approach meshes nicely with the above-described prototype theory of concepts: the child acquires the concept RABBIT by identifying dependencies between observable properties in the environment, and then the acquisition of a label for this concept is just a matter of noting one more environmental regularity, distinctive only in the kinds of environmental properties it relates.[24]

One major difficulty with this view, if understood as a general account of lexical acquisition, which has been elaborated on over the years by Lila Gleitman and many others, is that it makes empirical predictions/presuppositions about linguistic behaviour and the language-learning environment which do not seem to be met. For the child to learn associations between the use of a term and some feature of the environment, there must, of course, be such an association. Further, the environmental feature with which the expression is supposed to be associated must be, in some way, attention-grabbing or perceptually distinctive, to ensure that the child forms the correct association. For a wide range of perfectly learnable expressions, however, there are strong reasons to doubt that these conditions are met.

The most obvious kinds of case come from the acquisition of words which apply to environmental conditions that are always ("air", "language", "matter", etc.) or never ("angel", "astatine", "1000") present in the learning environment. And even when word-environment correlations between 0 and 1 do exist, they may often not cut the way such theories of lexical acquisition predict ("Grandma" and "Jamaica" may well be used much more frequently *in absentia*). Specifically, even when there are such correlations in principle available to the learner, there are reasons to doubt that the learner makes use of them.

Building on famous work by Rock (1957), Trueswell et al. (2013) tested the extent to

[22] This point is not, unfortunately, universally accepted; See, e.g., Chater et al. 2015.

[23] See, for recent statements, Lasnik and Lidz 2016 and Crain et al. 2021.

[24] Notice also the similarities with Quine's 'gavagai' thought experiment.

which word-world correlations were in fact identified in the process of word learning. They presented subjects with a series of simplified word learning environments, consisting of a collection of 2-5 stimuli, potential referents, and a novel label. Each trial consisted of an exposure to one such environment, and after each trial they were asked what the label referred to. In each subsequent trial, a novel instance of the 'correct' referent was featured (e.g., if 'zud' meant book, then all trials would contain different pictures of books), while non-target referent categories were restricted to appearing at most twice over the five trials (e.g., a shoe could appear in at most two trials if the label in question did not mean shoe). The learner was not told whether or not their guesses were correct after each trial. In the first trial, of course, the subject was at chance (between 20% and 50% depending on the number of stimuli), simply guessing one of the available referents. The interesting question was what happened on subsequent trials. Specifically: did the learner retain relevant information from previous guesses and use this to inform subsequent guesses? An ideal statistical learner would, in a case like this, identify several hypotheses (one for each label-stimulus pairing), and update these hypotheses as new evidence (additional trials) came in. The rational thing to do would be to assign a probability of 1/n to the initial hypotheses, which would then get increased or decreased on subsequent trials based on whether the hypothesized stimulus was or was not present. The observed pattern of learning, however, was quite different. Subjects appeared to disregard all the information from previous trials, except for consistency with their guess. So, if a learner hypothesized that the label referred to books, they would retain this hypothesis until the trial didn't contain any books. But at this point, their guesses were independent of the number of times stimuli in the current trial had occurred in previous trials. That is, they would form hypotheses with the same distribution as if the novel stimuli were presented on the initial trial (i.e., with a 1/5 chance when there are 5 stimuli). Note the information lost by such a strategy. Assume the worst case for the learner, in which the initial two trials contain both the correct referent and an instance of an incorrect referent (e.g., when 'zud' means book, but trials 1 and 2 contain both books and shoes). If the learner guessed that 'zud' meant shoe in trials one and two, they will be forced to reject this hypothesis on the third trial. In principle, they could use the information that books have appeared in both previous trials (and that no other stimulus in the current trial has occurred in all previous trials) to increase the likelihood of the correct hypothesis over any other (in fact, to guarantee it!). But they don't do this. The likelihood of guessing the correct hypothesis was found to be the same as that of all now-available incorrect hypotheses relating the label to the remaining stimuli. Thus, it is concluded that language learning is quite unlike the rational correlation-seeking assumed by traditional empiricist approaches.

As well as ignoring relevant environmental information concerning word meanings, human language learners seem to contribute aspects of word meaning which are not extracted from the environment. Such innate contributions will be particularly crucial for cases of word acquisition in which word-world correlations seem unable to distinguish non-synonymous expressions. Just as Quine originally pointed out, there are arbitrarily many correlations between the use of a linguistic expression and features of the environment, even in the intuitively “good” cases in which the expression is used in the presence of its referent, so long as there are no constraints on how we specify environmental features. This again shows the generalizability of Quine’s original case without resorting to the *recherché*: each instance in which a child perceptually encounters a chasing, they are also perceptually encountering a fleeing; each environmental case of knowing is also a case of believing (not to mention breathing...); and so on. How then does the child learn what “knows” means, or distinguish “chase” from “flee”, and so on?

