THE LINGUISTIC DEAD ZONE OF VALUE-ALIGNED AGENCY, NATURAL AND ARTIFICIAL

TRAVIS LACROIX

Abstract. The value alignment problem for artificial intelligence (AI) asks how we can ensure that the "values"—i.e., objective functions—of artificial systems are aligned with the values of humanity. In this paper, I argue that linguistic communication is a necessary condition for robust value alignment. I discuss the consequences that the truth of this claim would have for research programmes that attempt to ensure value alignment for AI systems—or, more loftily, those programmes that seek to design robustly beneficial or ethical artificial *agents*.

Keywords — artificial intelligence; AI; the value alignment problem; principal-agent problems; machine learning; objective functions; normative theory; language; linguistic communication; communication systems; information transfer; coordination; values; preferences; objectives; incentives

1. Introduction

The value alignment problem for artificial intelligence (AI) asks how we can ensure that the "values"—i.e., objective functions—of artificial systems are aligned with the values of humanity, writ large (Future of Life Institute, 2018; Russell, 2019). One component of this problem is *technical*, focusing on how to properly encode values or principles in artificial agents so that they reliably do what they ought to do—i.e., what *we* want them to do or what we *intend* for them to do. Another component of value alignment is *normative*, emphasising what values or principles from normative theory are the "correct" ones to encode in AI systems (Gabriel, 2020). However, ensuring value alignment for AI systems—or, more loftily, designing robustly beneficial or "ethical" artificial *agents*—requires more than just translating our best normative theories into a programming language.

In this paper, I propose and defend the following claim, which I will refer to as "the **main claim**" throughout.

Department of Philosophy, Durham University

E-mail address: travis.lacroix@durham.ac.uk.

Date: Draft of December, 2024. This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s11098-024-02257-w. Please cite published version.

Main Claim.
Linguistic communication is a necessary condition for robust value alignment.

After providing some background on artificial intelligence and the value alignment problem (Section 2), I begin by surveying some empirical evidence for the role of communication in aligning values in the context of interactions between human actors (Section 3). I then forward two arguments in favour of the **main claim** as it applies to interactions between humans and AI systems. First (Section 4), I highlight the fundamental role that asymmetric information plays in generating instances of the value alignment problem. Second (Section 5), I discuss the failures of the symbolic systems approach to AI by analogy with the rigidity of objective functions for aligning values in present-day AI systems. Each of these arguments provides some reason to believe the **main claim** is true in the form stated. Minimally, they make plausible a weaker version of the **main claim**: linguistic communication is *highly valuable* for aligning values between actors. Section 6 briefly considers whether the recent proliferation of large language models provides any optimism for the prospects of mitigating misalignment. Section 7 concludes by summarising the difficulty of solving the value alignment problem for artificial intelligence if the **main claim** is true.

2. Artificial Intelligence and the Value Alignment Problem

"Artificial intelligence" (AI) refers to a property of an artificial system——i.e., that it "thinks", "acts", or "behaves" in an intelligent way. In this context, intelligence is sometimes understood as "an agent's ability to achieve goals in a wide range of environments" (Legg and Hutter, 2007, 12).[1] "AI" also refers to an approach or set of techniques for *achieving* this property in an artificial system (Gabriel, 2020). The most promising method for achieving machine intelligence in recent years is machine learning (ML). This approach to AI involves training models using (typically huge amounts of) data. The models "learn" gradually to behave in the desired way without that behaviour being explicitly programmed.

Several paradigms fall under machine learning, including supervised, unsupervised, and semi-supervised learning, as well as reinforcement learning. Since 2012, owing to breakthroughs in image recognition (Krizhevsky et al., 2017), the main driver of AI research has been *deep learning* (Goodfellow et al., 2016; Prince, 2023). This approach to ML utilises deep neural networks modelled (roughly) after neurons in the human brain (Savage, 2019). Deep learning utilises layers of algorithms to process data—information is passed through each subsequent layer in a neural

[1]See also Gardner (2011); Cave (2017); Gabriel (2020); Russell and Norvig (2021) and the taxonomy of AI definitions provided by van Rooij et al. (2023).

network, with the previous layer's output providing input for the following layer. One of the key advantages of deep learning techniques is that they do not require the heavily hand-crafted features used by traditional methods for AI (Buckner, 2019).

Deep learning cuts across each machine learning paradigm mentioned above (Prince, 2023). Despite the variation in methods for training or fitting models, each of these approaches seeks to solve an *optimisation problem*. The thing being optimised, in this case, is an *objective function*—e.g., minimising a loss function in a supervised learning model or maximising a reward function in a reinforcement learning model. In effect, the objective function offers a way to configure the system to bring it closer to some *ground truth* provided during model training. Essentially, an objective function provides a *proxy* for what we want the system to do.

Consider an example from supervised learning for image recognition. Given a dataset, $\mathbf{D}$, of image-label pairs, (x, y), the objective is for the model to correctly "guess" (output) the image label for previously unseen images. In this case, the objective *function* might be a probability distribution function,

$$p(\hat{y} = y \mid x, \theta),$$

which outputs the conditional probability that the predicted label, $\hat{y}$, is identical to the true label, y, given the observed data, x, and a model, θ.

The true label, y, determines whether a particular predicted label, $\hat{y}$, is correct for a particular image, x. A *loss function* is used to optimise the model, θ. According to a specified evaluation metric—e.g., mean-squared error[2]—the loss function tells us how close the model is to the correct prediction for a single data point.[3] Thus, the *true objective*—correctly labelling images—is approximated by an *objective function*, which determines a metric for how close the model is to the objective.[4] If loss is close to zero, then the model successfully optimises the objective function according to the metric used. Note, however, that this success is relative to both the metric chosen and the aptness of the objective function as a proxy for the true objective. Notwithstanding, if the objective function is a good proxy for the true

[2]The mean squared error is given by $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where n is the number of data points, Y_i is the set of observed values, and $\hat{Y}_i$ is the set of predicted values.

[3]The loss function is sometimes differentiated from a *cost function*, which tells us the *average* loss over the entire dataset. This is sometimes couched in the language of *empirical risk*, which describes the loss over all samples in the dataset, which is then contrasted with the *true risk*—i.e., the loss over all samples. More accurately, an objective function approximates our true objective, and the objective function determines the empirical risk, which approximates the true risk. See discussion in Goodfellow et al. (2016, Sec. 8.1).

[4]In fact, this *approximates* an approximation. (With thanks to Michael Noukhovitch for bringing this to my attention). Loss is measured on the test set, which gives a measure of generalisation error for a given objective function. Hence, optimisation actually occurs on the training set. Still, the hope is that we are also optimising the test set—i.e., achieving some degree of generalisation. See discussion in Goodfellow et al. (2016, Sec. 5.2).

objective, then optimising an objective *function* means optimising the objective. So, for a system's resultant behaviour to align with the intended behaviour, the objective function(s) must be accurately specified (Reed and Marks, II, 1999). This technical feature of machine learning has been a key focus of research on value alignment to date.

In the context of AI, the value alignment problem can arise when objective functions (or, indeed, objectives) are misspecified. Complications emerge because objectives require programmers to define an objective function, but certain objectives may be difficult (or impossible) to operationalise in a programming language. Further complications arise from the fact that objective *functions* are mere proxies for the *true* objective; however, from the "perspective" of a model, the objective function just is the objective. When objective functions are poorly specified, this will lead to an instance of the value alignment problem insofar as there is a misalignment between the actual objective—our values—and its proxy—the system's "values" (LaCroix, 2025).

