

Knowledge and Artificial Intelligence:

A Foundational Framework for the Epistemology of AI

Christoph Kelp and Mona Simion

c.kelp@bham.ac.uk

mona.simion@philosophy.ox.ac.uk

Abstract

This Element develops an epistemological framework for Artificial Intelligence. As AI systems become increasingly integrated into our knowledge generating practices - from scientific research to everyday inquiry - we urgently need to understand their epistemic status. This Element takes the first steps towards filling this theoretical gap. It sketches the first foundational framework for the epistemology of Artificial Intelligence, comprising: a metaepistemology of AI (including demarcation criteria for distinguishing genuine artificial epistemic agents from mere data processors); an individual epistemology of AI (what it takes for an artificial system to inquire, achieve justification, knowledge, and understanding, and to produce genuine constative speech); and a social epistemology of AI (principles for when to trust AI outputs). In doing so, 'Knowledge and Artificial Intelligence' delivers new tools to better understand our new epistemic collaborators and their agency as knowledge-seekers and sharers.

For Mia and Max

Table of Contents

<i>Preface</i>	5
<i>Introduction</i>	6
<i>Part §1: Foundations of Artificial Epistemic Agency</i>	9
1.1 Epistemic Agency	9
1.2 Epistemic Agents, Biological and Artificial	11
1.3 Demarcating Artificial Epistemic Agency	13
<i>Part §2: Four for Artificial Epistemic Agents: Justification, Inquiry, Assertion, and Knowledge</i>	16
2.1 The Whole Hog	16
2.2 Argument §1: The Argument from Attributions of Justification, Evidence, and Reasons	21
2.3 Argument §2: The Argument from Inquiry	23
2.4 Argument §3: The Argument from AI Systems as Informants	26
2.5 Argument §4: LLMs, Knowledge, and Understanding	30
<i>Part §3. Should We Trust AIs?</i>	36
3.1 The Default View	37
3.2 Transparency and the Explainability View	38
3.3 Computational Reliabilism	41
3.4 Back to the Default View	43
3.4 Rational Trust and Trustworthiness	44
3.5 Trustworthy AI: An Ought-Based Account	47
<i>References</i>	51

Preface

Many thanks to all our philosophy friends for all the support for this research. In particular, many thanks to Herman Cappelen for his inspiring work, many discussions on the topic of this research, and his extraordinary support over the years. Special thanks for detailed comments on this work go to Adam Carter and Annalisa Coliva, as well as an anonymous referee for CUP. Many thanks for very helpful discussions also go to Jessica Brown, Paul Faulkner, Lizzie Fricker, Sandy Goldberg, John Greco, Jennifer Lackey, Susanna Schellenberg, Rachel Sterken, Ernie Sosa, and Tim Williamson.

Thanks also for very helpful feedback to the audiences at the BSTK 2026 in Oxford, Social Conflicts and Epistemic Aims Conference, University of Seville, Epistemology of AI Conference, University of Ioannina, Norms in Language-Based Human-AI Interaction conference, Jagiellonian University, Krakow, Research Seminar at the Centre for Philosophy and AI Research, Friedrich-Alexander-Universität, Erlangen, the 6th Biannual Philosophy of AI Conference, Society for the Philosophy of AI, VU Amsterdam, the Upsalla/Vienna Research Seminar, and the Trust and AI Workshop, University of L'Aquila. Special thanks for excellent critical discussions of our views go to the participants in the 'Ethics of AI' course in the 2026 Exeter College Summer Programme in Oxford.

At some junctures, arguments in this Element draw on and further develop ideas in previously published work: the discussion of epistemic agency is partially informed by (Simion Forthcoming); the discussion of trust and trustworthiness develops some ideas from (Kelp and Simion 2023), (Simion and Kelp 2020b) and (Simion and Willard-Kyle 2026). Many thanks to the editors of and referees for these venues for all the support.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme ('KnowledgeLab: Knowledge-First Social Epistemology' project, grant agreement No 948356, and 'Know-How: A New Framework for Theoretical and Practical Control Across Intelligence Domains' project, grant agreement No 101199255). We thank the European taxpayers and the ERC reviewers for the exceptional level of support.

As always, the biggest thanks go to our kids, Mia and Max, for making it possible for us to write this Element. They patiently and kindly sat in countless conferences and writing retreats while we were working on this topic.

Introduction

For the first time in the history of human knowledge, non-biological entities appear capable of generating, processing, and sharing knowledge in a fashion that closely resembles human epistemic agency. But what exactly is happening here, epistemologically? What is the epistemic status of AI systems? Are they genuine epistemic agents, or are they mere data processors? Can we trust their outputs in forming beliefs and making decisions? The urgency of answering these questions cannot be overstated: Artificial Intelligence systems have quickly become central parts of our epistemic lives. As such, the question concerning the epistemic status of AI transcends academic philosophy. AI systems now co-author scientific papers, with some journals reporting over 10% of submissions showing AI involvement. They diagnose diseases with accuracy sometimes surpassing human physicians. They analyse legal evidence, educate millions, and have the power to shape public discourse. Yet we operate without principled frameworks for understanding their epistemic contributions. This theoretical vacuum has profound practical consequences for how we make decisions based on AI generated information, how we structure our institutions, allocate responsibility, and navigate our collective epistemic future. For instance, without understanding whether AIs can bear epistemic responsibility, we cannot properly allocate accountability in AI-assisted decision-making: after all, there is an epistemic condition on responsibility in action. Without knowing whether AIs can provide genuine testimony, we cannot determine appropriate evidential standards for AI-generated content in courtrooms, classrooms, and laboratories: should the evidential treatment be on a par with that of human testifiers, or that of mere instruments?

While the past several years have witnessed the development of a burgeoning literature on the ethics of AI¹, the epistemology of AI remains woefully underexplored. Furthermore, a full, foundational framework for AI epistemology is missing in the literature. However, epistemological issues lie at the heart of both AI's main benefits (e.g. generating and sharing knowledge) and AI's harms (dangerous decision making based on AI generated misinformation, algorithmic amplification, etc.). At the roots of these benefits and harms lie distinctively epistemological questions, such as whether and to what extent AI should be understood as a source of knowledge, and if so why.

Three critical gaps characterize the current landscape:

- (1) **The Metaepistemology of AI: The Demarcation Problem.** While we routinely attribute knowledge to AI systems, we lack principled criteria for distinguishing genuine artificial epistemic agency from mere data processing. Are AIs like thermometers - reliable instruments that 'know' the temperature only metaphorically? Or are they genuine epistemic agents? Answering these questions requires principled demarcation criteria for artificial epistemic agency.
- (2) **The Epistemology of Artificial Agents: AI Inquiry, Justification, Knowledge, Understanding, and Assertion.** The very possibility of artificial epistemic agency raises a crucial epistemic alignment challenge for AI: how can we ensure that AI systems operate in accordance with genuine epistemic values and norms? Are AI systems capable of knowledge, and are they subject to inquiry norms like taking up relevant evidence and following rational updating processes? Are they bearers of epistemic justification and responsibility? Answering these questions requires careful epistemological theorizing.
- (3) **The Social Epistemology of Human/AI Interaction.** Existing social epistemological frameworks of testimony and trust often assume human psychology and struggle to accommodate artificial agents. Can AI systems be genuine testifiers? Are we justified to believe their outputs? Are they trustworthy? Answering these questions demands new social epistemological theoretical foundations.

¹ For an overview see Dubber et al. 2020.

This Element will sketch a foundational epistemological framework for understanding artificial epistemic agency, addressing all three knowledge gaps. The approach will rely on a key value-theoretic insight about emergent epistemic normativity that bridges functional and epistemic domains: both instruments and cognitive systems bear functional normative properties, but only cognitive systems' functional norms constitute genuine epistemic norms. The hypothesis we will defend is that, at some point on the functional normative spectrum, we encounter emergent epistemic normative properties.

Our approach will be informed throughout by relevant results on AI in neighbouring subdisciplines: ethics, philosophy of language, philosophy of mind, philosophy of technology, and philosophy of science, and will incorporate some of the most robustly established results in these fields. For instance, for the most part, we will assume an externalist picture of content, which allows AI outputs to be contentful² (however, our results will not stand or fall with this assumption: alternatively, AI outputs can carry information about the world in the way tree rings carry information about the age of trees).

That being said, we will venture to remain, as far as possible, broadly independent from substantive commitments in these fields, in order to ensure high robustness of results. For instance, we will not take a stance on what terms like 'Claude' or 'ChatGPT' pick out: we'll leave that task to metaphysicians. We'll also avoid taking on commitments about artificial minds: should it turn out that philosophers of mind conclude that AI systems lack mentality, they will, *a fortiori*, not be capable of justified beliefs or knowledge either. However, their status as epistemic agents will not be threatened by this, in that e.g. questions as to whether AIs are capable of e.g. making epistemically justified claims, or generating justified beliefs in us, remain open. Conversely, if we establish, on independent epistemological grounds, that AIs can know (for instance, because they are the proper bearers of *ex ante* justification, and the correct analysis of justification is in terms of knowledge), this will constitute strong evidence for research in the philosophy of mind for the existence of artificial minds.

Game Plan:

Part §1: Foundations of Artificial Epistemic Agency

This Part sketches a metaepistemology of Artificial Intelligence: it develops an account of, and a demarcation criterion for artificial epistemic agency. We argue that, at some point on the functional normative spectrum, we encounter emergent epistemic normative properties. We employ this insight in the service of proposing a demarcation criterion for artificial epistemic agency: if epistemic normativity is emergent on the spectrum of functional normativity, we should expect that the capacity for evidence responsiveness provides a principled way to distinguish genuine artificial epistemic agents from mere data processors.

Part §2: Four for Artificial Epistemic Agents: Justification, Inquiry, Assertion, and Knowledge

Part §2 offers four arguments in support of AIs as full epistemic agents. In this, we sketch a response to the epistemic alignment problem for AI. We first argue that widely spread justification and reasons responsiveness attributions are *prima facie* evidence of AI epistemic agency. We then argue that LLMs are genuine inquirers, genuine informants, and that they can have knowledge and understanding. As an immediate consequence for the extensive research programme in truth-tracking AI (in industry and beyond): if we are right, we already have truth-tracking AI.

Part §3: Should We Trust AIs?

² See e.g. Putnam 1975. See also Cappelen and Dever 2021 for discussion of content externalism and AI, and Hattiangadi (Forthcoming) for pessimism about AI content, and Sterken and Kirkpatrick (2024) for difficulties for human-AI conversational alignment.