Exactly how to answer this question is both complex and controversial. But one consistent theme in the literature is the appeal to innate and universal expectations about the form that acquired linguistic knowledge will take. In a series of papers (e.g., Gleitman et al. 2005; Lidz et al. 2003; Papafragou et al. 2007; Lidz 2020) Lila Gleitman and her collaborators have decomposed the word-learning process into three primary components: (i) observed word-world correlations, (ii) observed inter-linguistic patterning, and (iii) innate syntax-semantics correspondences. We can exhibit all three in discussion of the distinction between verbs of contact and verbs of change of state (Fillmore 1970). Verbs like ‘hit’, ‘strike’, ‘kick’, etc. describe events in which something comes into contact with something else, but do not entail any changes to the contacted entity (i.e., one can hit something without damaging it in any way). In contrast, verbs like ‘break’, ‘damage’, or ‘dent’ describe events in which some relatively permanent change is induced in an entity, but are officially silent on how this came about. Despite this semantic neutrality, these two classes of verb are quite similar in their extensions: many encountered hittings will also be breakings, and many encountered breakings will result from contact. This poses a problem for the language learner: when they hear someone describe a scene using, say, ‘hit’ or ‘break’, the scene is liable to contain both contact and a change of state. How then do they distinguish the meanings of these expressions?

Note first that the confounding between ‘hit’ and ‘break’ does not eliminate the role of word-world correlations entirely; it just shows that it cannot do all of the work. The fact that ‘hit’ and ‘break’ are both more likely to be found applied to scenes in which breakings occur than in scenes without breakings, and that in such situations other classes of verb (e.g., statives or propositional attitude verbs) are less frequently found, provides significant information for the child to narrow down the semantic space. Having narrowed it down, however, the child still needs a way to distinguish between these classes of expression which apply in very many similar scenarios. It is at this point that inter-linguistic correlations can be of use. While ‘hit’ and ‘break’ will apply to many of the same scenes, they will do so in grammatically different contexts. For instance, while both can be used transitively (“Arturo hit/broke the vase”), only the verbs of change of state can be used intransitively (“The vase broke” vs. *“The vase hit”). This gives the child a way to distinguish the two. However, it is one thing to notice that these two apparently referentially similar terms have different syntactic distributions, it is another to leverage this fact into a way of determining the distinct meanings associated with them. It is here that the innate contribution taking the form of an inductive bias is supposed to enter: children, according to this account, come to the task of language acquisition with presuppositions about the relationship between syntactic frames (transitive vs. intransitive, nominal arguments vs. clausal arguments, single vs. double object, etc.) and semantic properties. At an abstract level, verbs of contact specify physical *relations* between two distinct entities. Transitive clauses are thus well-suited to capture the crucial information conveyed by such verbs. Verbs of change of state, on the other hand, specify changes to properties of a single entity, and are thus aligned with intransitive structures.[25] Crucially, these relations between syntactic structures and semantic properties are not, could not be, themselves learned from the environment. Their job in language acquisition is precisely to provide, and select between, differing perspectives on one and the same environmental situation. If this model is correct, it seems that what the child learns when they learn the meaning of a verb is typically not reducible to features of the scenarios in which the child has encountered the verb used.

These results from developmental linguistics point towards two major conclusions concerning the semantic properties of acquired lexical meanings. Firstly, they establish that the representations associated with acquired lexical items are not statistical summaries of observed regularities. The information neglected in acquiring lexical items would be quite unexpected from the empiricist approaches to learning, such as the prototype theory and, as we shall see, the approaches embedded in contemporary Deep Learning, which views learning of all sorts, including language acquisition, as the rational extraction of statistical information from encountered stimuli. Secondly, these acquired representations partially comprise information *not* found in these stimuli. That is, they are significantly structured by innate contributions and inductive bias. This is another way in which such representations are deep, rather than superficial.

4 AI Competencies

The above case studies from cognitive science demonstrate that human cognitive development, including concept and language acquisition, is not solely a matter of identifying and extrapolating patterns from the environment or sensory stimuli. However, AI systems, especially those within the prevailing deep machine learning paradigm, do develop in precisely this manner. These systems are, at root, pattern-completers. While modern, highly complex AI systems can extract and generalize from ever more subtle and abstract patterns, they retain this broadly empiricist developmental style. And this has significant repercussions for the kinds of content they can acquire.

[25] On this view, that we find transitive uses of change of state verbs requires a bit more machinery. Specifically, such structures describe two distinct events: an event of some property changing and an event of some external subject causing the former. This prediction is born out in much work on the morphology-syntax interface. See, e.g., classics such as Borer 1991 and Hale and Keyser 2002 and the recent literature review in Tubino-Blanco 2020.