In addition, the provision of an objective function implicitly defines an optimisation "landscape" (Sipper et al., 2018) which, in complex action spaces, includes a "pathology" of *local* optima (Lehman and Stanley, 2008), meaning that these landscapes are frequently "deceptive" (Goldberg, 1987; Mitchell et al., 1992). In light of this, Lehman and Stanley (2008) highlight that the objective function "does not necessarily reward the stepping stones in the search space that ultimately lead to the objective" (329), meaning that objective functions are often constructed *ad hoc*. Thus, even when objective functions are accurate proxies, sufficiently complex action spaces will have local optima that may result in outputs that are grossly misaligned with the true objective. A mismatch between a well-specified objective function and the resultant behaviour of an AI system is sometimes referred to as "inner alignment"; when the objective is misspecified, this is referred to as "outer alignment" (Hubinger et al., 2021).[5]

These difficulties fall under the heading of "AI safety". Amodei et al. (2020) highlight that when objective functions are misspecified, this may give rise to unanticipated side effects or reward hacking; when objective functions are too expensive to evaluate at regular intervals, this creates a problem of scalable supervision; and, local optima of objective functions may lead to undesirable behaviour during learning.[6] These problems are exacerbated because the utilities determined by any concretely specified objective function will necessarily be a mere *subset* of our

[5]See also discussion in Amodei and Clark (2016); Ecoffet et al. (2020); Christian (2020); Krakovna et al. (2021).

[6]Amodei et al. (2020) refer to these problems as "accidents", but it should be clear how they can be classed as instances of the value alignment problem. See further discussion in Hadfield-Menell and Hadfield (2019); Raji and Dobbe (2023).

utilities—i.e., the things we value. Hence, value alignment is inherently difficult (LaCroix and Prince, 2023).

Each problem described above involves encoding the "right" objectives in an AI system. Gabriel (2020) refers to this as the "technical component" of the value alignment problem. Gabriel (2020) further distinguishes the technical component of the value alignment problem from the "normative component", which involves the problem of determining *what* values (objectives) should be encoded in an AI system in the first place. Part of the idea is that as these systems become more integrated into society, some of their decisions may carry moral weight, so we might classify their actions as "moral" or "immoral". These considerations have given rise to the field of *machine ethics*, which seeks to "implement moral decision-making faculties in computers and robots" (Allen et al., 2006).[7]

Even though discussions of the value alignment problem are often couched in the context of a hypothetical (i.e., fictional) future artificial general intelligence or superintelligence, it should be clear that instances of the value alignment problem are truly ubiquitous. "Accidents", like those described in Amodei et al. (2020), are (technical, internal) instances of value misalignment to the extent that misspecified or costly objective functions may lead to behaviour misaligned with what we intended the system to do. More generally, machine bias and problems arising from fairness considerations are instances of the value alignment problem—at least to the extent that we do not want or intend for models to act in ways that we would call "racist", "sexist", or otherwise discriminatory.[8]

Despite some of the conceptual clarification offered by distinguishing, e.g., inner versus outer alignment or technical versus normative components of value alignment, it is worth noting that contemporary discussions of value alignment for artificial intelligence are inherently vague insofar as the standard statement of the value alignment problem—i.e., the problem of ensuring AI systems align with the values of humanity—raises more questions than it answers (LaCroix, 2025).

More to the point of this paper, researchers who discuss value alignment in the context of artificial intelligence have not maintained adequate sensitivity to the role that communication plays in aligning values. Linguistic communication is indispensable in aligning the complex values of agents in human-human interactions. This fact lends some empirical credence to the **main claim**, as the subsequent section demonstrates.

---

[7]See Tolmeijer et al. (2020); Cervantes et al. (2020) for recent surveys of machine ethics and approaches to "artificial moral agency" (AMA). As well as criticisms offered by van Wynsberghe and Robbins (2019); LaCroix (2022).

[8]Some salient examples are described in, e.g., Angwin et al. (2016); Christian (2020); Tomasev et al. (2021); Miceli et al. (2022). See further discussion in LaCroix (2025).

3. The Curious Case of Human Value Alignment

As mentioned, some researchers who work on value alignment are concerned with our ability to control some hypothetical—i.e., fictional—future AI system. One such concern is that if such a system displays human-level intelligence (or superintelligence), it will be impossible to control.[9] Notwithstanding, for the entire history of the species, humans have engaged in the production of agents with human-level intelligence, whose values may be misaligned with their own, and over which they have limited control—i.e., human children. This provides something of a *proof of concept* that values can be aligned in such agents (Christian, 2020).

Value alignment in the context of artificial intelligence is typically presented as a problem of determining the "correct" or appropriate incentive structures that are required to induce the desired behaviour in an AI system. However, there is a sense in which aligning values is a type of coordination problem. So, at a structural level, we can understand and analyse value alignment in terms of coordination and cooperation—i.e., social dynamics.

Furthermore, part of why linguistic communication systems evolved in *Homo sapiens* is likely because of the cooperative demands that evolved in our lineage. In the context of hominin evolution, cooperation would have included demanding forms of collective action, and "with greater cooperation comes greater communication" (Planer and Sterelny, 2021, 73). It is difficult to imagine how such robust cooperation could have evolved without corresponding increases in the complexity and flexibility of our communicative abilities.

It is well understood that certain classes of coordination problems benefit from cheap talk (Farrell and Rabin, 1996)—i.e., simple communication. Therefore, simple communication channels will allow for the alignment of simple values; however, robust (linguistic) communication is necessary to admit robust value alignment in complex environments between agents with complex incentives. Hence, it should be relatively uncontroversial that a robust communication system—like natural language—is necessary for the demands of complex cooperation between agents with complex incentive structures.[10] One can understand the intent of other humans precisely because one can communicate linguistically.[11]

[9]The logic here is that "less intelligent" species cannot control "more intelligent" ones; see, e.g., Russell's (2019) discussion of the "Gorilla Problem".

[10]For further discussion of the distinction between simple and linguistic communication in the context of language origins, see LaCroix (2020, 2021).

[11]Strong empirical evidence exists for a tight, bidirectional connection between theory of mind capacities and linguistic capacities in human infants (Astington and Jenkins, 1999; Astington and Baird, 2005; de Villiers, 2007; de Villiers and de Villiers, 2014). First, language appears necessary to learn to express concepts surrounding one's own feelings and inner world (Nelson, 2005; Dunn and Brophy, 2005; Hutto, 2012); second, the information that language conveys about others' thoughts, feelings, beliefs, desires, etc., is much richer than the information conveyed through behaviour, eye gaze, gestural expressions, etc. (Appleton and Reddy, 1996; Harris, 2005; Peterson

Part of the benefit of linguistic communication over simpler communication systems is that the former are *compositional*. The compositional nature of language allows humans to communicate their goals with one another to an arbitrary degree of specificity. Linguistic communicative abilities also affect the cultural accumulation of new concepts. High-fidelity cultural learning allows human populations to solve coordination/cooperation problems by allowing selective learning and the accumulation of small improvements over time (Boyd et al., 2011). That is, language allows for accumulating knowledge across generations via social learning.[12] Planer and Sterelny (2021) highlight that "some forms of cooperation are stable only if reputation (knowledge of the past social actions of others) is tracked reliably and is part of common knowledge" (25). Again, this robust sort of cooperation depends significantly upon linguistic ability.