This Part sketches a comprehensive framework for the social epistemology of human-AI interaction, including an epistemology of AI testimony, developing a strong anti-reductionist position according to which comprehension of AI content provides default entitlement to believe, subject to defeat. This framework will explain when trust in AI is rational, how AI trustworthiness relates to its disposition to meet its functional norms, and why explainable AI is sometimes (but not always!) required for rational trust. Once more, this result comes with an immediate consequence for industry: the XAI research programme is unmotivated. What we should do is build trustworthy AI, not explainable AI.

Part §1: Foundations of Artificial Epistemic Agency

Introduction

It is well documented that we widely attribute knowledge to AIs such as Chat-GPT or Claude, just as we attribute knowledge to humans (Cheng et al. 2024). Do these knowledge attributions track reality? Are AIs genuine epistemic agents? A lot of research has attempted to make progress on this epistemological question by answering neighbouring questions in the philosophy of mind: do AIs have mentality (e.g. Goldstein and Levinstein Forthcoming)? Are they conscious (e.g. Chalmers 2023)? However, short of mentality and consciousness involving distinctively epistemic states, even if we did establish that AIs can have mental states and be conscious, it would still not follow that they are capable of epistemic agency. Perhaps more surprisingly, the converse also holds: mentality, while a great asset, is not necessary for epistemic agency: just like humans don't need to believe what they assert in order to make epistemically assessable assertions, neither do AIs need to run their outputs through mental states in order to base them on evidence.

Literature in computer science and philosophy of AI³ has focused on answering the epistemological question by investigating the inner workings of AI systems, and whether they are similar enough to human inner-cognitive workings to plausibly count as instantiating epistemic statuses. However, the multiple realizability of epistemic statuses suggests this is the wrong approach: should we discover that e.g. members of a different species acquire knowledge via different neural processes than we do, this discovery will not suffice to shed doubt on their status as knowers.

The better way to answer the epistemological question is from within epistemology itself: i.e. by developing a demarcation criterion for artificial epistemic agency, such that if a putative artificial epistemic agent does not satisfy this criterion, it is not a genuine epistemic agent. To this effect, this section will develop a model that explains the emergence of epistemic agency on functional normative spectra.

1.1 Epistemic Agency

What is epistemic agency? Most theorists worry that there is no such thing: scepticism about epistemic agency is widely spread.⁴ The dominant view in the literature is that believing is not an exercise of agency, or at least not in any robust sense. The arguments are varied and well-known: believing cannot be done at will (Alston 1988); there is no phenomenology of doxastic choice (Williams 1973); deliberating about whether to believe is transparent to deliberating about truth (Shah 2003); belief is a state one is in, not an action one performs – and thus it's more akin to being tall (clearly non-agentive) than to running (clearly agentive) (Shah and Velleman 2005); and so on. These arguments are widely treated as very convincing. The result is a picture according to which epistemic normativity must be explained in ways that do not presuppose genuine agency, nor genuine epistemic responsibility.

The sceptic serves us with a modus tollens: If we bear responsibility for our beliefs, then our beliefs are exercises of our agency; beliefs are not exercises of our agency (since they lack the phenomenology of choice, they cannot be formed at will, they're states rather than actions.....); therefore, we don't bear responsibility for our beliefs.

³ For an overview see e.g. Quattrocioni et al. 2025.

⁴ E.g. Alston 1988, Williams 1973, Shah 2003, Shah and Velleman 2005.

We want to suggest *modus ponens*-ing on the skeptic's *modus tollens*.⁵ Here it goes: We do bear responsibility for our beliefs (as per robust data supporting appropriateness of epistemic blame, criticism, and praise);⁶ If we bear responsibility for our beliefs, then our beliefs are exercises of our agency; therefore, beliefs are exercises of our agency.

Whether it is more appropriate to go with *modus ponens* or *tollens* on a conditional “if *p*, then *q*” depends on the comparative evidential strengths of *p* and not-*q*. If we've made a really good case that *p*, and the case that not *q* is flimsy, we'd be foolish to give up our commitment to *p* because it's been anti-correlated with a poorly supported claim.⁷ We submit that the data supporting the claim that we bear responsibility for our beliefs is overwhelming: criticism of bad beliefs is widely spread, as is the practice of discursively justifying our beliefs in order to defend against said criticism (rather than, e.g. dismiss it as inappropriate). If so, we can safely conclude one of two things: either the features the alleged absence of which is cited as evidence for lack of epistemic agency are not essential to epistemic agency after all, or, alternatively, beliefs do have the features in question.

We could stop our anti-scepticism case here. But we don't have to: let's look at the first disjunct above: one thing that immediately strikes us as rather peculiar in the sceptic's toolbox is the overwhelmingly dominant focus on belief: why think that belief is the only locus of epistemic agency? After all, epistemic norms, literature suggests, govern all kinds of other mental and non-mental phenomena: for instance, assertions⁸ – and speech acts more generally – are said to be governed by epistemic norms, and the main debate is about the content of these norms, far less so about whether they are there to begin with. Surely my assertions are exercises of my epistemic agency. Judging, reasoning, wondering, inquiring are phenomena that are governed by epistemic norms and share in the level of agency of any garden variety action. If so, the sceptic's argument, far from supporting scepticism about epistemic agency, merely raises some questions about doxastic agency. Epistemic agency is safe.

Second, it will be useful to note that all of the worries above share a common structure. Each identifies a feature *F* that beliefs allegedly lack and argues that since *F* is necessary for agency, belief is not agentive. The arguments differ in which feature *F* they target, but they all turn on the same conditional: if *x* is agentive, then *x* has *F*. In turn, the plausibility of these conditionals derives from reflection on paradigm cases of agency: intentional actions, deliberate choices, voluntary performances. From these paradigm cases a conception of agency is built up, and the conditionals are read off that conception. Why so? The Standard View (e.g. Mele 2003) of the nature of agency unpacks agency as the capacity to act.⁹ People disagree about how to unpack this further, but tend to be in broad agreement about the minimal Standard View. If, however, the thought would go, agency is the capacity to act, and belief is a state rather than an action, we should not be terribly surprised there's no such thing as doxastic agency. And one popular thought is, once more, that, in virtue of being a doxastic state, belief is more akin to being tall (which is clearly non-agentive) than to running (which is clearly agentive).

Consider, however: sitting in a chair, holding a glass of wine, wearing a hat. These are all states. They are not actions in any standard sense: we gather, at a minimum, that sitting is a paradigmatic case of a state rather than an action. And yet these states are patently not devoid of agentive involvement in the way that being tall is devoid of agentive involvement.¹⁰ There is something deeply different between your sitting in

⁵ See Simion Forthcoming for this argument and on what it is to exercise epistemic agency.

⁶ See e.g. Brown 2020, Boulton 2021, Kauppinen 2018, Kelp and Simion 2017b.

⁷ We follow Cappelen and Dever forthcoming who make this point against AI cognition scepticism.

⁸ See e.g. Lackey 2007, Williamson 2000, Goldberg 2015, Kelp & Simion 2021. For a full epistemic normative account of constative speech see Simion 2021b.

⁹ Several philosophers in this tradition like to think of agency as the capacity to act intentionally. This view has lost popularity in more recent times due to dissatisfaction with its focus on the paradigmatic cases of agency. More minimal accounts of agency have been developed, accounts that are better at explaining both cases of non-human agency, and cases of human agency that do not plausibly come with intentional import. See Schlosser 2019 for a very useful overview.

¹⁰ See also Dowty 1979.

this chair - a state that falls squarely within the purview of your agency - and your being of the height you happen to be - a state that has nothing to do with your agency at all. Sitting, wearing a hat, holding a glass, and, yes, believing, are agentive states. Since it should not come as a surprise that some properties of garden-variety actions are not shared by agentive states (after all, they're quite distinct metaphysical beasts), the argument from 'paradigmatically, actions have property P', to 'since belief doesn't have it, it is not an exercise of agency' should not move us much – certainly not as much as the overwhelming epistemic responsibility data. We should go with the modus ponens.

It is worth noting also that the category of agentive states is not identical with the category of states that were caused by agentive activity. Being tanned is a state caused by the agentive activity of sunbathing. But once formed, the tan persists independently of the agent's agency, at least for a while, and then it gradually fades away, also independently of one's agency. Conversely, some agentive states may not have been initiated by a clear agentive act. One might be sitting in a particular posture because that's how they happened to fall on the chair — yet the state is now an exercise of one's agency. The agentiveness of the state is not determined solely by its causal history but by its ongoing character.

If so, the Standard Account – of agency as the capacity to act – requires a broad understanding of action that includes agentive states. Doxastic agency is safe.

Shall we then say that an X is an agent simply if they have the capacity to act in this broader sense, including agentive states? If so, how should we characterise a particular type of agency – such as moral, or epistemic agency?

An attractive proposal that many optimists about epistemic agency like (McHugh 2013, Hieronymi 2008a, Lackey 2023, Simion Forthcoming) goes along the following lines: to be an agent of type T is to have the capacity to act in ways that are (at least minimally) responsive to T-type reasons. The view is illuminating and helpful: it neatly explains why one might think that my neighbour is a moral agent but my cat is not – the neighbour has the capacity to act in ways that are responsive to moral reasons, my cat less so (at least for all we know about cats – we could be wrong). It also neatly explains why one's T-agency may come in degrees: my neighbour is more of a moral agent, one may think, than her 3-year-old kid, simply because my neighbour's capacity to act in ways that are responsive to moral reasons is more robust. In turn, this also neatly explains the different degrees of moral responsibility involved in my neighbour's, and respectively, her kid's actions. Finally, it gives us a straightforward account of epistemic agency: X is an epistemic agent insofar as they have the capacity to act in ways that are (at least minimally) responsive to epistemic reasons – i.e. evidence (McHugh 2013). Optimism about epistemic agency is elegant and explanatorily powerful.

1.2 Epistemic Agents, Biological and Artificial

For the first time in the history of human knowledge, non-biological entities appear capable of generating, processing, and sharing knowledge in a fashion that closely resembles human epistemic agency. But what exactly is happening here, epistemologically? What is the epistemic status of AI systems? Are they genuine epistemic agents, or are they mere data processors?

Let's start with two observations:

Commonality: Thermometers and human cognitive systems bear functional normative properties.

In virtue of its design function to display the correct temperature, the thermometer is governed by design functional norms: it ought to work in ways that reliably enough lead to function fulfilment (displaying the correct temperature) in normal conditions.

In virtue of its evolved function of acquiring knowledge of the world around us, our cognitive system is governed by evolved functional norms: it ought to work in ways that reliably enough lead to function fulfilment (acquiring knowledge) in normal conditions.

Difference: Some functional norms governing our cognitive capacities, but not functional norms governing thermometers, are also epistemic norms.