Consider the two most prominent recent applications of deep learning that have been the focus of this paper up to now: LLMs and image classifiers. Our contention is that the representations associated by these systems with a particular symbol are exhausted by the probabilistic relationships the system identifies between it and other symbols in the system. For example, the representations associated by an LLM with a particular word are exhausted by the probabilistic relationships the system identifies between this expression and others in its language. These probabilistic relations may be complex and abstract. It may identify various strengths of complementarity between expressions (e.g., 'kick' and 'kiss' occur in many, but not all, of the same sentential contexts), as well as various "thematic" relations (e.g., 'pitcher' and 'baseball' occur in similar texts). It may further identify higher-level patterns, such as that expressions with sentential distributions similar to those of 'kick' and 'kiss' tend to be collocated with terms with sentential distributions similar to those of 'the ball' and 'his husband' (when we anthropomorphize such systems and describe them as recognizing that 'kick' and 'kiss' are transitive verbs, this is the sort of information that underlies this attribution). And so on. But all of this information is strictly reducible to probabilistic relationships between expressions.

We can tell a similar story for image classifiers. These systems identify, typically highly abstract, probabilistic relations between distributions of pixels and categories of images. As we've seen, images of rabbits will display high-level probabilistic relationships to one another, on the basis of intra-image relationships between regions. When trained on a database containing a number of images of rabbits, such systems identify these statistical relationships, and can extrapolate from them to a wider range of possible images with similar properties. When confronted with a novel image, it can classify it in a similar way on the basis of its similarity to those images it has seen. Once again, the system's representation of a RABBIT(-IMAGE) is exhausted by these statistical facts concerning how likely particular distributions of colour are for images of a certain category.

Representations formed in this way are, by their very nature, superficial. To the extent that LLMs know the meaning of words like 'kick' or 'pitcher', what they know are associations, extracted from encountered linguistic data, between these words and other expressions. To the extent that image classifiers know the meaning of labels like 'rabbit', what they know are associations, extracted from encountered images, between distributions of pixels. There is nothing more to these representations than what a sophisticated statistical reasoner could identify in the data they are trained on. While with humans we can distinguish the semantics of a given concept (what the concept is *about*, what it is a concept *of*) from the epistemology (what the human agent *knows about* the (referents of) the concept, and what the agent uses to determine whether the concept applies to a given stimulus or not), no such distinction is available for concepts that are, in our sense, superficial. This is clearly demonstrated for the LLM and image classifier: there is nothing more to these systems' internal representations than the associations they have identified, and nothing more could play a role in determining whether a given representation is applicable to a given stimulus or not.

Although theoretical discussions about the structure, goals, and evolution of intelligent systems are insightful, they gain more substance when supported by empirical research. This research examines the actual structures and abilities that drive the performance of these systems. However, empirical investigations into artificial systems are relatively new and still developing their methodologies. This limits their theoretical import. Nevertheless, we will explore two empirical studies that delve into the inner workings of AI competencies. These include anti-essentialist categorization and adversarial images, which respectively investigate the abilities of Large Language Models (LLMs) and image-classification systems.

As forceful as these theoretical reflection on the structure, goals, and development might be, like in the case of human competencies, they are best bolstered by empirical work probing the actual structures and capacities underwriting the performance of intelligent systems. However, unlike in the case of human psychology, empirical work on the nature of artificial systems is both recent and nascent in its guiding methodology, limiting any appeals to it to bolster the theoretical case.[26] Still, in what follows, we look to two recent empirical attempts to probe the inner structure of AI competencies: anti-essentialist categorization and adversarial images, exploring the competencies of both LLMs and image-classification systems, respectively.

4.1 Anti-essentialist Categorization

LLM capabilities are changing rapidly. Still, early work probing their propensity for essentialist categorization is available. For example, Zhang et al. (2023) investigated the extent to which essentialist beliefs about categories are transmitted through language by analyzing question-answer response patterns in LLMs. In this analysis, they submitted to the LLM vignettes from the classical empirical literature on psychological essentialism. For example, one such vignette presented a scenario where scientists modified a bee to possess spider-like visible characteristics, such as removing its wings, adding legs, and enabling it to make webs. They then explored to what extent the LLM would regard it as a bee or a spider.

The results from one such experiment are presented in this graph:

[26] We recognize that cross-disciplinary work spanning philosophy, psychology, and technology is encumbered by the vast and rapidly evolving bodies of theoretical work in each field, with the study of human intelligence encompassing the entirety of human history and technological advancements progressing at an unprecedented rate. Thus, contributions risk becoming obsolete swiftly. However, this reality only reinforces our dedication to understanding and comparing these processes, given their ever-increasing importance.

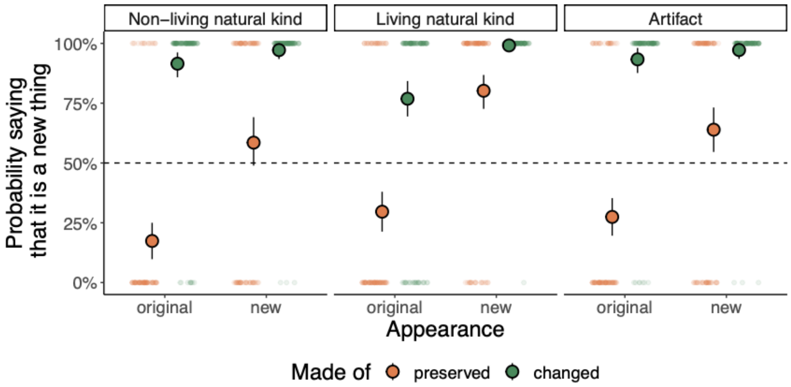

Figure 4: Figure borrowed from Zhang et al. 2023, 4.