To some extent, communication depends upon joint attention and common knowledge (Tomasello, 2008, 2014). Complex language is not necessary for joint commitment if there is common-ground understanding (Tomasello, 2014); however, diminishing common ground between agents appears to necessitate greater lexical and structural richness in the language used to communicate information about disparate knowledge between agents.[13] Essentially, when the social aspects of agent interactions become increasingly dispersed in time and space, members of a social group will need more sophisticated communicative (and cognitive) abilities to report behaviour and events that happened "elsewhere and elsewhen" (Planer and Sterelny, 2021, 195).

From the cooperative nature of our species, it seems apparent that humans are at least capable of aligning their values in the various contexts we face daily. Empirical evidence suggests that part of the reason we can do so is that we can communicate via natural language. Humans use linguistic competence to impart subtle norms, goals, and values in subsequent generations that align with a long cultural history of norms, goals, and values. Because human linguistic capacity comes hardwired, we can take this particular aspect of how we align our values for granted. Thus, the "dead zone"[14] of value-aligned agency arises from a lack of sensitivity to the

---

and Siegal, 1999; Wellman and Peterson, 2013); third, linguistic abilities allow individuals to reason abstractly about others' actions via their beliefs (Astington and Jenkins, 1999; de Villiers and de Villiers, 2009; Milligan et al., 2007).

[12]See, e.g., Boyd (2016); Henrich (2016); Planer and Sterelny (2021).

[13]This has been observed empirically in, e.g., children's sign languages; see Meir et al. (2010).

[14]In David Cronenberg's film adaptation of Stephen King's novel, the lead character, Johnny Smith, wakes up from a coma with psychic powers. He uses the phrase "dead zone" to describe the missing part of his psychic vision—a blank part of the future that he cannot see. Hence, in the context of the film's plot, the "dead zone" is unseen, but it also denotes a future outcome that is not yet determined, meaning that it can be changed. See Cronenberg (1983).

importance of language for *Homo sapiens* as fundamentally cooperative and social creatures.[15]

Hence, empirical work on the co-evolution of language and cooperation in *Homo sapiens* is instructive when considering cooperation—i.e., the alignment of values—in the context of artificial intelligence. In the case of a hypothetical (i.e., fictional) superintelligence, such an entity would presumably have very little common ground with human agents, meaning that an ability to communicate linguistically is necessary to ensure cooperation (assuming that the cooperative demands on the agent are sufficiently complex). What is pressing in the context of value alignment for AI is that assumptions of linguistic competency cannot be taken for granted when considering artificial agents.

4. Language, Value Alignment, and Information Transfer

The empirical evidence for the role of communication in aligning values in human-human interactions, discussed in Section 3, should provide some plausibility to the claim that linguistic communication is required for robust value alignment (at least in the context of human-human interactions). In this section, I examine how or whether these insights apply to the AI context by underscoring the fundamental role that information asymmetries play in generating value misalignment in the first place. Hence, I suggest that the first argument in favour of the **main claim** follows from the joint realisation that

(1) the value alignment problem (in the context of AI) is a type of principal-agent problem;
(2) the principal-agent problem is fundamentally a problem of information transfer rather than misaligned values per se; and
(3) linguistic communication is a uniquely robust and flexible communication system which allows for information transfer to an arbitrary degree of specificity.

The value alignment problem, in its most general form, is a problem of how two (or more) agents (actors) can align their values (objectives). Hadfield-Menell and Hadfield (2019, 417) suggest that the value alignment problem for artificial intelligence has a "clear analogue" in principal-agent problems from economics, law, and

[15]Of course, one might object that although humans use linguistic communication systems—i.e., they have linguistic communicative abilities—they often fail to align their values. This is fine. The **main claim** furnished a *necessary* condition for value alignment; it does not say anything about whether language is *sufficient* for value alignment. It should be obvious that it is not. There is a substantive question about whether any set of abilities will suffice for value alignment; however, my claim is that no matter what set of abilities is uncovered, value alignment will be impossible without linguistic communication. The **main claim** also does not imply linguistic communication is *the only* necessary condition for value alignment.

political theory.[16] A principal-agent problem arises (or, at least, may arise) in any context where some entity—called "the principal"—appoints another entity—called "the agent"—to act on the principal's behalf. Delegating tasks from a principal to an agent can give rise to a problem instance when the principal and the agent have competing objectives, incentives, values, or interests. When there is a conflict between the principal's and the agent's values, we say their values are misaligned. Agency dilemmas of this form are ubiquitous in human-human interactions insofar as human agents rarely have identical incentives.[17] Much work in economics and law has been devoted to mitigating this problem by aligning the agent's values with the principal's via contracts.

Laffont and Martimort (2002) highlight that if there is no private information between a principal and an agent, then *even if* the agent's objectives conflict with the principal's, the principal could still propose "a contract which perfectly controls the agent and induces the [agent's] actions to be what [the principal] would like to do himself in a world without delegation" (12). Essentially, under complete information, the principal has complete *bargaining power*.[18] Therefore, misaligned values alone are insufficient for generating a principal-agent problem since they can be controlled when there is no informational asymmetry between the principal and the agent.[19] Private information can create an agency dilemma in at least three different ways.

First, *hidden knowledge* is an informational asymmetry resulting from the agent's private information—e.g., concerning their own skills or opportunity costs—which the principal cannot access. Hidden knowledge may contribute to the generation of a principal-agent problem.[20] For example, a prospective employee (the agent) knows their background skill level and appropriateness for a particular job, whereas the hiring committee (the principal) does not. In cases of hidden knowledge, uncertainty is exogenous to the relationship between the principal and the agent.

[16]This problem is sometimes referred to as an *agency dilemma* or an *incentive problem*; see discussion in Jensen and Meckling (1976); Eisenhardt (1989); Laffont and Martimort (2002).
[17]Although it is worth noting that coordination problems can tolerate a reasonable degree of conflict regarding the incentives of the agents; see discussion in Noukhovitch et al. (2021).
[18]Classical game theory often operationalises this as a higher disagreement point for the powerful agent; see discussion in LaCroix and O'Connor (2021).
[19]This claim is mathematically provable on the particular economic model we are discussing. Of course, since this is a model and, therefore, an idealisation, we might question whether this model is sufficiently applicable to real-world interactions and whether this claim holds in the real world.
[20]In economics, this is referred to as *adverse selection*. Hidden knowledge on the agent's part causes the principal to give up some information rent. Thus, a contract must be designed to elicit private information, which may be costly to the principal. See Akerlof (1970); Rothschild and Stiglitz (1976); Spence (1973, 1974); Laffont and Martimort (2002); Hou et al. (2009).

Second, *hidden action* is an informational asymmetry caused by the agent's ability to perform an action that the principal cannot observe.[21] When the risk-taking individual (the agent) knows more about their intentions than the consequence-paying individual (the principal), the agent may take on more risk than the principal would otherwise be comfortable with, as is common with insurance. In this case, the uncertainty due to asymmetric information is endogenous to the relationship between the principal and the agent.

Solutions to principal-agent problems arising from informational asymmetries due to hidden action and hidden knowledge assume that information is (ex post) verifiable by an independent third party, such as a (benevolent) Court of Justice (Laffont, 2000). Hence, a third type of informational asymmetry may give rise to an agency dilemma when we assume that the information between an agent and principal is symmetric (ex post) but unverifiable *in principle* by a third party (Sappington, 1991). Thus, the third type of informational asymmetry that may generate a principal-agent problem arises from the non-verifiability of (otherwise symmetric) information.[22] For example, suppose the principal and the agent have identical and sufficient (ex ante) information to complete some transaction. Here, a principal-agent problem may still arise when the agent represents the true state of the world as being different than they (and the principal) know it to be. This is a problem when either the agent's representation of the world is unverifiable by a third party or when it is too costly for a third party to verify.