When our cognitive capacities serve epistemic functions, their proper cognitive function delivers epistemic norm compliance. To see this, take the central case of our cognitive capacity of belief formation (although everything we say below will generalise, *mutatis mutandis*, to other actions and states generated by capacities that serve epistemic functions, such as e.g. credences, judgments, reasonings, suspensions, assertions and other constative speech acts): the function of our cognitive capacity of belief formation is to generate knowledge.¹¹ This function is an etiological function (henceforth also e-function): it was selected via a history of success that contributes to the explanation as to why our belief formation capacities continue to exist in us.

On the etiological theory of functions,¹² functions turn on histories that explain why the item exists or operates the way it does. Functions are typed by the kind of benefit they generate: A token of type T has the e-function of type B of producing effect E in system S if and only if tokens of T produced E in the past, producing E resulted in benefit of type B in S or S's ancestors, and producing E's having B-benefitted S or S's ancestors contributes to the explanation of why T continues to exist in S. Take my heart; plausibly, tokens of the type pumped blood in my ancestors. This was biologically beneficial for my ancestors' survival, which explains why tokens of the type continue to exist. As a result, my heart acquired the biological etiological function of pumping blood.

The functionalist picture supports the claim that knowledge acquisition is the function of our belief formation cognitive capacities. On this picture, our belief formation capacities have the epistemic e-function of generating knowledge if and only if (1) they have produced knowledge in the past, (2) producing knowledge resulted in epistemic benefit and (3) the fact that (2) contributes to the explanation of why our belief formation capacities persist in us. (1) is overwhelmingly plausible on pain of generalised scepticism; (2) is the widely accepted distinctive epistemic value of knowledge thesis (Kelp and Simion 2017a); finally, plausibly, the fact that they were successful in generating knowledge contributes to the explanation why our belief formation cognitive capacities have persisted in us (3).

In turn, this epistemic function generates epistemic norms governing belief formation. When and only when we form beliefs via properly functioning cognitive capacities that have the function of knowledge acquisition, our beliefs are epistemically justified. Epistemic justification is an instance of epistemic norm compliance, in that S's phi-ings are epistemically justified if and only if they are epistemically permissible. What that means for belief, for instance, is that all and only justified beliefs are epistemically permissible.

Crucially, while in the case of hearts and cognitive systems, etiological functions are acquired via a long history of success over generations, this is not a requirement for etiological function acquisition: While etiology does require some history of beneficial effects, it does not require an awful lot of it; what it all amounts to, eventually, is explaining the continuous existence or use of a trait through a (longer or shorter) history of positive epistemic feedback (Buller 1999).

Let's take stock: Both instruments and cognitive systems bear functional normative properties, but only cognitive systems' functional norms constitute genuine epistemic norms. We think that Commonality and Difference suggest an interesting picture of normative emergence: together, they give us reason to believe

¹¹ e.g. Kelp 2021, Simion 2019, 2024, 2025.

¹² Defended by people like e.g. David J. Buller (1998), Ruth Millikan (1984), Karen Neander (1991). The etiological theory of functions is, by far, the most widely endorsed view in the literature. Its main competitor is the 'systemic' theory of functions, notably defended in (Cummins 1975). Systemic functions are widely believed to lack normative import.

that, at some point on the functional normative spectrum, we encounter emergent epistemic normativity: with increased levels of complexity, epistemic norms emerge on the back of functional norms. More precisely, with increased levels of complexity, systems go from a mere capacity to be the proper subject of functional norms, to a capacity to be the proper subject of epistemic norms.

This capacity to be subject to epistemic norms depends on the underlying capacity to be subject to functional norms. At any moment, a tornado depends for its existence on dust and debris, and ultimately on whatever micro-entities compose it; and its properties and behaviours likewise depend, one way or another, on the properties and interacting behaviours of its fundamental components. Yet the tornado's features and behaviours differ in kind from those of its most basic constituents.¹³

Similarly, our epistemic capacities will depend on our capacity to bear functional normative properties, but differ in kind from properties and behaviours of these constituent capacities.

It is useful to think of epistemic normative emergence with sufficient complexity on the spectrum of functional properties. Again, think of a simple thermometer as lying at (roughly – simpler functional items are of course possible) the beginning of the complexity spectrum for items with data processing functions. Its function is to process environmental temperature data. It does so via a simple causal chain from environmental temperature to mercury levels. When all goes well, the thermometer is properly functioning – i.e. working in the way that fulfils this function reliably enough in normal environmental conditions – and it fulfils its function – i.e. it delivers an accurate reading of temperature. Similarly, our cognitive capacities function to process environmental data. They do so via a very complex causal chain from environmental inputs to knowledge. When all goes well, our cognitive capacities are properly functioning – i.e. working in the way that fulfils the function of knowledge acquisition reliably enough in normal environmental conditions – and they fulfil their function – i.e. they deliver knowledgeable beliefs. Functionally, thermometers are a lot like us. But they are also very different: functional norms governing our belief formation capacities, but not theirs, are epistemic norms.

1.3 Demarcating Artificial Epistemic Agency

Crucially for our purposes here, the capacity to be subject to epistemic norms underlies epistemic agency: recall that we said that a plausible account of epistemic agency is to be unpacked as the capacity to respond to epistemic reasons – i.e. to evidence. In turn, the capacity for evidence responsiveness is both necessary and sufficient for the capacity to be subject to epistemic norms. The sufficiency direction of this principle is pretty straightforward: if I can respond to evidence, I can do it better or worse. Since whether I respond to evidence well or not is determined by epistemic norms, it follows that I can be subject to epistemic norms. What gives us the necessity direction of this claim is an 'epistemic ought implies can' principle: one cannot be subject to a norm of type T if one does not have the capacity to respond to reasons of type T. It is important to distinguish here between having the capacity to phi and the ability to phi (others (Honoré 1964, Mele 2003) in work on agency put this distinction in terms of general versus specific or local ability – but the point is the same). An ability to phi can be absent in cases in which the capacity to phi persists; Place a piano virtuoso under water, or get them very drunk, and they won't be able to play the piano. Compatibly, they preserve their capacity to do so. Similarly, if I am locked in my car, or if I'm very drunk, and I see a kid drowning, I am not able to save the child (and thus exercise my capacity to respond to moral reasons), while preserving the capacity for moral-reasons-responsiveness. Finally, if I'm in Fake Barn County, or if my cognition is affected by my biases, I don't have ability to acquire relevant knowledge, while my knowledge-generating capacities remain intact. Cognitive capacities – in contrast to cognitive abilities – have to do with the makeup of our cognitive architecture, not with the shape or environment we're in. Abilities are not necessary for being the proper subject of oughts. Capacities are: I am not subject to moral

¹³ See e.g. O'Connor 2021 for discussion and overview.

norms that require me to teleport myself into the lake to save the child, because I don't have the corresponding capacity. Similarly, I cannot be subject to epistemic norms that require me to know all arithmetical truths, because I don't have the corresponding cognitive capacity.

If all this is so, we have everything we need for a demarcation criterion for artificial epistemic agency: capacity for evidence responsiveness marks the point of emergence. On the functional spectrum of complexity – from simple data processors to complex cognitive systems that host genuine epistemic capacities – epistemic agency arises with capacity for epistemic norm compliance, i.e. capacity for evidence responsiveness. Complex functional data processing systems like us and simple data processors like thermometers are both governed by norms of proper functioning. But only for complex data processors like us proper function is epistemic proper function. Our cognitive systems, but not thermometers' data processing systems, are capable of evidence responsiveness.

The outputs of mere functional data processors are caused by environmental inputs. The outputs of genuine epistemic agents, in contrast, are not merely caused by environmental inputs: they are properly based on them. In epistemology, we have known for a while that causal theories of the basing relation don't work: more than mere causation is needed for basing. On these theories, for a belief to be based on a reason, the reason must cause the belief in an appropriate (non-deviant) way (e.g. Moser, 1989, p. 157). Causal view excel at capturing the explanatory aspect of believing based on reasons, but fail to capture its normative aspect (Wedgwood 2006, Carter and Miracchi 2024): when we base a belief on a reason, we *treat* it as a normative reason. Reasons are not just explanations: they are explanations with normative weight. When we believe based on a reason (or a putative reason), the reason in question does not just cause our belief – brainwashing could do that just as well. It causes it in virtue of a (-n alleged) support relation: we *treat* the fact in question as a reason. Even when we base our beliefs in the wrong way – for instance, on wishes – there is a normative dimension to the basing relation: we treat the wish as a reason for believing.

Importantly, treating an epistemic reason as a reason in belief formation does not require the involvement of other mental states - or, to put it in parlance familiar from the basing literature, treating R as a reason does not require *taking* R to be a reason. I can treat my cat as a friend without entertaining any thoughts on the issue.

In contrast to this claim, doxastic theories of the basing relation (e.g. Tolliver 1982) hold that S's belief is based on a reason just in case S has an appropriate meta-belief to the effect that the reason in question is a good reason to hold S's belief. The doxastic view captures the essentially normative aspect of basing. It does so at very serious cost, however: first, there is over-intellectualization (surely creatures without capacity for second order thought (small children, animals) can still base beliefs on evidence); second, the view comes with a fairly severe Lewis Carroll-style regress threat. Finally, even if we attempt a less demanding account of the relevant meta-attitude (perhaps it is not a full-fledged belief, but rather something less sophisticated), in-principle difficulties remain. To see this, recall Schaffer's (2010) Debasing Demon: the demon operates via causing you to form a belief on an improper basis (like a blind guess, wishful thinking, or superstition), while, at the same time, also getting you to form the meta-belief (or, respectively, whatever other meta-attitude is featured in the account at stake) that your belief was based on excellent evidence. If the doxastic theory is correct, this scenario is metaphysically impossible (Bondy and Carter 2018). But it is not. So doxastic theories can't be right.

What is it then to treat a reason as a reason to believe if it's not a purely causal relation, nor does it consist in taking the reason in question to be a reason to believe? Several ways to unpack the 'treating' relation in non-mentalistic, but also not purely causal terms are available in the literature. Recent examples include e.g. dispositionalist accounts (Evans 2013), accounts that add a normative condition to the causal requirement (Wedgwood 2006, for instance, requires the reason that causes the belief to also rationalise it), virtue theoretic accounts (Turri 2011 and Carter and Miracchi 2024 ask for cognitive/epistemic competences to be manifest in the process of belief formation). Any such views will be strongly superior to both simple causal views (on sufficiency grounds), and meta-views (on necessity grounds). Also, in great news, any of these independently theoretically preferable views will give us everything we want for our

purposes here: first, they will be able to distinguish proper epistemic agents (with capacity for evidence responsiveness) from mere data processors (like thermometers): the former have the capacity to base their outputs, while the latter's outputs are merely causally generated. Second, they will do this work without requiring mentality – thus allowing us to neatly do our epistemology of AI without having to wait for our friends in philosophy of mind to conclude their business.