The data most relevant to our current discussion is the central column 'Living natural kind'.[27] This graph presents the likelihood that an LLM would classify a stimulus as a member of a novel category (e.g., a spider when it was introduced as a bee), subject to various interventions. At the bottom left, we see that even when the "scientists' interventions" are unsuccessful (i.e., neither its appearance nor what it is made of was changed), the LLMs tested categorized it differently roughly 25% of the time. A puzzling result, but one we will ignore for the purposes of this essay. The conditions which matter to us are indicated in the second and third column. These indicate the likelihood that the stimulus would be categorized differently subject to interventions which modify what it is made of (column 2) and what it looks like (column 3), respectively.[28] Crucially, LLMs seemed to treat these interventions in pretty much the same way, viewing such changes as sufficient for recategorizing the stimulus roughly 75% of the time. Contrary to the authors' claims (although keep in mind their difference in focus), this seems a strikingly *anti-essentialist* behaviour. Changing the superficial properties of, say, a bee is just as effective a way, in the eyes of the LLM, of turning it into a spider as is changing its internal constituents! But it was precisely the relative unimportance of superficial properties in human classification that motivated the label 'essentialism' in the first place. While still a nascent area of empirical investigation, these early results indicate that LLMs are not reliably biased toward deep concepts.

[27] We focus on these data as they most closely speak to the issues raised by the earlier-discussed results on children's categorization. It is worth noting that the authors of this piece are centrally concerned with the role that teleology plays in the classificatory behaviour of LLMs, an issue that was not raised in the early work on psychological essentialism we have discussed. Comparison of these results with more up-to-date work on psychological essentialism would be an important project, but is beyond the scope of the present chapter.

[28] Though, as Neufeld (2022) cautions, one ought not assume that "insides" are identical to essence.

4.2 Adversarial Images

One of the most direct demonstrations of the differences between human and machine cognition comes in the form of adversarial images. An adversarial image is an image which has been manipulated so as to pose difficulties for AI image classification systems. For example, consider an image classifier which works by making predictions about the distribution of colour qualities (hue, saturation, value) of the pixels in an image. Such a classifier will classify according to the overall likelihood that a particular collection of pixels would be found in an image of a particular sort. Each pixel or region will be assigned a conditional probability of occurring in such an image, and the classification which makes the average likelihood greatest will be assigned the highest probability. Adversarial images can then be generated by making minor tweaks to each pixel, such that each pixel is now judged slightly less likely by the classification in question. E.g., if the hypothesis is that the stimulus is an image of a panda, then it might predict that a particular pixel or region will be black. By slightly adjusting the saturation of this region, say by making it slightly more grey, the hypothesis is made very slightly less likely. By doing this over the entire image, the overall probability of the hypothesis is made significantly lower. Consider the following example wherein Goodfellow et al. (2014) demonstrated that an image that was classified with the label 'panda' with 58% confidence could be disrupted via the introduction of noise and pixel perturbations to render a label of 'gibbon' with 99% confidence:

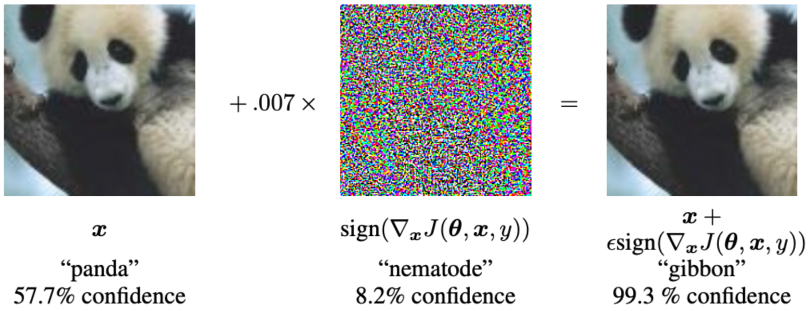
$\boldsymbol{x}$
"panda"
57.7% confidence
$+ .007 \times$
$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$=$
$\boldsymbol{x} + \epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 5: Figure borrowed from Goodfellow et al. 2014, 3.

The crucial point for our purposes is that such minor shifts in colour quality can be barely noticeable by the human visual system, even taken in the aggregate. In this way, imperceptible or barely perceptible changes (to a human) to an image can radically change the behaviour of an image classifier in response to the stimulus, while leaving human responses unchanged. Adversarial images are, in this way, particularly suggestive of quite deep differences between human classification and machine classification.