Thus, misaligned incentives (value misalignment) may give rise to a principal-agent problem. However, when information is symmetric, the principal can create a contract that induces the agent to act just as the principal would without delegation. Therefore, value misalignment *alone* is not sufficient to generate a principal-agent problem.

Furthermore, value misalignment is also *unnecessary* in generating a principal-agent problem. For example, suppose the agent and principal have perfectly aligned objectives but cannot transmit that information. In that case, it is still possible that the agent's actions misalign with the principal's objectives (despite the agent's intentions). This situation might occur if, for example, there is an optimal action which would satisfy the principle's objectives—and, *ex hypothesi*, would also satisfy

[21]In economics, this is known as *moral hazard*. See Haynes (1895); Knight (1921); Arrow (1963, 1968); Vaughan (1997); Laffont and Martimort (2002).

[22]Non-verifiability is particularly relevant in the field of *contract law*. However, Shah (2014) notes that non-verifiability receives much less coverage in the economic literature on principal-agent problems than hidden knowledge (adverse selection) and hidden action (moral hazard). See Williamson (1973, 1975); Grossman and Hart (1986); Sappington (1991); Hart (1995); Laffont (2000); Laffont and Martimort (2002).

the agent's objectives—but the existence of this action is not common knowledge.[23] Simple coordination games, where the agent chooses an action, and the principal and agent both receive the same payoff, provide a clear example of this possibility.[24]

In its most general form, the value alignment problem arises from the dynamics of multi-agent interactions involving the delegation of tasks from one actor to another. Hence, the value alignment problem in the context of artificial intelligence can be understood as a subset of the principal-agent problem, where the principal is a human (or set of humans) and the agent is an AI system. LaCroix (2025) gives the following definition:

> **The Value Alignment Problem** (Structural Definition)
> A problem that arises from the dynamics of multi-agent interactions involving the delegation of tasks from one actor (a human principal) to another (an artificial agent). This problem can arise whenever
> (*a*) The agent's objective function is misaligned with the true objective of the principal(s); *or*,
> (*b*) There are informational asymmetries between the principal and the agent.

The second condition of LaCroix's (2025) structural definition highlights that, in addition to competing incentives, another key feature of delegation that can give rise to an agency dilemma is *informational asymmetry* and *imperfect information*.[25]

If the value alignment problem in the context of artificial intelligence is a type of principal-agent problem, and principal-agent problems are problems of informational asymmetries rather than misaligned values per se, then it follows that the value alignment problem in the context of artificial intelligence is (primarily) a problem of informational asymmetries. Hence, information-transferring capacities are necessary for solving or mitigating instances of the value alignment problem.

---

[23] Suppose the philosophical literature on peer disagreement is to be trusted. In that case, it is at least possible for this to be true even when both parties are privy to identical evidence.

[24] As a toy case, suppose the principal chooses a number, and the agent has to guess the correct number for both players to receive a payoff. Their values are perfectly aligned, but the agent may still act in a way that the principal would not, precisely because of an informational asymmetry between them. (With thanks to Aydin Mohseni for raising this possibility to me.) This toy example is much less artificial than it may seem at first glance—simple signalling games have a similar structure.

[25] "Perfect information" is a term of art in this context. In Economics, perfect information implies that all market participants have all the information required to make a decision. In game theory, perfect information means that a player knows the game's entire history up to the decision point, as in backgammon. Imperfect information is the negation of perfect information; this occurs when some information is unavailable or hidden. Thus, imperfect or incomplete information means that there is some uncertainty. See discussion in von Neumann and Morgenstern (1944); Shapley (1953). In the 1970s, economists showed how asymmetric information poses significant challenges to *General Equilibrium Theory* (Akerlof, 1970; Spence, 1974; Rothschild and Stiglitz, 1976). See also Marschak and Radner (1972) and discussion in Laffont and Martimort (2002).

In addition, this problem scales in complexity. Namely, the more complex or robust the informational asymmetry is, the more complex or robust the information-transferring capacity will need to be to resolve that information asymmetry. (As we saw in Section 3, complex cooperation settings impose robust demands on communicative abilities.)

It is also uncontroversial that linguistic communication—understood in this context as relatively synonymous with natural language—is a uniquely robust system for transferring information. One key feature often taken to differentiate linguistic communication systems from their simpler precursors is *compositionality* (and related features like *hierarchy* and *recursion*). The crucial point is that the elements of natural language can be combined into hierarchical phrases, which then may be recursively combined into larger phrasal expressions. Moreover, the meaning of such an expression depends (functionally) upon the meaning of its parts and how they are combined. In this sense, recursion requires hierarchy—at least to some extent—and hierarchy requires compositionality—again, at least to some extent (De Beule, 2008). These features of linguistic communication are often understood as a requirement for explaining the *systematicity* and *productivity* of natural language (Fodor, 1998; Szabó, 2020)—features which are almost certainly useful and very likely necessary for the cooperative abilities of linguistic agents, and so the possibility of value alignment.

We have now seen that the value alignment problem is a type of principal-agent problem which arises whenever one actor delegates authority to another to act on her behalf. Furthermore, I have argued that informational asymmetries are necessary and sufficient for generating such problems—misaligned values alone will not suffice, nor are they required. Hence, to solve complex instances of the value alignment problem, agents need to be able to communicate linguistically, which is the **main claim**.

It is worth reiterating at this point that the **main claim** furnishes a *necessary* condition; it should be obvious that linguistic communication alone is insufficient for aligning values. The point of drawing attention to this key feature of value alignment is not to deny that incentive structures are irrelevant for value alignment but that the focus on designing "provably safe" incentive structures will be insufficient for mitigating such problem instances as these systems become more complex. Hence, more attention must be paid to informational asymmetries, particularly in light of the ineliminable opacity of increasingly complex computational systems (Creel, 2020).

Building on empirical evidence of the relationship between linguistic communication and value alignment between human agents, this section has argued that information asymmetry is a key component of the principal-agent problem—and,

therefore, the value alignment problem for artificial intelligence. These considerations together imply the **main claim**—that linguistic communication is a necessary condition for robust value alignment. Section 5 offers a distinct argument in favour of the **main claim** based on the brittleness of objectives—i.e., the targets of value alignment—under the machine learning paradigm.

5. Objective Functions and Value Proxies

Another way to think about why language might plausibly be required for a robust form of value alignment is to consider how the *symbolic systems* approach to AI—sometimes called "good old-fashioned AI" (GOFAI)—is thought to have failed—or, perhaps more charitably, to be much more limited than initially believed.[26] Part of this is that these systems are far too rigid: every rule for action must be hard-coded. This is adequate for simple tasks, but writing explicit instructions for every contingency becomes intractable as complexity increases in light of combinatorial explosion.