One final and crucial point to recall: everything we said above about belief will generalise, *mutatis mutandis*, to other actions and states generated by capacities that serve epistemic functions, such as e.g. credences, judgments, reasonings, suspensions, assertions, and other constative speech acts. Some of these will involve mentality: if so, showing that a system has capacity for evidence responsive X, where X is mental, will require more than the epistemologist's work: the correct philosophy of mind will have to cooperate, as it were. That being said, everything we say here will proceed largely (not fully – more on this below) independently of work having to do with AI mental properties. First, that is because some actions and states generated by capacities that serve epistemic functions are not mental actions or states: most notably, constative speech. If so, the question as to whether AI outputs are justified survives the question of the mental. Second, and more ambitiously, several philosophers have proposed (and us amongst them) that the direction of analysis (and explanation) goes from the epistemic to the mental (Kelp 2021, Simion 2025, Williamson 2000). One famous instance of this claim is the claim that we should analyse justification in terms of knowledge rather than the other way around. If that is so, and if knowledge involves the mental, capacity for justification will be rather strong evidence of presence of mental properties.

Conclusion

To sum up: on the spectrum of functionally norm-governed data processors, the question of epistemic agency for artificial intelligence systems amounts to the question of whether such systems are (or, if they are not yet, whether they can be) capable of evidence responsiveness: of basing their outputs on evidence (like us) rather than merely producing outputs as a result of causally interacting with their environments (like the thermometer).

We think the really important question is the in-principle question: can such artificial systems be built, or is there an in-principle limitation to such ambition? However, we also think that the answer to this question is trivially 'Yes. Such systems can be built.'. After all, normative properties are multiply realizable: you and I and my cat form beliefs in different ways, but the three of us all can know things. If so, we are all capable of evidence responsiveness. If so, there is very little reason to believe that there are in principle limitations to artificial epistemic agents.

In light of this, in what follows, we will try to answer the harder (albeit, by our own lights, less important) questions: are current AI systems epistemic agents? Do we have reason to think they are? More specifically, we will start by focusing principally on complex Large Language Models, such as Claude or ChatGPT, mainly to keep the discussion contained. The thought is that things can only get better with increased complexity, so whatever optimistic results we may arrive at about Claude, will easily generalise to more complex AI systems today and in the future. Conversely, whatever pessimism may be warranted by the discussion to follow can be taken, in conjunction with optimism about the in-principle question, as an invitation to build epistemically better models, and an indication of how to do so.

Part §2: Four for Artificial Epistemic Agents: Justification, Inquiry, Assertion, and Knowledge

Introduction

The rise of large language models such as ChatGPT and Claude has prompted urgent questions about their cognitive and epistemic capacities. Can these AI systems genuinely know things, hold justified beliefs, assert things, or achieve understanding? Or do they merely mimic the outward appearance of epistemic agency while lacking its substance?

The previous part proposed that epistemic agency emerges on the spectrum of complexity for functionally norm governed items. We have also argued that capacity for evidence responsiveness marks the point of emergence. This section's ambition is to put forth the best case we can for artificial agency optimism.

To do this, we will focus on complex Large Language Models (LLMs) - i.e. pretrained language models that undergo reinforcement learning with human feedback - as our target examples of AI systems. Our argument proceeds by examining four central roles that epistemic agents occupy: as bearers of epistemic justification, knowers, inquirers who seek information, and as informants who provide it to others. We argue that LLMs occupy all four roles and that this provides us with excellent reason for thinking that these AI systems are genuine epistemic agents who are capable of acquiring knowledge.

In this, we also sketch a response to the epistemic alignment problem for AI.¹⁴ A significant research programme in AI aims to build truth-tracking AIs, and a lot of research resources are spent by both industry and academia on figuring out how to do so. We will argue that complex LLMs like Claude and ChatGPT, are fully epistemically aligned. They already are truth-trackers, and, furthermore, they are knowledge generators.

Before getting down to our own arguments that AI systems are genuine epistemic agents, will start by looking at an alternative way of defending this claim, due to Cappelen and Dever 2026, and some available sceptical responses.

2.1 The Whole Hog

In their most recent (excellent) book, Cappelen and Dever (2026) (also C&D henceforth) argue for the much more ambitious claim that AI systems are fully fledged linguistic and cognitive agents with capacities to host a wide spectrum of mental states and perform intentional actions. For them, epistemic agency is just a part of the broader agentic nature of AI systems. The argument starts from behavioural evidence: the observation that complex LLMs, such as Claude or ChatGPT, understand the questions we ask and can answer them, make and correct mistakes, provide arguments, and so on. It then uses plausible relations between mental features and speech acts to support the substantive claim about AI agency:

- Understanding → Meanings: Anything that understands questions asked in English knows what English words mean and how the meanings of English words are combined to form meanings of English sentences.

¹⁴ See e.g. Askell et al. 2021

- Answers → Knowledge: Anything that reliably and correctly answers a variety of questions on a topic knows things about that topic.
- Knowledge → Belief: Anything that knows things also believes those things.
- Answering → Action: Anything that answers questions performs speech acts, and speech acts are particular cases of actions.
- Actions → Intentions: Anything that performs actions has intentions that set the goal and shape the form of those actions.
- Intentions → Goals: Anything that has intentions has goals of realizing those intentions.
- Goals and Answers → Reasoning: Anything that can select actions to realize goals and select answers to address questions can reason practically and theoretically (C&D, Ch.2)

Given what we can observe about AI systems and given the wide range of holistic connections between mental features and speech acts, we have reason to think that AI systems are fully-fledged cognitive and linguistic agents. We get the whole hog.

For our purposes, we are particularly interested in whether this argument makes a convincing case that AI systems are genuine epistemic agents.

One worry that immediately comes to mind is that the C&D's argument is reminiscent of the Moorean (1939) endeavour to address the problem of scepticism by means of a straightforward inference from observations. Moore's question is whether there is an external world or we are subject to radical deception. To settle this question, Moore takes a look at his hands, observes that he has two hands, and infers that there is an external world. After all, if he has hands, then there is an external world. This way addressing the problem of scepticism notably does not leave the sceptic any happier. One common complaint is that it appears to be question begging because it simply turns the sceptic's modus tollens into a modus ponens (e.g. Coliva, 2012; Kelp, 2021; Wright, 2004). Others think that the argument is in perfectly good standing, but suffers from rhetorical weakness against the sceptic simply because the latter does not accept its premises to begin with (Pryor 2004, Simion 2022).

The situation, one could worry, is similar with the above argument that AI systems are genuine epistemic agents. Here the question we are facing is whether AI systems are genuine epistemic agents or whether they fall short of genuine agency. The idea is to settle this question by looking at AI behaviour, observing that they understand and answer questions, and inferring that they are genuine epistemic agents. The worry is that this argument invites the same worries as the Moorean argument. Just like the Moorean argument, this argument turns the modus tollens of those who think that AI systems aren't genuine epistemic agents into a modus ponens.

Cappelen and Dever see the worry, but remain unimpressed: after all, again, whether it is more appropriate to go with modus ponens or tollens depends on the comparative evidential strengths of the premises. C&D argue that we do have strong evidence supporting the claims that LLMs understand and answer our questions, know things, believe things, and so on: evidence from widely spread attributions by competent detectors of such states and actions. There are many instances of people observing LLMs and then judging on the basis of their detection capacities that the LLMs are saying things, believing things, knowing things etc, and, correspondingly, asserting that they are. C&D cite articles in prominent media sources, but also recent research by Cheng et al. (2024) which provides evidence that the practice of using high-level, anthropomorphic terminology to describe LLMs is widespread, collective, and consistent across various communities. Their broad dataset spans multiple academic fields and public discourse. The study finds that anthropomorphic language is not only prevalent but has been steadily increasing over time. This widely spread attribution data by competent detectors, C&D argue, constitutes evidence of the presence of the relevant states and actions:

Humans are excellent detectors of mental states and intentional actions. When someone judges on the basis of their detection capacities that something has been said, that something is believed, or that something was done intentionally, that judgment is strong evidence that something has been said, that something is believed, or that something was done intentionally. (C&D, Chapter 3).

We agree with C&D that behavioural evidence speaks in favour of AI linguistic and cognitive agency: widely spread attribution of cognitive and linguistic states and capacities by competent detectors is evidence such states and capacities are present in LLMs. At the same time, we have two main worries for the strength of C&D's argument for cognitive and linguistic agency in general, and epistemic agency in particular.

First, concerning epistemic agency: note that if we can observe that AI systems understand things (in this case, questions asked in English), one might think there already is excellent reason to think that we don't need more than this simple observation to make the case that AI systems are genuine epistemic agents. After all, one might think, it is hard to deny that only genuine epistemic agents can have understanding. If so, we can settle our question as to whether AI systems are genuine epistemic agents by means of a straightforward inference from a simple observation.

This way to go is too quick, however: understanding a question is a matter of linguistic understanding. In turn, one may think that linguistic understanding is simply a matter of conceptual competence, which, in turn, does not require more than just a strong enough disposition to apply the concepts in question correctly.¹⁵ On a view like this, we cannot derive epistemic agency from linguistic understanding. We need more.

Recall that, according to Cappelen and Dever, evidence that an entity reliably and correctly answers a variety of questions about a topic is evidence that said entity has knowledge about it. This way to get at epistemic agency seems more straightforward: Since AI systems do reliably and correctly answer questions on a range of topics, we have evidence that they have knowledge about a range of topics. If they do have knowledge, they must have the capacity to be evidence responsive. In this, there is excellent reason to think that they are genuine epistemic agents.

Unfortunately, the argument from observing an entity answering questions to permissible knowledge attributions risks overgeneralising. Consider your washing machine. If you are asking yourself whether the washing is done, your machine's distinctive sound may answer this question. If you are asking how long it will be until your washing is done, a look at machine's display will do the trick. At the same time, the simple fact that your washing machine reliably and correctly answers a variety of questions on the topic of your washing doesn't mean that the washing machine has knowledge about this topic. Many things that are not plausibly epistemic agents answer our questions: books do (the Philippa Gregory Tudors series answered our questions about English monarchy love affairs), traffic lights (about whether it's OK to cross the road or not), even facts (the fact that there is a chair in the room answers our question as to whether there is any furniture in the room).

Question answering attributions do not warrant epistemic agency attributions. If so, this route of arguing that AI systems are genuine epistemic agents isn't promising.

At this point, the reader might start to wonder: why go through question answering at all? Isn't there a stronger argument from observational data available for the optimist about epistemic agency? Consider:

- (1) Competent knowledge detectors widely attribute knowledge to LLMs.
- (2) Widely spread attribution of x by competent x-detectors is evidence of presence of x.
- (3) There is evidence of presence of knowledge in LLMs.