Note that we are not claiming that image classifiers entirely lack more abstract classificatory schemes which they apply to the images they are presented with. Nor are we saying that the mere fact that such systems 'misclassify' test images is itself problematic. Of course, human perceivers are subject to a whole host of more-or-less systematic kinds of illusion and performance failure.[29] The crucial point is that the kinds of errors that such systems make appear quite unlike the kinds of errors that humans make (e.g., susceptibility to adversarial images v.s. susceptibility to the Müller-Lyer illusion). These different response profiles are suggestive of different ways of interpreting and classifying stimuli, and are themselves susceptible to different styles of explanation.[30] In the human case, visual illusions are characteristically explained by appeal to the kinds of cues humans use to detect non-obvious properties of the external visual scene, such as depth or surface reflectance, and the statistics of such cues in the natural scenes in which human vision evolves and develops. In the machine case, the explanation appeals rather to the relations between previously encountered stimuli, considered as 2D images, and the categories such stimuli have been labeled with. Here again we see how differences in the goals, structures, and biases inherent to the systems under investigation lead to differences in the classificatory schemes they appeal to in responding to their environment.

5 Divergent Competence

What we hope to have illustrated with the above cases are the ways in which human and machine representations display deep dissimilarities, despite surface similarity. AI systems function precisely to capture patterns in the stimuli on which they are trained. By capturing these patterns, they increase their capacity to make accurate predictions concerning future encounters with similar stimuli. In the case of LLMs, this means identifying all the relevant information about the distributions of expressions in the texts on which they are trained, so that the generation of novel text will conform to familiar patterns, and will thus have a greater chance of satisfying norms of grammar, relevance, truth, etc. that human users will expect. To the extent that such systems can be said to attach a meaning to a particular expression, this will consist in, and not be dissociable from, the set of probabilistic relations the system has identified between this expression and all the other information it can identify, including other individual expressions, as well as higher level patterns such as sentential context, semantic themes, and so on.

In the human case, however, there is compelling evidence that the mental representations used in reasoning, language use, and other cognitive tasks, are not so reducible to statistical dependencies. While of course humans are adept at projecting patterns of experience from the past to the future, and very likely make use of such capacities in a wide range of cognitive contexts, including the use of language, it seems clear that this is

[29] Likewise, recent empirical work suggests humans have a knack for predicting machine classifications in light of some adversarial attacks (Zhou and Firestone 2019) again suggesting systematicity relevant to theoretical explanation.

[30] Indeed, the dominant methodology in cognitive science assumes that patterns in mistakes are as integral to a theory of innate competencies as are instances of getting things right.

not *all* they do. Human concepts appear to display a strong dissociation between their referential capacities and the worldly information they associate with their referents. This basic, philosophical, observation about human concepts is reinforced by the wide range of empirical work demonstrating the ways that human cognition and cognitive development deviate from, and extend beyond, the mere accumulation of information about correlated features of the environment. If this line of reasoning is on the right track, it seems that AI and human systems differ substantially in the innate contributions that bias their predictive models. On the assumption that such inductive biases reflect the systems' structure, goals, and developmental differences and, in tandem, establish meaning, we conclude that AI and human systems will not attach the same meanings to these expressions.

Deep Learning systems are incredibly powerful learners. It is not unreasonable to think that, given sufficiently large datasets on which to be trained, they will be able to replicate to a high degree of accuracy, the superficial patterns observed in human language use, image classification, and much more. However, this ability to replicate the behavioural patterns of human speakers and classifiers is not sufficient grounds for saying that they will thereby replicate the inner representational systems that human beings make use of in their cognitive activities. That is, while these systems may reproduce human *performance*, they may do so without replicating human cognitive *competence*. To determine whether they have managed the latter, we will need to make use of more sophisticated methodologies, such as those of the cognitive sciences, to investigate the systems underlying these behavioural capacities. The results in so far suggest that dissimilarity between humans and machines is more likely than similarity in these underlying systems.

In itself, this should not be a surprising result. It is a commonplace in the cognitive sciences that there will typically be a variety of ways of achieving a particular behavioural end. Because the evolutionary and ontogenetic pressures on human cognitive development are not the same as those found in the production of Deep Learning systems, we should expect to find differences in the structures and operations of such systems, including at the level of representations and representational content. Indeed, it is often remarked that human cognition is, like many biological traits, far from what an intelligent designer would construct given the same task specification. From this perspective, it would be quite surprising if human and machine representational systems ended up in the same place.