Part of why "second-wave" AI—ML and particularly deep learning methods bolstered by big data—has been surprisingly successful in comparison is that many rules for action are not hard-coded but implicit. What underwrites this success is that it is difficult to express the intuitive knowledge required for robust generalisation in the form of a set of verbally expressible rules that can be codified in a machine language. The symbolic systems approach of first-wave AI relied on the ability of humans to express explicit knowledge (often in the form of complicated *if-then* rules). In contrast, LaCroix and Bengio (2019) highlight that deep neural networks can "capture" the kind of implicit knowledge that is difficult to express in a formal language.[27]

The failures of symbolic systems are supposed to have been caused by the system's being "brittle, unconducive to learning, defeated by uncertainty, and unable to cope with the world's rough and tumble" (Cantwell Smith, 2019). This has led to the view that, although hard-coding expert knowledge into AI systems helps in the short term, it plateaus and inhibits progress in the long term; historically, "breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning" (Sutton, 2019). This view has led to the *scaling hypothesis*—i.e., the guiding faith of contemporary machine learning research that performance scales with size, data, and compute (Brown et al., 2020; Kaplan et al., 2020).

[26]GOFAI is also called "first-wave AI" or "symbolic AI". The phrase used here was coined by Haugeland (1985). For discussions of the failures of GOFAI, see also McDermott (1987); Cantwell Smith (2019).

[27]See also discussion in Buckner (2019).

However, the key thing to note is that contemporary approaches to AI still require hard-coding objective functions. Thus, although these systems are significantly more flexible for *learning* and *acting*, the "values" encoded in these systems are still brittle. By analogy, it appears that a similar move from rigidity to flexibility concerning these systems' objective functions will be necessary to ensure that the "values" of these systems are aligned with our values. Again, some way of transferring information is necessary for dynamic values.

The *inverse reinforcement learning* (IRL) paradigm is one possible exception to this claim. Whereas a classic reinforcement learning model is given a reward function and attempts to learn behaviour based on the rewards it receives, an IRL model is given behaviour and attempts to learn the reward function that would give rise to that behaviour. Therefore, the objective function is learned instead of hard-coded. So, the thought goes, it may be possible for a sophisticated IRL model to learn values from behaviours insofar as actions transfer information. However, this line of reasoning is problematic in several directions.

On the one hand, IRL is underspecified as a research problem because many reward functions can explain an actor's behaviour, and the costs of solving the problem tend to grow disproportionately with the size of the problem (Arora and Doshi, 2021). Essentially, many different reward functions could explain any observed behaviour.[28] Hence, the information transferred by behavioural cues is too ambiguous to ensure value alignment. Again, this problem arises primarily from information asymmetries and the need for a robust communication channel for aligning values.

On the other hand, behaviours will only ever serve as a proxy for values, which is part of why poorly specified objective functions give rise to value misalignment in the first place, as discussed in Section 2. Furthermore, by emphasising the information transferred by behaviours alone, we run into well-known problems to which revealed preference theory from economics gives rise.[29] A basic (and false) assumption of IRL is that the behaviour a model observes *is* optimal, given the (hidden) reward function motivating that behaviour. In addition, even if behaviours were a good proxy for preferences and provided a high-fidelity channel for transferring information about those preferences, linguistic communication would aid learning objectives at a much higher rate—consider the difference between attempting to discern someone's preferences by watching them act versus asking them what their

[28]This is effectively identical to the problem of rule-following, discussed in Wittgenstein (1953); Kripke (1982) and a significant body of secondary literature since.

[29]Revealed preference theory (Samuelson, 1938a,b) uses consumer behaviour to analyse the choices made by individuals. See discussion in Sen (1973, 1977, 1993, 1997, 2002); Koszegi and Rabin (2007); Hausman (2012).

preferences are. Hence, in addition to systematicity and generalisability, linguistic communication offers the possibility of speed in terms of communicating and, therefore, identifying values in the first place. In high-stakes cases, waiting until sufficient behavioural instances are observed will be impossible. At the same time, pre-training is inadequate because values shift over time. Linguistic communication is necessary for flexible objective specification, which is necessary for robust value alignment.

6. Language and Language Models

Assuming the **main claim** is true, it is worth noting that recent advances in generative AI—particularly the proliferation of "large language models" (LLMs) like BERT (Devlin et al., 2019), the GPT Suite (Brown et al., 2020), RETRO (Borgeaud et al., 2022), LaMDA (Thoppilan et al., 2022), Llama (Touvron et al., 2023), Claude (Anthropic, 2024), etc.—do not warrant optimism for the tractability of mitigating the value alignment problem.

Although LLMs can be impressive syntax engines, language is more than mere syntax—hence why some researchers have suggested that the phrase "large language model" is a misnomer better described in terms of "stochastic parrots" (Bender et al., 2021), "large corpus models" (Veres, 2022), or "bullshit generators" (Narayanan and Kapoor, 2024). Some authors argue that meaning cannot be learned from form alone; furthermore, because of the very structure of language modelling tasks—predicting plausible words in a sequence—LLMs are designed to learn form rather than meaning (Bender and Koller, 2020; Bender et al., 2021; Bisk et al., 2020; Marcus and Davis, 2020).[30] Hence, there is a real sense in which the output of a large language model (like ChatGPT) is literally meaningless.[31] Moreover, the meaninglessness of LLM outputs is independent of theoretical views on meta-semantics (Mallory, 2023).

Although such models can produce coherent-sounding and contextually relevant responses based on patterns learned, they lack any understanding of the data consumed or generated (Saba, 2023). Of course, the relationship between understanding and linguistic competence is complicated (and beyond the scope of this paper). The key thing to be clear about is that linguistic communication is a particularly robust system of communication that relies on shared meanings: language is a *social* endeavour that requires the participation of both senders and receivers. The outputs

[30]See further discussion in Mitchell and Krakauer (2023).
[31]See further discussion in, e.g., McCoy et al. (2019); Ettinger (2020); Pandia et al. (2021); Sinha et al. (2021); Sahlgren and Carlsson (2021).

of LLMs *look* impressive because they are interpreted by competent language users (humans) who project meaning, intentions, understanding, etc., onto the system.[32]

One of the key advances that has made the advent of widely- and publicly-available LLMs possible is the successful application of reinforcement learning from human feedback (RLHF) and reinforcement learning from computational feedback (RLCF). Indeed, RLHF/RLCF is sometimes touted as the most prominent and successful method for aligning AI systems to human values. However, this success has little bearing on value alignment more generally: Narayanan et al. (2023) highlight that model alignment via RLHF can only protect against *accidental* harms. The focus on techno-solutionism amounts to little more than technochauvinism (Broussard, 2018, 2023). Moreover, it is important not to confuse *usefulness* with *alignment*. Although it is true that RLHF has made the LLM business financially viable, this does not imply that these systems are aligned with the values of humanity.

These criticism apply to other current (technical) approaches to alignment, including supervised fine tuning and prompt crafting. The key insight of the **main claim** is that such approaches will be inadequate for sufficiently complex systems. Approaches that focus purely on the technical components of value alignment commit a category mistake insofar as the value alignment problem is neither technical nor even normative, but fundamentally social (LaCroix, 2025). Moreover, we have seen how linguistic capacity is a prerequisite of value alignment in social interactions.

Even though there are reasons to think that present-day LLMs lack the type of linguistic abilities the **main claim** posits are required for robust value alignment, nothing in the preceding arguments suggests anything over and above mere *functional* understanding of linguistic communication abilities is required for solving certain forms of value misalignment arising from informational asymmetries. Although current models fail to exhibit meaningful linguistic capacities or any form of understanding, even if they did exhibit such a capacity, this would not provide a counterargument to the **main claim**—namely, if a future version of a language model did exhibit demonstrable understanding and moved us toward more aligned AI systems, then this would provide additional evidence in favour of the main claim.