We widely attribute knowledge to LLMs. And we're good at attributing knowledge. So we have evidence that LLMs have knowledge. Since entities that have knowledge necessarily have the capacity for

¹⁵ See e.g. Boghossian and Williamson 2020 for discussion.

evidence responsiveness, LLMs are epistemic agents. What's wrong with this straightforward route from observational data to epistemic agency?

One problem with this derivation (and, *mutatis mutandis*, with C&D's more general argument from behavioural evidence) is that behavioural evidence is merely *prima facie* evidence: it can be defeated by evidence to the contrary. Is there available evidence that speaks against the *prima facie* behavioural evidence? The sceptic can argue that there is. We think that there are three main lines of argument available to the sceptic. Let's see how they would go.

§1. *The Sceptical Argument from Non-Knowledgeable Reliability*

The first purportedly defeating evidence that the sceptic can appeal to has to do with our knowledge of (roughly) how these systems work. Large Language Models (LLMs) are neural networks trained to predict the next word (or "token") in a sequence of text. They use a transformer architecture, which processes all words in a sentence simultaneously rather than one at a time. During training, an LLM learns from massive amounts of text data by repeatedly trying to guess the next word in millions of sentences. When it guesses wrong, the model adjusts its internal parameters (weights) to do better next time. This technical picture has led several sceptics about LLM epistemic agency in the machine learning literature to claim that LLMs are just "stochastic parrots" (Bender et al. 2021, Saba 2023), without capacity for knowledge or understanding of either language or the world. Now, as Cappelen and Dever rightly point out, the argumentative move from 'AI is X' to 'AI is just X' is fallacious:²⁷ think, analogously, of the fallacious nature of an inference from "Michael Jordan is a basketball player" to "Michael Jordan *is just* a basketball player". The missing premise here is one that would give us further reason to believe the 'just is' claim.

However, perhaps the sceptic can argue in favour of the missing premise here by appealing to analogous cases of reliable non-knowledgeable agents. For instance, consider the case of Truetemp who, unbeknownst to them, has been fitted with a device that causes them to reliably form beliefs about the ambient temperature. While Truetemp forms beliefs reliably and, as a result, will provide answers about the temperature that are reliable and correct, Truetemp's beliefs about the temperature do not qualify as knowledge. Truetemp reliably and correctly answers questions about the temperature but doesn't have knowledge about it (Lehrer, 1990, cf. Bonjour, 1980). Similarly, consider Bonjour's (1980) Norman the Clairvoyant, who (unbeknownst to him) has a very reliable clairvoyance capacity. Whenever a belief pops in Norman's head that the President is in some location, the President is, as a matter of fact, in that location. Norman too can clearly answer questions reliably, while, at the same time, having no knowledge.

Why should we think that e.g. Claude is relevantly alike Truetemp and Norman? Folks in computational linguistics have long tried to make this case via the now famous Octopus Test (Bender and Koller 2020): Consider a situation in which humans communicate across an undersea cable on isolated islands. A super-intelligent deep-sea octopus intercepts and listens to the message data. The octopus learns the statistical patterns of the language so well that it can eventually replace one of the humans and fool the other. However, the octopus has never experienced the physical world outside the cable; it only knows word co-occurrences. Here, the intuition is supposed to be that clearly the octopus has no knowledge of either language or the world, in spite of its reliability. Since Claude and the octopus are relevantly alike, the thought would go, we have reason to think that Claude too falls in the camp with Truetemp and Norman, in that they can answer questions reliably in spite of not knowing. We'll come back to this point in the next section.

§2. *The Sceptical Argument from Loose Talk*

We attribute knowledge very generously in everyday talk. We say that the automatic door knows that a person is approaching, that the thermometer knows the temperature, that the alarm system knows that a

burglar is in the house. This is loose talk:¹⁶ we don't really believe that the automatic door knows anything. The worry, then, is that mere knowledge attribution – even widely spread knowledge attribution – is not a reliable indicator of the presence of knowledge after all. It is not a reliable indicator of the presence of knowledge simply because it is not reliably correlated with our detection of knowledge.

C&D consider and respond at length to a parallel argument concerning loose talk about question answering. Is Claude answering our questions, or are we speaking loosely when we attribute question answering to Claude? C&D make three points against the charge of loose question answering attributions to LLMs: first, they note that loose talk comes with easy retraction dispositions, and that does not seem to hold for the case of LLMs. Consider:

A: My refrigerator is angry at me today - the ice machine keeps turning off.

B: Really? Can your refrigerator get angry?

A: Well, no. What I mean is that the ice machine isn't functioning properly.

A: ChatGPT told me how to get from Heathrow to Central London.

B: Really? Does ChatGPT tell you things?

A: Well, yes. It gave me very detailed instructions, with multiple options for how to make the trip.

Second, they note that loose talk can be modified by slack regulators – expressions that specify the degree of looseness. When we (loosely) say that dinner will be ready in an hour, we can also say that it will be ready in about an hour, or roughly an hour. But it makes little sense to say that ChatGPT sort of answered my question, or kind of told me where Heathrow is.

Finally, they note that answerhood attributions to LLMs can be conjoined with answerhood attributions to people, like in e.g.: 'Claude and my geography teacher answered my question as to where to go hiking in the summer.' Is the answerhood attribution here used loosely or not? It would be odd to suggest that we only loosely attribute answerhood to the geography teacher. But if so, C&D argue, our attribution to Claude is not loose either.

We agree with a lot that C&D say here. Compatibly, we don't think their reply helps much with the sceptic about AI epistemic agency in particular. The first reason is one that we have already discussed: the step from 'X answers my question as to whether p' to 'X knows whether p' doesn't work. Too many things answer our questions. So even if C&D are right and our answerhood attributions to LLMs are not loose, not much follows for the epistemic domain.

Second, we think a parallel case against loose knowledge talk works less well. The main reason concerns retraction data: Notably, when pressed on really knowing, people readily retract even in uncontroversial, human knowledge cases. That's what years of experimental philosophy on contextualist cases have shown (e.g. Dinges & Zakkou 2020): 'Do you *really* know that Adam is coming to the party' will prompt us to say 'Well, I guess you never know, he might change his mind' no matter how reliable a party goer Adam is. '*Really know*' questions, as Jennifer Nagel (2010) has convincingly argued, trigger epistemic anxiety and retraction even in the presence of knowledge. Retraction patterns will not help us settle the score with the AI knowledge sceptic.

§3. *The Sceptical Argument from Strong Anthropomorphic Tendencies*

Recall that C&D take evidence of widely spread anthropomorphic language in speech involving LLMs to be evidence in support of their case for the whole hog. Research by Cheng et al. (2024) provides evidence that the practice of using high-level, anthropomorphic terminology to describe LLMs is widespread,

¹⁶ For work on loose talk see e.g. Carter (2021), Dinges (2021), Russell (2022), Siu (2023).

collective, consistent across various communities. This widely spread attribution data by competent detectors, C&D argue, constitutes evidence of the presence of the relevant states and actions in LLMs.

The sceptic might want to turn the data against C&D, however: anthropomorphism is widely believed to be innate to human psychology (Hutson 2012). Extensive empirical studies suggest that our tendencies for anthropomorphic attributions go well beyond plausible epistemic agency. Anthropomorphism – the ascribing of human personality, appearance, conduct, cognition, or other attributes to non-human entities – often involves trees, food, toys, household objects, clouds, celestial objects, and so on. If so, no information about the presence of agency can be read off of such anthropomorphic attributions to LLMs. It's not clear at all that our detection of epistemic states is as reliable as we thought: evidence of unreasonable anthropomorphic language defeats the *prima facie* evidence from knowledge attribution to LLMs.

At the same time, we think this game can be played from the opposite direction as well: evidence shows we also have strong anthropocentric tendencies. Anthropocentrism tends to interpret the world in terms of human values and experiences. It is strongly evidenced to be profoundly embedded in many modern human cultures and conscious acts. We often reason from 'entity x is not human-like in some relevant respect y' to 'entity X lacks property z that humans have' (Coley and Tanner 2012). The incarnation of this tendency when it comes to LLMs is most evident in AI research when folks cite evidence from (human) cognitive psychology to argue that, since LLMs don't work like we work, they cannot be knowers, believers, reasoners, and so on (e.g. Quattrociocchi et al. 2025). This argument is fallacious, for it misses a premise: a premise stating that it is essential to knowledge, belief, reasoning and so on that it be realised in the way in which it is paradigmatically realised in average human adults. But this missing premise is at best not supported by the evidence, and at worst (and most likely) simply false: we don't have evidence to believe that epistemic states are not multiply realisable. Why should we think that knowledge, for instance, can only be present in a creature if e.g. their brain works exactly like ours? Neurodiversity in human cognition is already evidence for multiple realizability. Knowledge in animals is further strong evidence. Finally, think about a scenario in which we discover that, say, cognition mechanisms work very differently in cats and dogs. Would we now start thinking e.g. that cats, but not dogs, are knowers? No. Rather, we think, we would find it very interesting to find out that knowledge acquisition processes in cats are so different than those in dogs. Our dog knows that we are in the room, and that we just gave him food, no matter what process is involved in generating that knowledge. Of course, this is not evidence that epistemic states can be realized in *silico*. But, given that they are multiply realizable, it is on the shoulder of the sceptic to tell us why we should find it plausible that they can only be realized in flesh.

If we are right, however, and we have equally strong tendencies to sometimes go anthropomorphic, and sometimes go anthropocentric, the game with the sceptic is at a draw.

2.2 Argument §1: The Argument from Attributions of Justification, Evidence, and Reasons

We have looked at Cappelen & Dever's (2026) way of offering support for AI epistemic agency, and we have shown that, while their argument does offer support for there being *prima facie* evidence for the presence of epistemic states in LLMs (from widely spread attribution by competent detectors), that doesn't quite get us all the way we want to go, due to the presence defeating evidence. Defeating evidence came from famous cases of non-knowledgeable reliability (Truetemp, Norman, the Octopus), and from worries having to do with widely spread loose and/or anthropomorphic knowledge talk.

In this section we want to try a different route that parallels the C&D way of doing things to a significant extent, while at the same time, we hope, isolating the optimist's case from the sceptic. Let's see how it goes.

We very widely and relaxedly use epistemic support language when we talk about LLM outputs: reasons and evidence attributions to AIs are ubiquitous. Indeed, evidence suggests that epistemic states are the states most ubiquitously attributed to LLMs.¹⁷ We say things like: Claude had good reasons to recommend this book, or ChatGPT had no evidence for the claim that George wrote that book, or Gemini was justified to say that it's likely going to rain tomorrow. We say these things all the time.