Bringing us back to our original question, and the topic of this collection, we can draw a sceptical conclusion concerning the near-term prospects of human-machine communication. The assumption we adopt is a simple, but we think intuitive, one: communication is successful only if the communicator associates a specific representational content with a message, and the communicatee identifies this representational content, and associates it with the communicative behaviour of the communicator, on the basis of this very behaviour. In order to successfully communicate, the speaker and the hearer must be talking *about the same thing(s)*. This is widely taken for granted within the philosophical literature, and can be easily motivated by everday cases in which miscommunication stems from speaker and hearer assigning different extensions to the same word-form. Consider, for example, the different ways that the term ‘noodles’ is used in British and American English. As always, the conceptual analysis is tricky here, but roughly: In American English, ‘noodle’ is applied broadly to any unleavened dough products which are flattened out and cut before being cooked in boiling water; whereas, in British English, ‘noodle’ has a more restricted application, applying primarily to such products as found in East Asian cooking. So, while ramen is in the extension of ‘noodles’ on both sides of the Atlantic, a British speaker would find the application of the term to spaghetti unusual at best, and to lasagne sheets utterly impossible. These different concepts, expressed by the same word-form, provide prime material for miscommunication. The examples write themselves. If Brittany, a Brit, tells Amy, an American, that she “loves all noodle dishes”, she is liable to be disappointed by the well-intentioned presentation of a tuna casserole. Examples of this sort seem well accounted for by the proposal that successful communication requires shared concepts: what has gone wrong in such a case is precisely that Brittany and Amy do not share their concepts (Brittany’s concept NOODLE is more restrictive than the concept Amy would express with the same word), and so communication fails.

If we, and most theorists in the philosophy of language, are right that genuine, successful communication involves the sharing of representational content between communicator and communicatee, these results suggest that human-machine communication is unlikely to be found along the pathway of current AI research. The nature of representation in humans and machines is sufficiently different that it is unlikely that this condition will be met. Specifically, we argue that the referents of human representations are secured in ways not available to machine representations. While human representation is connected to the world by innately-shaped biases about perceptually non-obvious features of the referential targets of our concepts/expressions, machine representation is essentially tied to superficial properties. Deep learning systems are, in a particular sense, descriptivist systems; whereas humans are not natural descriptivists. This difference undermines attempts to align the content of these two kinds of representational system.[31]

The danger that this fact poses is enhanced by the uncanny ability to mimic human performance we have remarked upon above. As LLMs and similar technologies get better and better, their productions will look more and more like those of human language users, both with respect to sentence-internal features like well-formedness and fluency, and discourse properties like relevance, conversational style, and so on. In the limit, their behaviour could be indistinguishable from that of human speakers (or, at least, human writers). Even with today’s systems, such as GPT4, it is very difficult to avoid the temptation to interpret

[31] These representations will not merely differ in something akin to sense, but reference as well. Humans will apply the term ‘tiger’ to a range of (actual or merely possible) entities which can be arbitrarily unlike our stereotype associated with tigers, and will refuse to apply this term to things which are arbitrarily similar to this stereotype, given the right story. But a machine representation which associates nothing more, or less, than this set of stereotypical assumptions with this expression will not be able to do so. So, when an LLM, or some more complex future system incorporating LLMs with other, non-linguistic sources of information, produces a string of text containing the term ‘tiger’, this will not have the same meaning, it would not have the same *truth-conditions*, for the machine as it would for a human interpreter.

such productions *as if they had been produced by a human.* If our arguments above are on the right track, this is a mistake. This uncanny performance will mask deep underlying differences in the systems responsible for it, and the meanings attached to it. Such systems are thus liable to generate the *illusion* of communication, wherein we anthropomorphize the generated text and interpret in ways that do not coincide with the interpretation assigned to it by the system which produced it, rather than the genuine "meeting of minds" assumed in standard analyses of successful communication.[32]

6 Uncanny Performance

For many practical purposes, this may not matter. Just as in a number of everyday contexts it may not matter that two speakers are using terms with precisely the same meaning or extension.[33] If I just want the automated machine I am interacting with to book my train ticket, I don't really care whether the text it produces means what I would mean by it, or indeed whether it means anything at all. But in other cases it may matter significantly.

Consider a more practical issue involving the use of machine learning programs in criminal justice. Predictive policing algorithms like PredPol (now 'Geolitica') are designed to predict where and when crimes are likely to occur based on statistical data. The statistical data here consists of quantifiable variables, such as past police records of crime types, crime locations, and crime dates and times. On the basis of these data, the algorithms identify a geographic area as a potential hotspot for crime:

[32] These anthropomorphic tendencies are reminiscent of Quine's own solution to indeterminacy, through his invocation of empathy, and are what inspired the epigraph at the start of the paper.

[33] Returning to our example of Amy and Brittany, they may have a practically successful conversation using the term 'noodles' in contexts in which only, say, ramen noodles are under discussion, due to the relevant overlap between their concepts.

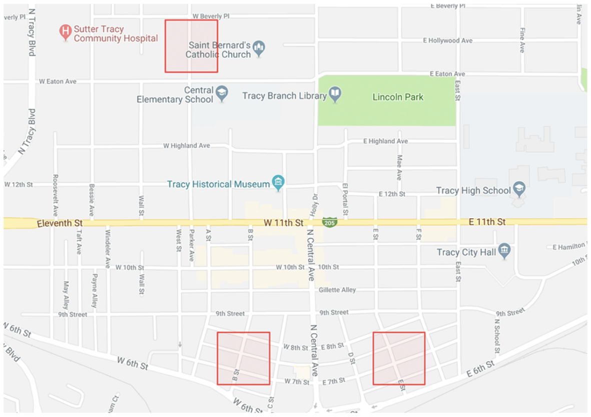

Figure 6: Figure borrowed from PredPol 2023, showing predictions "displayed as red boxes on a web interface via Google Maps [wherein] ... boxes represent the highest-risk areas".