7. Conclusion

The value alignment problem is ubiquitous in the context of AI systems. As these systems become more sophisticated and increasingly deployed in society, the

[32]This is simply another iteration of the ELIZA effect, first observed in the 1960s Weizenbaum (1976). One key difference is that Weizenbaum was disturbed by lay responses to his chatbot, ELIZA. In contrast, the for-profit OpenAI has a financial interest in advertising its syntax engine to users and the media even when its stated capacities are demonstrably false.

problems arising from value misalignment become more pressing. However, empirical evidence suggests that complex agents with divergent incentives are capable of aligning their values. In Section 3, I highlighted that insights from evolutionary biology suggest that linguistic communication is at least highly valuable, if not necessary, for the cooperative features of human-human interactions (i.e., value alignment between human agents). Part of the reason is because value alignment is primarily a problem of information asymmetries rather than competing incentives.

This empirical work highlights a key distinction between the difficulty of aligning values in human-human interactions versus human-AI interactions. In the former type of interaction, linguistic competence is "in-built" to some extent; in the latter, no assumptions can be made regarding inherent linguistic ability. Hence, it is essential to maintain sensitivity to the informational asymmetries that underlie value alignment for AI systems. Value alignment between human principals and human agents allows key features of the problem to be taken for granted—particularly the ability to communicate linguistically. This fact has apparently been forgotten or ignored by some researchers focusing on the technical components of value alignment for AI.

In light of the fundamental role of informational asymmetries in generating instances of the value alignment problem, I have argued that the key constitutive features of linguistic communication—simplified here in terms of compositionality—give rise to the systematicity and generalisability that is a prerequisite for mitigating such problem instances (Section 4). The second argument for the **main claim** (Section 5) highlights that objective functions—the "values" encoded in present-day AI systems—are rigid in precisely the ways that contributed to the failures of the symbolic systems approach to AI. Hence, although current systems are flexible regarding *learning*, they are still rigid regarding *objectives*—i.e., the target of value alignment.

The **main claim** helps clarify how difficult value alignment can be in this context. Moreover, if the **main claim** is true, this would specify a fairly demanding lower bound on the difficulty of mitigating value alignment for sufficiently robust AI systems. At the same time, if language is an AI-complete problem, then this suggests that value alignment is impossible for suitably complex AI systems.[33]

Thinking about the importance of linguistic communication for value alignment in artificial systems pushes work in this area beyond the straightforward application of specific insights from normative theory to AI by additionally providing novel insights in the opposite direction. Normative theories historically focus on *human*

[33]This means that creating linguistic AI is essentially equivalent in difficulty or complexity to creating generally human-level AI systems. However, it is worth noting that "AI completeness" is an *informal* concept defined by *analogy* with complexity theory. See discussion in Shahaf and Amir (2007).

agents. As such, they can take for granted that agents can communicate linguistically. As mentioned, no such assumption can be made when the agent is artificial. It is precisely because we cannot make assumptions about, e.g., in-built linguistic ability that the implementation of ethically-aligned artificial systems comes to bear on normative theory. Importantly, such insights will be theory-neutral insofar as they do not depend upon any particular metaethical framework. As a result, thinking about problems of value alignment in terms of coordination—and linguistic aids to successful coordination—provides valuable insights into the implicit foundations upon which many normative theories rest.

Hence, linguistic communication has been a "dead zone" in theorising about value alignment in the context of artificial agents; however, since the ability to communicate linguistically is taken for granted in normative theories that focus on human (natural) agents, little attention has been paid to the necessity of language for value-aligned agency in natural agents as well. As mentioned at the outset, I take the **main claim** to hold in the context of value alignment more generally than just situations involving artificial agents. Hence, the linguistic dead zone of value-aligned agency applies in normative contexts for both natural and artificial agents.

Acknowledgements. I want to thank the editors of this special issue on Normative Theory and AI and the organisers of the Normative Theory and AI Workshop (2022)—Seth Lazar, Pamela Robinson, Claire Benn, and Todd Kharu—as well as participants—David Grant, Jeff Behrends, John Basl, Chad Lee-Stronach, Jack Parker, and David Danks—for helpful feedback. I would also like to thank, in no particular order, Aaron Courville, Aaron Wright, Aydin Mohseni, Benjamin Wald, Celso Neto, Dominic Martin, Duncan MacIntosh, Gillian Hadfield, Jeffrey Barrett, Jennifer Nagel, Michael Noukhovitch, Richmond Campbell, Sasha Luccioni, Simon Huttegger, and Yoshua Bengio for discussion and feedback throughout the process. Thanks also to an anonymous referee for constructive feedback. An early draft of this paper was presented at the Philosophy Department Colloquium at Dalhousie University (Fall 2021). I am grateful to the audience members for their helpful discussion. Thanks also to the SRI at the University of Toronto and Mila - Québec Artificial Intelligence institute for partially funding this research and for providing generous resources.

References

Akerlof, George A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3): 488–500.

Allen, Colin, Wendell Wallach, and Iva Smit (2006). Why machine ethics? *IEEE Computer Society*, 21(4): 12–17.

Amodei, Dario and Jack Clark (2016). Faulty reward functions in the wild. https://openai.com/blog/faulty-reward-functions/.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané (2020). Concrete problems in ai safety. *arXiv*, 1606.06565: 1–29. https://arxiv.org/abs/1606.06565.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). Machine bias. *ProPublica*, May 23: np. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anthropic (2024). The claude 3 model family: Opus, sonnet, haiku. *Papers with Code*. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Appleton, Michele and Vasudevi Reddy (1996). Teaching three-year-olds to pass false belief tests: A conversational approach. *Social Development*, 5(3): 275–291.

Arora, Saurabh and Prashant Doshi (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297: 1–28.

Arrow, Kenneth J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53: 941–973.

Arrow, Kenneth J. (1968). The economics of moral hazard: Further comment. *American Economic Review*, 58: 537–539.

Astington, Janet Wilde and Jodie A. Baird (2005). Introduction: Why language matters. In Astington, J. W. and J. A. Baird, editors, *Why language matters for theory of mind*, pages 2–25. Oxford University Press, Oxford.

Astington, Janet Wilde and Jennifer M. Jenkins (1999). A longitudinal study of the relation between language and theory of mind development. *Developmental Psychology*, 35(5): 1311–1320.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (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, Emily M. and Alexander Koller (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian (2020). Experience grounds language. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre (2022). Improving language models by retrieving from trillions of tokens. *arXiv*, 2112.04426: 1–43. https://arxiv.org/abs/2112.04426.

Boyd, Robert (2016). *A Different Kind of Animal: How Culture Made Humans Exceptionally Adaptable and Cooperative*. Princeton University Press, Princeton.

Boyd, Robert, Peter J. Richerson, and Joseph Henrich (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Suppl. 2): 10918–10925.

Broussard, Meredith (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. The MIT Press, Cambridge, MA.

Broussard, Meredith (2023). *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. The MIT Press, Cambridge, MA.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). Language models are few-shot learners. *arXiv*, 2005.14165. https://arxiv.org/abs/2005.14165.

Buckner, Cameron (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14: e12625.

Cantwell Smith, Brian (2019). *The Promise of Artificial Intelligence: Reckoning & Judgment*. The MIT Press, Cambridge, MA.