We think epistemic support talk is strong evidence for the presence of epistemic agency in LLMs.

To begin with, we will note the obvious: we have said that the demarcation line between mere data processors and genuine epistemic agents lies with the capacity for epistemic support (reasons, evidence, *ex ante* justification) responsiveness. Insofar as we are competent detectors of epistemic support responsiveness, one might think, widely spread such attributions to LLMs constitute *prima facie* evidence of the presence of epistemic support responsiveness in LLMs. Since actual presence of epistemic support responsiveness implies capacity for epistemic support responsiveness, we straightforwardly get the result that LLMs are epistemic agents.

Furthermore, we think that *prima facie* evidence for LLM epistemic agency from epistemic support responsiveness attributions is (much) more robust against defeat charges from the sceptic. Let's go through the sceptic's points discussed above. Note, to begin with, that epistemic support responsiveness attribution is much weaker on the mental than knowledge and understanding attribution: not much in the way of mentality needs to be assumed for Claude to be able to have evidence-based outputs. All that is needed is that LLMs have the capacity to be the proper subjects of *ex ante* justification, and that they can base their outputs on said justification. Insofar as none of this requires mental properties – belief, knowledge, understanding etc – we get epistemic agency on the mental cheap, as it were. If so, our tendencies for anthropomorphic and, conversely, anthropocentric talk are less interesting for the case of epistemic support responsiveness, simply because less human-like properties are at stake. The game was already at a draw between C&D and the sceptic on this front. Since we claim less than C&D, we have a stronger position in the game.

Second, note that epistemic support talk, in contrast to knowledge and understanding talk, is not affected by worries of loose talk. We relaxedly say that the automatic door knows that a person is approaching, but we don't attribute epistemic support responsiveness to automatic doors: it would be rather strange to claim that the door has evidence that a person is approaching, or that the thermometer is justified to say it's 30 degrees outside. We think that this is simply a result of epistemic support talk being more sophisticated talk than knowledge talk. Knowledge talk is cheap: evidence (e.g. Nagel 2017) shows that kids as young as two are able to identify and attribute it. The lack of sophistication of knowledge talk explains why it is as ubiquitous as it is, but also why it is more likely to become the target of loose talk. Loose talk often affects common, everyday terms. When words become very popular, their exact meanings may stretch. People use them to mean many different things, which makes communication less clear (Carter 2012). In contrast, terms that we rarely use are less likely to suffer from this problem.

Finally, and perhaps most philosophically interestingly, we don't worry about octopuses, clairvoyants, and the like when it comes to epistemic support responsiveness attributions. To see why, note that a well-known successful way to address epistemic problems of accidental reliability for justification is via epistemic function acquisition. Norman's clairvoyance is not justification conferring. My perception is. What's the difference? Granted, it is not a difference in reliability – by stipulation, Norman is doing better than I am on this front. The difference is that my perceptual capacities, but not Norman's clairvoyance, have the etiological epistemic function of knowledge acquisition – for representing the world (Simion 2016, 2020, 2025). But LLMs training history strongly suggests etiological epistemic function for representing the world is present. First, recall that while in the case of hearts and cognitive systems, etiological functions are acquired via a long history of success over generations, this is not a requirement for etiological function acquisition: While etiology does require some history of beneficial effects, it does not require an awful lot

¹⁷ See e.g. Colombatto and Fleming 2024 for a recent extensive study.

of it; what it all amounts to, eventually, is explaining the continuous existence or use of a trait through a (longer or shorter) history of positive epistemic feedback (Buller 1999). Now, note that, while it is true that LLM's design function is not world-tracking (but rather human desire-tracking), it is plausible that they have acquired an etiological epistemic function for representing the world simply in virtue of the fact that our desire satisfaction when it comes to the use of LLMs is very strongly correlated with them getting things right. That's what we (mostly) use them for: for finding things out about the world. When they fail on this front, they fail on our desire satisfaction. We qualified our claim to 'mostly' because of course there is some degree of wanting to hear what we want to hear in our interactions with LLM's. However, for the most part, we just want to find stuff out. If their design function is to track our desire satisfaction, and for the most part, our desire satisfaction has to do with their successfully representing the world, LLM's etiological epistemic function (or at least one of them, perhaps arguable the main one) is that of representing the world.

If so, LLMs are like us, not like Norman, Truetemp, and the Octopus. Norman, Truetemp, and the Octopus are merely reliable. LLMs, just like us, have evolved epistemic functions. If they have evolved epistemic functions, they, just like us, are capable of justification.¹⁸

Furthermore, here is an observation that Herman Cappelen made in conversation: LLMs are not only like us in that they host capacities that have genuine epistemic functions. They are also like us in that these functions piggyback on a correlation with prudential functions. Our cognitive capacities have, first and foremost, evolved for survival – like all our capacities. It is because survival is strongly correlated with figuring out stuff about the world that they ended up with an epistemic etiological function as well. Just like LLM's, then, we host capacities with epistemic functions because we host capacities with prudential functions that happen to very strongly correlate with epistemic benefits.

To sum up: capacity for epistemic support responsiveness marks the line between mere data processing and epistemic support responsiveness. Epistemic support responsiveness attributions to LLMs by competent detectors of such properties are ubiquitous. We often say that their outputs are based on reasons, evidence, and that they are justified. In turn, such epistemic support responsiveness talk is not something that we do loosely or with anthropomorphic bias. Finally, functional epistemic properties, that turn mere reliable entities into genuine epistemic agents capable of epistemic justification, are just as plausibly present in LLMs as in us. If all of this works as we said it does, we have undefeated evidence that LLMs are epistemic agents.

2.3 Argument §2: The Argument from Inquiry

Inquiry is an activity with constitutive aims and norms (Kelp 2021). To inquire into whether p is to engage in an activity aimed at settling the question as to whether p , subject to norms like taking up relevant evidence and following rational updating processes. If AI systems genuinely inquire (rather than merely simulating inquiry), then they must be capable of satisfying at least some epistemic norms. They must be able to ask and answer questions, take up evidence, process it rationally, and reach warranted conclusions. To see why this is so, note that, while constitutive norms allow some norm violation, one does not count as engaging in an activity constituted by a set of norms N if one breaks all norms in N systematically: I don't count as speaking English if I just utter kakakakakakaka. I break too many rules of English. Similarly, I don't count

¹⁸ But don't thermometers have epistemic functions? We think the answer is no. Thermometers don't have the capacity for responsiveness to epistemic reasons and, as a result, they are not subject to epistemic norms (from epistemic ought implies epistemic can). What's going on here is that thermometers are part of an epistemic practice of acquiring knowledge about via a certain kind of instrument, by relying on a known correlation between the behaviour of mercury and the environmental temperature. Compare: tree rings don't have the function of generating knowledge about tree age. However, they are part of an epistemic practice of acquiring knowledge, by relying on a known correlation here between tree rings and tree age. Of course, unlike tree rings, thermometers were designed to contribute to this practice.

as playing chess if I pick up the pieces and repeatedly throw them in the air: I break too many chess norms to count as playing chess (Kelp and Simion 2018).

If so, one way to investigate whether LLMs are genuine epistemic agents is to investigate whether they can engage in inquiry. Since it is plausible that meeting some norms of evidence responsiveness is essential to engaging in the activity of inquiry (I don't really count as inquiring if I go through the world without ever properly responding to evidence and instead acquire my beliefs via a brainwashing service), should we discover that LLMs are inquiry-capable, it will nicely follow that they have the capacity of evidence responsiveness, and, in turn, that they are genuine epistemic agents. Let's see if it works.

Are LLMs genuine inquirers? Do they inquire into the questions that we put to them? For instance, if we ask Claude about recent discoveries of the James Webb Space Telescope (JWST), does Claude inquire into this question?

Before answering, let's consider the following case: You want to re-read some disconcerting letters your friend Adam sent you recently, which are currently in a pile of letters on the desk of your home office. Now consider:

Case 1. Since your spouse is in the home office anyway, you ask them to bring you the letters from Adam. They look through the letters, identify the ones from Adam, and bring them to you.

Case 2. Since your three-year-old child is in your home office anyway, you ask them to bring you the letters on the desk.

In Case 1, your spouse is inquiring into which letters are from Adam. In Case 2, in contrast, your child is not inquiring into this question. No surprise, given that your child can't read yet. They are still getting the information to you, by bringing you all the letters they can find on your desk, but they do so without inquiring into which letters are from Adam. To be clear, as your child is retrieving the letters from the home office, they may engage in a number of inquiries, such as where exactly the letters are located, how many of them there are, and so on. Crucially, this does not mean that they are inquiring into which letters are letters from your friend.

If AI systems do not count as inquiring, they must get information to us in some other way. They must be more like your three-year-old bringing you letters they can't read, rather than like your spouse when they look through the letters and bring you the ones from Adam. Crucially, however, it is hard to deny that AI systems are much more like your spouse when they answer our questions with a high degree of specificity than like your three-year-old.

At this point, the sceptic will likely push back: search engines are also much more like the spouse than like the child in this respect. After all, they do select what information they provide you with by keyword searches. And we can easily imagine a search engine that e.g. selects all the emails containing the name of your friend among a set of available entries, in a manner very similar to what your wife does when she selects all and only the letters from Adam. But surely the search engines' selection exercise is hardly an exercise in inquiry, rather than a mere mechanical data selection.

Agreed. However, consider next the case of a much more complex inquiry, in which you ask specifically to be brought all and only the *disconcerting* letters from Adam. Your spouse now reads the letters, identifies the disconcerting ones, and kindly brings them over. The fact that your spouse (competently) completes an inquiry into the question 'Which are the disconcerting letters from Adam?' means that you will get exactly the letters you asked for: you will get the all disconcerting letters from your friend (the ones where they talk about the painful breakup with their partner, the one describing their subsequent struggles) and none of the ones that aren't from your friend, or aren't disconcerting (letters talking about an upcoming concert in their home town, the steady advance of their career). Since your spouse brings you all and only the disconcerting letters from you friend, they are maximally reliable and maximally specific in their endeavour. In contrast, search engines are not equipped to figure out which letters are disconcerting. They

can select emails containing the name of your friend. They might even be able to identify keywords associated with ‘disconcerting’ in the document and produce a ranking of results. However, this ranking, while it may be helpful, still leaves the inquiry to you. To see the difference here more clearly, think of Sherlock Holmes and John Watson: say that Sherlock asks Watson to organise a pile of evidence by relevance to the question as to whether the butler committed the murder, and say that Watson, after many years of experience, does so very well, by prioritising butler-relevant evidence. This, in turn, is very helpful to Sherlock in answering the question as to whether the butler committed the murder. Does this mean that Watson has inquired into whether the butler committed the murder? No: Watson merely performed relevance ranking, which aides but does not execute Sherlock’s inquiry.