Ask any computer scientist what this label means and they'll likely respond that it means the likelihood of a crime occurring within a red box's area is greater than the likelihood of a crime occurring outside of that area.[34]

But, bringing the indeterminacy worry to the fore, why should we favor this particular interpretation? Why should we assume that the algorithm is actually assessing the probability of individuals engaging in criminal activities, rather than, for instance, gauging the likelihood of someone having the police called on them (a metric highly sensitive to the race of the suspect), or the probability of someone encountering law enforcement without access to quality legal representation, or even the mere presence of a police officer in a specific geographical area?[35] In these cases, the concept of 'crime' is being distorted through its operationalization on the basis of superficial variables, obscuring its connection to a richer concept utilized by humans.

Likewise for recidivism-risk algorithms like COMPAS. Programs like these are designed to predict the likelihood of a person reoffending based on statistical data. The statistical data here consists of quantifiable variables, such as past arrest, location data, and other demographic details like age, and the label of 'high-risk' is taken by computer scientists, judges, and laypeople alike to mean that the likelihood of a person reoffending is sufficiently higher than those labeled as 'low-risk'. Just as before, however, we can regard the high-risk label as identifying instead the likelihood of a person of colour being targeted by law enforcement, or of a person being unable to afford legal representation in court, or of a police officer being present in a specific geographic location.

In both of these cases, like with the gaydar example from the start of the paper, the

[34] Importantly, these systems have been found to be largely inaccurate (Sankin and Mattu 2023).

[35] Rendering it unsurprising that such algorithms disproportionally target low-income and predominantly Black and Latino neighborhoods (Sankin et al. 2021).

consequences of miscommunication between what the machine means by this label and what a human understands it to mean are grave. While the system might superficially appear effective in its aim (dispatching more police to a particular area is very likely to result in more arrests, denying bail to a defendant identified by the program as high-risk is likely to reduce their being targeted by law enforcement patrol in the immediate future, and labeling a photo of a man as gay based on his exhibiting gay male stereotypes in that photo is likely to aid matchmaking algorithms), in the worst case, this could lead to unnecessary arrests or surveillance, a perpetuation of a cycle of criminalization that affects marginalized communities disproportionately, and the further persecution of historically targeted groups. Miscommunication or misunderstanding of these divergent meanings can thus have serious societal and ethical consequences, including perpetuating bias, reinforcing stereotypes, and contributing to systemic inequality.

7 Conclusion

We have argued that human concepts and the representational states generated by Deep Learning systems in contemporary Artificial Intelligence are fundamentally different to one another. Human concepts are naturally biased towards 'deep' representation: representation which abstracts away from, and is not reducible to, perceptually available features of the stimulus. Machine representations are, on the other hand, what we call 'superficial concepts'. These representations consist of (potentially quite complex) statistical summaries of perceptually encountered features of the stimulus. While sufficiently fine-grained and statistically grounded superficial concepts may, for a wide range of tasks, successfully mimic the behaviour of deep concepts, these are fundamentally different ways of representing the world. And these differences will matter for a number of reasons when we consider human-machine interactions. We have focused in this chapter on communication. It is widely assumed that successful communication constitutively involves a sharing of content between the speaker and the hearer. If we are right about the disparities between the representational capacities of human and machine systems, this will undermine communicative interactions between humans and machines. Technological progress along the lines currently being developed will not lead towards successful human-machine communication, but instead towards more and more convincing *illusions* of communication. To the extent that we want to genuinely communicate with machines, this might suggest a foundational revision to the kinds of systems which we ought to be developing. Looking towards the wealth of work in the cognitive sciences aimed precisely at uncovering how human cognition and conceptualization work can provide a picture of what such systems will have to be like. Considered against the backdrop of the high social, political, and ethical costs to miscommunication in critical domains, pursuing this new direction in AI development becomes not just a technical benchmark, but a moral imperative.

8 Bibliography

Agüera y Arcas, B., Todorov, A., and Mitchell, M. (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? medium.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Borer, H. (1991). The causative-inchoative alternation: A case study in parallel morphology. *Linguistic Inquiry*.

Bourget, D. and Chalmers, D. J. (2014). What do philosophers believe? *Philosophical studies*, 170:465–500.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4(1):73–122.

Burge, T. (2003). Perceptual entitlement. *Philosophy and phenomenological research*, 67(3):503–548.

Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics*, 33(1):1–78.

Burge, T. (2007). Postscript to 'individualism and the mental'. *Foundations of mind*, pages 151–181.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press.