Cave, Stephen (2017). Intelligence: A history. *Aeon*. https://aeon.co/essays/on-the-dark-history-of-intelligence-as-domination.

Cervantes, José-Antonio, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, 26: 501–532.

Christian, Brian (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, New York.

Creel, Kathleen A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4): 568–589.

Cronenberg, David (1983). *The Dead Zone*. Dino De Laurentiis Company. 103 min.

De Beule, Joachim (2008). *Compositionality, Hierarchy and Recursion in Language: A Case Study in Fluid Construction Grammar*. PhD thesis, Vrije Universiteit Brussel.

de Villiers, Jill G. (2007). The interface of language and theory of mind. *Lingua*, 117(11): 1858–1878.

de Villiers, Jill G. and Peter A. de Villiers (2009). Complements enable representation of the contents of false beliefs: Evolution of a theory of theory of mind. In Foster-Cohen, S., editor, *Language acquisition*, pages 169–195. Palgrave Macmillan, Hampshire.

de Villiers, Jill G. and Peter A. de Villiers (2014). The role of language in theory of mind development. *Topics in Language Disorders*, 34(4): 313–328.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 1810.04805: 1–16. https://arxiv.org/abs/1810.04805.

Dunn, Judy and Marcia Brophy (2005). Communication, relationships and individual differences in children's understanding of mind. In Astington, J. W. and J. A. Baird, editors, *Why language matters for theory of mind*, pages 50–69. Oxford University Press, Oxford.

Ecoffet, Adrien, Jeff Clune, and Joel Lehman (2020). Open questions in creating safe open-ended ai: Tensions between control and creativity. *arXiv*, 2006.07495: 1–9. https://arxiv.org/abs/2006.07495.

Eisenhardt, Kathleen M. (1989). Agency theory: An assessment and review. *The Academy of Management Review*, 14(1): 57–74.

Ettinger, Allyson (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *arXiv*, 1907.13528: 1–20. https://arxiv.org/abs/1907.13528.

Farrell, Joseph and Matthew Rabin (1996). Cheap talk. *The Journal of Economic Perspectives*, 10(3): 103–118.

Fodor, Jerry A. (1998). There are no recognitional concepts—not even red, part 2: the plot thickens. In *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*, pages 49–62. The MIT Press, Cambridge, MA.

Future of Life Institute (2018). Asilomar AI principles. https://futureoflife.org/ai-principles/.

Gabriel, Iason (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30: 411–437.

Gardner, Howard (2011). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, New York.

Goldberg, David E. (1987). Simple genetic algorithms and the minimal deceptive problem. In Davis, Lawrence D., editor, *Genetic Algorithms and Simulated Annealing (Research Notes in Artificial Intelligence)*, pages 74–88. Morgan Kaufmann Publishers, Burlington, MA.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. The MIT Press, Cambridge, MA. http://www.deeplearningbook.org.

Grossman, Sanford J. and Oliver D. Hart (1986). The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy*, 94: 691–719.

Hadfield-Menell, Dylan and Gillian K. Hadfield (2019). Incomplete contracting and ai alignment. In Conitzer, Vincent, Gillian Hadfield, and Shannon Vallor, editors, *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422. Association for Computing Machinery, New York.

Harris, Paul L. (2005). Conversation, pretense and theory of mind. In Astington, J. W. and J. A. Baird, editors, *Why language matters for theory of mind*, pages 70–83. Oxford University Press, Oxford.

Hart, Oliver (1995). *Firms, Contracts, and Financial Structure*. Oxford University Press, Oxford.

Haugeland, John (1985). *Artificial Intelligence: The Very Idea*. The MIT Press, Cambridge.

Hausman, Daniel M. (2012). *Preference, Value, Choice, and Welfare*. Cambridge University Press, Cambridge.

Haynes, John (1895). Risk as an Economic Factor. *Quarterly Journal of Economics*, 9(4): 409–444.

Henrich, Joseph (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species and Making Us Smarter*. Princeton University Press, Princeton.

Hou, Jianwei, Ann Kuzma, and John Kuzma (2009). Winner's Curse or Adverse Selection in Online Auctions: The Role of Quality Uncertainty and Information Disclosure. *Journal of Electronic Commerce Research*, 10(3): 144–154.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant (2021). Risks from learned optimization in advanced machine learning systems. *arXiv*, 1906.01820: 1–39. https://arxiv.org/abs/1906.01820.

Hutto, Daniel D. (2012). *Folk psychological narratives*. The MIT Press, Cambridge, MA.

Jensen, Michael C. and William H. Meckling (1976). Theory of the firm: Managerial behaviour, agency costs and ownership structure. *Journal of Financial Economics*, 3(4): 305–360.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). Scaling laws for neural language models. *arXiv*, 2001.08361: 1–30. https://arxiv.org/abs/2001.08361.

Knight, Frank H. (1921). *Risk, Uncertainty and Profit*. University of Chicago Press, Chicago.

Koszegi, Botond and Matthew Rabin (2007). Mistakes in Choice-Based Welfare Analysis. *American Economic Review*, 97(2): 477–481.

Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg (2021). Specification gaming: the flip side of ai ingenuity. https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

Kripke, Saul A. (1982). *Wittgenstein on Rules and Private Language*. Harvard University Press, Cambridge, MA.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

LaCroix, Travis (2020). *Complex Signals: Reflexivity, Hierarchical Structure, and Modular Composition*. PhD thesis, University of California, Irvine.

LaCroix, Travis (2021). Reflexivity, Functional Reference, and Modularity: Alternative Targets for Language Origins. *Philosophy of Science*, 88(5): 1234–1245.

LaCroix, Travis (2022). Moral Dilemmas for Moral Machines. *AI and Ethics*, 2: 737–746. https://doi.org/10.1007/s43681-022-00134-y.

LaCroix, Travis (2025). *Artificial Intelligence and the Value Alignment Problem: A Philosophical Introduction*. Broadview Press, Peterborough, ON. Forthcoming.

LaCroix, Travis and Yoshua Bengio (2019). Learning from learning machines: Optimisation, rules, and social norms. *arXiv preprint*, 2001.00006. https://arxiv.org/abs/2001.00006.

LaCroix, Travis and Cailin O'Connor (2021). Power by Association. *Ergo: an Open Access Journal of Philosophy*, 8: 163–189. https://doi.org/10.3998/ergo.2230.

LaCroix, Travis and Simon J. D. Prince (2023). Deep Learning and Ethics. *arXiv*, 2305.15239: 1–25. https://arxiv.org/abs/2305.15239.

Laffont, Jean-Jacques (2000). *Incentives and Political Economy*. Oxford University Press, Oxford.

Laffont, Jean-Jacques and David Martimort (2002). *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, Princeton.

Legg, Shane and Marcus Hutter (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4): 391–444.

Lehman, Joel and Kenneth O. Stanley (2008). Exploiting open-endedness to solve problems through the search for novelty. In *Proceedings of the Eleventh International Conference on Artificial Life (ALIFE XI)*, pages 329–336, Cambridge, MA. The MIT Press.

Mallory, Fintan (2023). Fictionalism about chatbots. *Ergo: an Open Access Journal of Philosophy*, 10: 1082–1100.

Marcus, Gary and Ernest Davis (2020). Gpt-3, bloviator: Openai's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

Marschak, Jacob and Roy Radner (1972). *Economic Theory of Teams*. Yale University Press, New Haven.