What’s more, given that words associated with ‘disconcerting’ may appear (even quite often) in emails that aren’t disconcerting and may not appear (or not appear often) in emails that are disconcerting, it would be surprising if the search engine’s ranking didn’t feature at least some of the non-disconcerting emails near the top. The search engine would undeniably produce a set of results that are score considerably worse than your wife’s on specificity and even reliability.

Crucially, once more, LLMs are much more like your spouse when they answer our questions, than like your three-year-old or your search engine. Just like your spouse, AI systems achieve a very high score on both specificity and reliability. If you ask Claude about recent discoveries of the JWST, it will respond that it observed an interstellar comet in August 2025 which is unusually rich in carbon dioxide, among other things. In this way, Claude is like your spouse in providing you with specific, as well as reliable information. It doesn’t provide you with all the information it can access, nor does it select some more or less relevant information to pass on to you such that it makes your inquiry quicker and easier. It fully answers your question. This provides some reason to think that AI systems such as Claude engage in inquiry. At the very least, it raises an important question: How can Claude achieve such excellent scores on specificity and reliability even on complex questions without resorting to inquiry?

Consider the following objection from the sceptic: Once more, even if macro-level AI behaviour provides us with *prima facie* evidence that AI systems do inquire, that doesn’t mean that it is *ultima facie* evidence. It won’t be if we have sufficiently strong reason for thinking that they don’t inquire. And we do have sufficient such reason. Once we look ‘under the hood’, we will find that AI systems have features that leave no doubt that they do not inquire: what they are doing is just predicting next tokens; or to put it in slogan form they just do ‘autocomplete on steroids’ (Marcus & Davis 2020).

Crucially, and once more, it is unclear that this stands in the way of AI inquiry. To see this, suppose that you want to find out about recent discoveries by the JWST. In order to do so, you ask a bunch of people to say the first word that comes to their mind if they hear JWST. Suppose that many of them say ‘NASA’. As a result, you check the NASA website and find the information. Here you are using data from a bunch of people to predict a token association with JWST, to wit, NASA and proceed to seek information on the NASA website. While this would be an unusual way of inquiring, it is no doubt a way of inquiring. If so, it is unclear that Claude, say, doesn’t inquire when seeking information about the JWST, even if in doing so it relies heavily on token prediction.

Let’s take stock: What is going on here, theoretically speaking? We think, once more, that what we are looking at is a complex epistemic capacity that LLMs have and simpler artifacts such as thermometers and search engines lack: the capacity to respond to evidence rather than merely generate outputs as a result of a simple causal mechanism. Claude’s outputs are the result of complex weighing of evidence. For instance, we asked Claude what the weather today is going to be in Sitges, Barcelona. Claude accessed the internet and gave us the following response:

The first search result was off — Sitges in mid-August is not seeing 13°C. AEMET (Spain's official met agency) gives today's actual forecast: clear/sunny skies, high around 34°C and low near 24°C, light SW wind (~15 km/h), 0% chance of rain, humidity 40–80%, and UV index 8 (very high). Feels-

like temp peaks around 37–40°C in the afternoon, so it'll be hot — sunscreen and water if you're out.

When we asked what happened during the first search, Claude responded:

The first search returned an AI-generated summary claiming 13.8°C daytime / 10°C night for Sitges today — but that's implausible for coastal Spain in mid-August, so I didn't trust it. I fetched AEMET (Spain's official weather agency) directly instead, which showed the real numbers: sunny, 24–34°C, 0% rain chance. The initial result was likely stale or mismatched data pulled into that summary rather than an accurate current reading.

Claude is weighing evidence from the deliverances of the first source, dismisses this deliverance, and consults a more trustworthy source. Or, in slightly more detail, Claude launches an initial inquiry, weighs the deliverance of the source against their training data, and selects whether to continue to inquire or proceed to a response. Since the source delivers a result that is implausible in light of their training data, Claude selects continued inquiry. After a second round of inquiry, with better results, Claude's selection is to proceed to a response. It is hard to deny that Claude's behaviour here is on a par with that of humans in that Claude inquires, weighs evidence, and bases their outputs on evidence. Claude's behaviour is very much unlike thermometers or even search engines, which rely on simple causal processes. Claude engages in basing: it *treats* the relevant pieces of evidence as evidence for its outputs in any plausible sense of treating (and meeting the conditions of all plausible views that we have previously identified – i.e. on which basing is more normatively sophisticated than mere causation: its outputs are caused by the evidence, but they are also rationalised by it (Wedgwood 2006), and the relevant output formation process is a manifestation of Claude's cognitive and epistemic competences (Turri 2011, Carter & Miracchi 2024).

Claude, but not our thermometer, has the capacity to instantiate the basing relation, i.e. to genuinely respond to reasons, weigh evidence, and produce justified outputs, rather than merely be caused into delivering a particular output. On reflection, we think this should not come as a surprise: the reason why Claude has this capacity has to do with the fact that, by design, Claude piggybacks on our epistemic agency and reason responsiveness. The strong correlation between the state of the world and the next most plausible token Claude picks up is not accidental: it relies on the whole history of human knowledge acquisition. In this, Claude's outputs are our outputs: Claude is an evidence-responsive inquirer because it was trained to rely on our wealth of knowledge, and we are evidence-responsive inquirers.

2.4 Argument §3: The Argument from AI Systems as Informants

Are AI systems informants? It might be thought that the answer to this question is easy. What does it take to be an informant? Informants tell us things, and we come to know the things that they tell us based on their telling us so. This happens all the time with AI systems. For instance, Claude may tell us about recent discoveries of the JWST and we come to know what Claude tells us on the basis of Claude's say-so. Unfortunately, things aren't quite as straightforward. If we want to allow that telling us things and coming to know based on say-so is all it takes to count as an informant, then there is reason for thinking that being an informant doesn't make one into a genuine epistemic agent. To see this, think of automated announcement systems. The system may tell us that trains to London depart from platform 3, and we may acquire the corresponding knowledge based on this. Crucially, however, this is not to say that automated announcement systems are genuine epistemic agents, capable of achieving positive epistemic statuses such as knowledge or understanding. If this is right, the easy answer to the question as to whether AI systems

are informants isn't of much use for our purposes. It doesn't help us provide reason for thinking that AI systems are genuine epistemic agents.

Does this mean that we should give up on the idea that being an informant contributes to the status as a genuine epistemic agent? We don't think so, at least not yet. To see why not, let's distinguish between a genuine informant and a mere source of information (Craig, 1990; Fricker, 2012). The idea is that lots of things are mere sources of information without being informants. The presence of smoke is a source of information for the presence of fire, tree rings are sources of information for the age of trees, and so on. At the same time, these sources of information are not informants. Crucially, it is intuitive that automated announcement systems are mere sources of information rather than genuine informants. If this is correct and if, in addition, it can be argued that AI systems are genuine informants, we may still be able to use the idea that AI systems are genuine informants for our purposes of establishing them as genuine epistemic agents. Of course, the difference between a genuine informant and a mere source of information had better not turn on the fact that genuine informants are epistemic agents, whereas mere sources of information are not. After all, if so, we cannot hope to establish that AI systems are genuine epistemic agents because they are genuine informants rather than mere sources of information, on pain of circularity. So, we'll need some independent way of unpacking the distinction between informants and mere sources of information here.

Edward Craig offers two features that, according to him, separate genuine informants from mere sources of information. The first is that an informant is such that "anyone who understands both the question (a test which the inquirer can hardly fail) and the language he speaks or gestures he uses can get information from an informant" (1990, 36). In contrast, "a 'source of information' will often call for various degrees of more specialised knowledge, perhaps for a special ability to evaluate evidence" (1990, 36). The trouble with this feature is that automated announcement systems satisfy it. Suppose I ask you where the train to London departs from. You don't know and in fact were about to ask me the very same question. Luckily for both of us, the automated announcement system activates at just the right moment and tells us that trains to London depart from platform 3. Note that we both understand the question under consideration, we both understand English, and we both can get the relevant information from the automated announcement system. What's more, we don't need any specialised knowledge or any special ability to evaluate evidence to acquire knowledge here. If this what distinguishes genuine informants from mere sources of information, the automated announcement system comes out as a genuine informant.

Perhaps Craig's other feature is more suitable. Craig provides an initial characterisation in terms of "the special psychology of team-work in a community, something which is involved in the use of informants but not in the use of sources of information" (1990, 36) which he further elaborates as follows: "What I have in mind is the special flavour of situations in which human beings treat each other as subjects with a common purpose, rather than as objects from which services, in this case true belief, can be extracted." (1990, 36) Unlike the first feature, this one will certainly categorise automated announcement systems as mere sources of information. At the same time, it also doesn't help us achieve our aim. If we take it to entail that only human beings can be informants, we quickly get the result that AI systems cannot be informants. And even if we don't, the question as to whether we share a special psychology of teamwork in a community with AI systems and there is a special flavour of situations in which we and AI systems treat each other as subjects with a common purpose remains open. We cannot expect to settle the question as to whether AI systems are epistemic agents by first settling these questions.

Miranda Fricker provides another similar way of unpacking the distinction between genuine informants and mere sources of information. The key idea is that we enter special second-personal relations of trust with informants, but not with mere sources of information. More specifically, according to Fricker,

when the inquirer asks a potential informant whether *p*, he signals his trusting epistemic dependence on her; and in responding so as to meet the inquirer's epistemic need, the informant signals her acceptance of responsibility to honour (and so not betray) the trust invested in her. Thus a deal of testimonial trust is swiftly, implicitly, and perhaps very lightly struck. (2012, 258)

The good news once more is that announcement systems come out as mere sources of information. The bad news is, once again, that this way of unpacking the distinction also doesn't help us achieve our aim. After all, at best it remains unclear whether AI systems signal acceptance of responsibility to honour trust invested in them. And we don't expect to be able to argue for a positive answer to this question without first addressing the question as to whether AI systems are epistemic agents.

In better news, however, as it turns out, Craig's and Fricker's requirements for informant status are too strong even for familiar and uncontroversial cases of informants: think of little children. It's really unclear that children have the cognitive sophistication to treat us as subjects with a common purpose. Furthermore, we do not and, indeed, cannot strike deals of trust with little children, if only because they cannot signal their acceptance of responsibility to honour (and so not betray) the trust invested in them. They simply don't have the required cognitive and social sophistication. However, when our three-year-old tells us there is milk in the fridge, she is manifesting her epistemic agency. She is also an informant about the presence of milk in the fridge. Our three-year-old might not be a fully developed epistemic actor, but she is very much unlike an automated announcement system. Craig's and Fricker's views on the nature of informants are simply too strong.