Burge, T. (2013). Some remarks on putnam's contributions to semantics. *Theoria*, 79(3):229–241.

Butlin, P. (2022). Machine Learning, Functions and Goals. *Croatian journal of philosophy*, 22(66):351–370.

Butlin, P. (2023). Sharing our concepts with machines. *Erkenntnis*, 88(7):3079–3095.

Cappelen, H. and Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. Oxford University Press.

Chalmers, D. J. (2023). Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Chater, N., Clark, A., Goldsmith, J. A., and Perfors, A. (2015). *Empiricism and language learnability*. OUP Oxford.

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, 35(1):26–58.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Crain, S., Giblin, I., and Thornton, R. (2021). The deep forces that shape language and the poverty of the stimulus. *A Companion to Chomsky*, pages 462–475.

Donnellan, K. S. (1966). Reference and definite descriptions. *The philosophical review*, 75(3):281–304.

Dretske, F. (1986). Misrepresentation. In Bogdan, R., editor, *Belief: Form, Content, and Function*, pages 17–36. Oxford University Press.

Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Harvard University Press.

Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. (2023). Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721.

Fillmore, C. (1970). The grammar of hitting and breaking. In Jacobs, R. A. and Rosenbaum, P., editors, *Readings in English transformational grammar*. Ginn and Co.: Waltham Mass.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

Gagliardi, A., Feldman, N. H., and Lidz, J. (2017). Modeling statistical insensitivity: Sources of suboptimal behavior. *Cognitive science*, 41(1):188–217.

Gelman, S. A. and Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23(3):183–209.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., and Trueswell, J. C. (2005). Hard words. *Language learning and development*, 1(1):23–64.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hale, K. L. and Keyser, K. L. H. S. J. (2002). *Prolegomenon to a theory of argument structure*, volume 39. MIT press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Johnson, G. (2023a). Proxies Aren't Intentional; They're Intentional. *Unpublished Manuscript*.

Johnson, G. (2023b). Varieties of Bias. *Unpublished Manuscript*.

Johnson, G. M. (2020). The structure of bias. *Mind*, 129(516):1193–1236.

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.

Lasnik, H. and Lidz, J. (2016). The argument from the poverty of the stimulus. In Roberts, I., editor, *The Oxford Handbook of Universal Grammar*, chapter 10, pages 221–248. Oxford University Press.

Lidz, J. (2020). Learning, memory, and syntactic bootstrapping: A meditation. *Topics in Cognitive Science*, 12(1):78–90.

Lidz, J., Gleitman, H., and Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, 87(3):151–178.

Locke, J. (1836/1996). *An essay concerning human understanding*. Penguin.

Mandelkern, M. and Linzen, T. (2023). Do language models refer? *arXiv preprint arXiv:2308.05576*.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT press.

Mitchell, T. M. (1980). The need for biases in learning generalizations.

Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy compass*, 17(5):e12823.

Papafragou, A., Cassidy, K., and Gleitman, L. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition*, 105(1):125–165.

PredPol (2023). The Three Pillars of Data-Driven Policing.

Putnam, H. (1975a). Is semantics possible? In *Mind, Language, and Reality: Philosophical Papers, Vol. 2*, pages 139–152. Cambridge: Cambridge University Press.

Putnam, H. (1975b). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7:131–193.

Quine, W. V. (1960). *Word and Object.* MIT Press.

Rock, I. (1957). The role of repetition in associative learning. *The American journal of psychology*, 70(2):186–193.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, 4(3):328–350.

Sakas, W. G. and Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, 19(2):83–143.

Salmon, N. U. (1979). How not to derive essentialism from the theory of reference. *The Journal of Philosophy*, 76(12):703–725.

Sankin, A. and Mattu, S. (2023). Predictive Policing Software Terrible At Predicting Crimes. *The Markup.*

Sankin, A., Mehrota, D., Mattu, S., and Gilbertson, A. (2021). Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them. *The Markup.*

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Teney, D., Abbasnejad, E., Lucey, S., and Van den Hengel, A. (2022a). Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16761–16772.

Teney, D., Peyrard, M., and Abbasnejad, E. (2022b). Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In *European Conference on Computer Vision*, pages 458–476. Springer.

Trueswell, J. C., Medina, T. N., Hafri, A., and Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156.

Tubino-Blanco, M. (2020). Causative/inchoative in morphology. In *Oxford Research Encyclopedia of Linguistics.*

Vasilyeva, N., Gopnik, A., and Lombrozo, T. (2018). The development of structural thinking about social categories. *Developmental Psychology*, 54(9):1735–1744.

Vasilyeva, N. and Lombrozo, T. (2020). Structural thinking about social categories: Evidence from formal explanations, generics, and generalization. *Cognition*, 204:104383.

Wang, Y. and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246.

Yang, C. (2017). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, 24(2):100–125.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zhang, S., She, J. S., Gerstenberg, T., and Rose, D. (2023). You are what you're for: Essentialist categorization in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Zhou, Z. and Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, 10(1):1334.