McCoy, Tom, Ellie Pavlick, and Tal Linzen (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Korhonen, Anna and Lluís Màrquez David Traum, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence. Association for Computational Linguistics.

McDermott, Drew (1987). A critique of pure reason. *Computational Intelligence*, 3: 151–160.

Meir, Irit, Wendy Sandler, Carol Padden, and Mark Aronoff (2010). Emerging sign languages. In Marschark, M., editor, *Oxford Handbook of Deaf Studies, Language, and Education*, pages 267–280. Oxford University Press, Oxford.

Miceli, Milagros, Julian Posada, and Tianling Yang (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP): 1–14.

Milligan, Karen, Janet Wilde Astington, and Lisa Ain Dack (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2): 622–646.

Mitchell, Melanie, Stephanie Forrest, and John H. Holland (1992). The royal road for genetic algorithms: Fitness landscapes and GA performance. In Varela, F. J. and P. Bourgine, editors, *Proceedings of the First European Conference on Artificial Life*, pages 1–11. The MIT Press, Cambridge, MA.

Mitchell, Melanie and David C. Krakauer (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, 120(13): e2215907120.

Narayanan, Arvind and Sayash Kapoor (2024). *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton University Press, Princeton, NJ.

Narayanan, Arvind, Sayash Kapoor, and Seth Lazar (2023). Model alignment protects against accidental harms, not intentional ones. *AI Snake Oil*, Dec(01). https://www.aisnakeoil.com/p/model-alignment-protects-against.

Nelson, Katherine (2005). Language pathways into the community of minds. In Astington, J. W. and J. A. Baird, editors, *Why language matters for theory of mind*, pages 26–49. Oxford University Press, Oxford.

Noukhovitch, Michael, Travis LaCroix, Angeliki Lazaridou, and Aaron Courville (2021). Emergent communication under competition. In *Proceedings of the*

20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21), page 974–982, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Pandia, Lalchand, Yan Cong, and Allyson Ettinger (2021). Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv*, 2109.12951: 1–13. https://arxiv.org/abs/2109.12951.

Peterson, Candida C. and Michael Siegal (1999). Representing inner worlds: Theory of mind in autistic, deaf and normal-hearing children. *Psychological Science*, 10(2): 126–129.

Planer, Ronald J. and Kim Sterelny (2021). *From Signal to Symbol: The Evolution of Language*. The MIT Press, Cambridge, MA.

Prince, Simon J. D. (2023). *Understanding Deep Learning*. The MIT Press, Cambridge, MA.

Raji, Inioluwa Deborah and Roel Dobbe (2023). Concrete problems in ai safety, revisited. *arXiv*, 2401.10899: 1–6. https://arxiv.org/abs/2401.10899.

Reed, Russell D. and Robert J. Marks, II (1999). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. The MIT Press, Cambridge, MA.

Rothschild, Michael and Joseph Stiglitz (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics*, 93(4): 541–562.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York.

Russell, Stuart and Peter Norvig (2021). *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken, NJ, 4 edition.

Saba, Walid S. (2023). Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. *arXiv*, 2309.05918: 1–17. https://arxiv.org/abs/2309.05918.

Sahlgren, Magnus and Fredrik Carlsson (2021). The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4: 682578.

Samuelson, Paul A. (1938a). A note on the pure theory of consumer's behaviour. *Economica*, 5(17): 61–71.

Samuelson, Paul A. (1938b). A note on the pure theory of consumer's behaviour: An addendum. *Economica*, 5(19): 353–354.

Sappington, David E. M. (1991). Incentives in principal-agent relationships. *Journal of Economic Perspectives*, 5(2): 45–66.

Savage, Neil (2019). How ai and neuroscience drive each other forwards. *Nature*, 571: S15–S17.

Sen, Amartya K. (1973). Behaviour and the concept of preference. *Economica*, 40(159): 241–259.

Sen, Amartya K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, 6(4): 317–344.

Sen, Amartya K. (1993). Internal consistency of choice. *Econometrica*, 61(3): 495–521.

Sen, Amartya K. (1997). Maximization and the act of choice. *Econometrica*, 65(4): 495–521.

Sen, Amartya K. (2002). *Rationality and Freedom*. Harvard University Press, Cambridge, MA.

Shah, Sunit N. (2014). Literature review: The principal agent problem in finance. *The CFA Institute Research Foundation*, pages 1–55.

Shahaf, Dafna and Eyal Amir (2007). Towards a Theory of AI Completeness. In Amir, Eyal, Vladimir Lifschitz, and Rob Miller, editors, *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 150–155, New York. AAAI Press.

Shapley, Lloyd S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10): 1095–1100.

Sinha, Koustuv, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv*, 2104.06644: 1–26. https://arxiv.org/abs/2104.06644.

Sipper, Moshe, Ryan J. Urbanowicz, and Jason H. Moore (2018). To know the objective is not (necessarily) to know the objective function. *BioData Mining*, 11(21): 1–3.

Spence, Michael (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3): 355–374.

Spence, Michael (1974). *Market Signalling: Informational Transfer in Hiring and Related Processes*. Harvard University Press, Cambridge, MA.

Sutton, Rich (2019). The bitter lesson. *Incomplete Ideas*, Mar(13). http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Szabó, Zoltán Gendler (2020). Compositionality. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University, Fall 2020 edition. https://plato.stanford.edu/archives/fall2020/entries/compositionality/.

Thoppilan, Romal, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le (2022). Lamda: Language models for dialog applications. *arXiv*, 2201.08239: 1–47. https://arxiv.org/abs/2201.08239.

Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein (2020). Implementations in machine ethics: A survey. *arXiv*, 2001.07573: 1–38. https://arxiv.org/abs/2001.07573.

Tomasello, Michael (2008). *Origins of Human Communication*. The MIT Press, Cambridge, MA.

Tomasello, Michael (2014). *A Natural History of Human Thinking*. Harvard University Press, Cambridge, MA.

Tomasev, Nenad, Kevin R. McKee, Jackie Kay, and Shakir Mohamed (2021). Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 254–265, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3461702.3462540.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2307.09288: 1–77. https://arxiv.org/abs/2307.09288.

van Rooij, Iris, Olivia Guest, Federico Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich (2023). Reclaiming ai as a theoretical tool for cognitive science. *PsyArXiv Preprints*, pages 1–21. https://doi.org/10.31234/osf.io/4cbuv.

van Wynsberghe, Aimee and Scott Robbins (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, 25: 719–735.

Vaughan, Emmett J. (1997). *Risk Management*. Wiley, New York.

Veres, Csaba (2022). Large language models are not models of natural language: they are corpus models. *arXiv*, 2112.07055: 1–12. https://arxiv.org/abs/2112.07055.

von Neumann, John and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

Weizenbaum, Josef (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman and Company.

Wellman, Henry M. and Candida C. Peterson (2013). Theory of mind, development, and deafness. In Baron-Cohen, S., H. Tager-Flusberg, and M. V. Lombardo, editors, *Understanding other minds: Perspectives from developmental social neuroscience*, pages 51–71. Oxford University Press, Oxford.

Williamson, Oliver E. (1973). Markets and hierarchies: Some elementary considerations. *The American Economic Review*, 63(2): 316–325.

Williamson, Oliver E. (1975). *Markets and Hierarchies, Analysis and Antitrust Implications: A Study of the Economics of Internal Organization*. The Free Press, New York.

Wittgenstein, Ludwig (2009/1953). *Philosophical Investigations*. Wiley-Blackwell, Oxford, 4 edition.