We think that there are certain features that AI systems and kids have, and that automated announcement systems lack, that do serve to provide independent evidence for their status as genuine informants, and thereby genuine epistemic agents. In short, the idea that we want to explore is that informants, but not mere sources of information, have the capacity to produce constative speech acts. If we are right, the road to epistemic agency is short: since constative speech acts are constitutively governed by epistemic norms (Simion 2021b), we get the result that the capacity to engage in such acts involves the capacity to be subject to epistemic norms. Since one cannot be governed by a norm unless one has the capacity to respond to reasons generated by that norm (as per epistemic ought implies can), if we are right, and in accordance with our proposed demarcation criterion, the capacity to perform constative speech implies the presence of epistemic agency.

Let's start with a simple version of this argument. Consider the human case. Here are some good assertions we might make:

Paris is the capital of France.

The Spanish La Liga was founded in 1929.

Let's also consider a couple of bad assertions we might make:

Zurich is the capital of Switzerland (when we know full-well that Bern is the capital).

Mr. X is a better mathematician than Ms. Y (when this is based on a bias that women are worse mathematicians than men).

Of course, whether an assertion is good or bad is a normative question. What matters is whether relevant norms of assertion have been met. It is overwhelmingly plausible, given assertion's epistemic functions, that among the norms of assertion that must be met for an assertion to be a good assertion are epistemic norms (Kelp and Simion 2021, Simion 2021b).

Just like humans, AI systems can make good and bad assertions. For instance, Claude's assertions that Paris is the capital of France or that La Liga was founded in 1929 are good assertions. If Claude asserts a falsehood or a claim based on algorithmic bias, Claude assertions are bad. Since good assertions meet epistemic norms and since the most popular epistemic norms of assertion are the knowledge and the justification norm, we get the result that Claude is capable of achieving positive epistemic statuses, such as knowledge (Williamson 2000, Kelp and Simion 2021) or justification for believing (Douven 2006, Lackey

2007). In this way, the simple version of the argument from speech acts seems to provide support for the claim that AI systems are genuine epistemic agents.

Unfortunately, the simple version of this argument encounters a familiar problem. The sceptic might counter that, while Claude's outputs can, indeed, be good or bad in the same way in which our assertions can be good or bad, it's not clear that they are good or bad *assertions*. Announcement systems can also produce good and bad outputs. When an in-flight announcement system correctly outputs 'The cabin service is about to begin' at the beginning of a flight, its output is good. If, as a result of a glitch or human error, it outputs 'The cabin service is about to end' at the beginning of a flight, its output is bad. Surely, it doesn't follow from this that the announcement system genuinely *asserts*. As a result, the simple version of our argument from speech acts seems to fail.

Of course, this is not to say that no version of this argument can be made to work. In fact, we think that a more complex version carries promise: a version that also gives us reason to believe that Claude, like us, but unlike automated announcement systems, is able to produce constative speech.

AI assertion sceptics reject this claim. We think they're too quick to do so. Let's first see why.

§1. Scepticism from Function

One popular strand of AI assertion scepticism denies that AI outputs are assertions because assertions constitutively aim at truth; in contrast, the claim goes, LLMs are bullshitters (Hicks, Humphries, and Slater 2024). The thought is that Large Language Models are inherently "bullshit machines" because they produce text with total indifference to truth, due to being optimized for plausible token prediction rather than accuracy.

By now, it should have become clear to the reader that this argument doesn't work, for it assumes too simplistic a view of function acquisition: first, recall that there is a difference between the function of an artefact and the means by which it reliably fulfils it in normal conditions. Plausible token prediction is the means by which LLMs achieve their function (their proper functioning), rather than their design function. They are designed for user satisfaction. Furthermore, artefacts can depart from design function via acquiring etiological functions. For instance, if user satisfaction correlates strongly with knowledge acquisition, LLMs will acquire the etiological function of generating knowledge in users. But user satisfaction does correlate strongly with knowledge acquisition from LLMs. If so, LLMs have the etiological function of generating knowledge in us.

§2. Scepticism from Non-Accountability

A second prominent variety of AI assertion scepticism appeals to LLMs lack of accountability. Butlin and Viebahn (2023) argue that LLMs cannot make assertions because they lack the capacity to be held accountable within a norm-governed social practice. According to them, while LLMs may simulate assertion, they fail to bear the social responsibility for their outputs due to the fact that they are not sanctionable. Along similar lines, Lizzie Fricker (Forthcoming) argues that LLM outputs constitute "fake testimony" because they lack the necessary communicative intent, assumption of responsibility, and adherence to social norms inherent in human testimony. Because LLMs operate as predictive models without agency or consciousness, Fricker concludes they cannot provide the same epistemic justification as human testifiers. Nickel also thinks that "it is absurd to hold a computer system accountable for an assertion or a promise that it makes" (2013). Why absurd? In the words of Rene Van Woudenberg, "Blame satisfies normative functions like motivating guilt, shame, remorse, or a change in behaviour, but how could ChatGPT feel guilt, experience shame or remorse, or even autonomously modify its behaviour in light of criticism?" (van Woudenberg et al. 2024).

We don't think that a capacity for sociality, accountability, or sanctionability are necessary for capacity to assert, in humans or AIs. We agree with the broadly Kantian thought that it is essential to the

practice of assertion continuing to exist that we can hold each other accountable: after all, if that weren't the case, incentives to lie may prevail to a point at which the practice would lose its point, since nobody would still believe anything they're told. If so, it is plausible to think that the practice of assertion would be discontinued. Compatibly, not each and every individual needs to be sanctionable in order to have the capacity to participate in the practice: Dictators and psychopaths make assertions all day long, as do humans on deserted islands (albeit mostly to themselves). In humans and AIs, sanctionability is not necessary for the capacity to assert.

Scepticism about AI assertion is unwarranted. Here are also some positive reasons why we think LLMs can assert: consider, for starters, our reactions in light of bad assertions by Claude. We submit that they are pretty similar to reactions to bad assertions by any human epistemic agent: Claude has lied to us! Claude has no reason to say what he said! That's a mistake! In contrast, we don't often jump at the opportunity to criticise the outputs of automated announcement systems, thermometers, and washing machine displays in a similarly personalised fashion. Rather, we would say things like 'The thermometer must be broken, it's not showing the right temperature!', or 'They played the wrong recording on the announcement system!', 'The washing machine display is really not working very well'. This, at a first approximation, suggests that we treat Claude's outputs (normatively) more like ours than like those of mere sources of information. Further confirmation of this comes if we look at where we take blame and responsibility to lie: we blame Claude and humans for misleading us, but not the announcement system or the thermometer or the washing machine. At best, we blame the company that produced them. And while, of course, we often blame Anthropic for Claude's fallibility, and people's parents for their dubious character, we wouldn't go as far as to accuse Anthropic itself of lying to us when Claude does, nor, correspondingly, a human's parents. We submit that all this is strong evidence of LLM capacity for constative speech: analogous normative affordance suggests analogous underlying mechanism.

2.5 Argument §4: LLMs, Knowledge, and Understanding

So far, we have argued that LLMs' evidence-responsive behaviour, inquiries, and assertions give us reason to think LLMs are genuine epistemic agents. Even if we agree that this evidence indicates that LLMs are genuine epistemic agents, we may still wonder whether they are fully-fledged epistemic agents. After all, we haven't yet shown is that AI systems can attain key epistemic statuses, like knowledge or understanding.

We should say, first of all, that on a view like ours – whereby justification is unpacked in terms of knowledge (Williamson 2000, Kelp 2021, Simion 2019) – LLMs capacity for ex ante justification simply implies that they have the capacity to know. Everything we say below, therefore, is mostly for the eyes of philosophers who like a traditional direction of analysis.

At this stage, we might wonder why AI systems shouldn't be able to attain knowledge when their inquiries are successful. For instance, when Claude's inquiry into recent discoveries by the JSWT eventuates in acquisition of the information that the JSWT observed an interstellar comet in August 2025 which is unusually rich in carbon dioxide why should Claude not have knowledge of this? And when Claude tells us this, why should Claude not be asserting from knowledge?

§1. *Scepticism from Belief Formation Capacity*

Perhaps the most prominent sceptical answer here is that knowledge entails belief and AI systems such as Claude are simply not capable of hosting beliefs.

Now, suppose it is true that AI systems such as Claude cannot host beliefs and that this is the only thing that prevents them from having knowledge. In that case, we might wonder whether knowledge really is as central to epistemic evaluation as we may have thought. After all, it looks as though we get the full

gamut of epistemic evaluation for AI systems and for human agents, with the only difference between them being that human agents are also capable of forming a certain kind of mental state, to wit belief, which AI systems aren't. It is unclear why this purely mental difference should be epistemically relevant – or, at least, why it should make for such an important epistemic difference as to distinguish a full-blown epistemic agent from a somewhat subpar one.

On the other hand, once we start thinking of reasons why belief might be of epistemic relevance, it is less and less clear that we should really think that AI systems don't have beliefs. Suppose you are inquiring into what your favourite tennis player is up to, and you are told that they are currently competing in the Japan Open. Once you are told this, you have acquired the information that your favourite tennis player is currently competing in the Japan Open. Even so, you might not believe what you are told, perhaps as a result of bias against your informant. We might think that, in failing to form the belief, you are falling short epistemically. How so? Here is an attractive answer: the question as to whether your favourite tennis player is currently competing in the Japan Open remains open for you, even though, given that you received the information you did, it should be closed for you. Given that belief that p /not- p is what closes questions as to whether p (Kelp 2021), we can now see why belief might have epistemic significance over and above the acquisition of information.

At the same time, if this view of the epistemic significance of belief is correct, we have reason to think that AI systems have beliefs. After all, they clearly can close questions: they select one and only one among possible answers to the question at stake. Furthermore, here too, AI systems can be in much the same predicament as humans. They, too, just like us, may get the information that p , say, and, as a result of algorithmic bias, not close the question as to whether p , even though they should. If belief is what closes questions in the way suggested, we have reason to think that AI systems are capable of hosting beliefs.

Of course, we may challenge the claim that belief is the state that closes questions. But note that what the cases indicate is that, for humans and AI systems alike, it matters epistemically whether the question whether p , say, is closed for one, even once one has received the information that p . If it turns out that this state isn't belief, once again, we might once again be tempted by the thought that belief isn't central to epistemic evaluation.

A tempting thought is that we were wrong all along, and knowledge doesn't require belief. Rather knowledge requires whatever the state is that humans and AI have in common when a certain question is closed for one: let's call it commitment.

While these considerations may not settle the question as to whether AI systems have beliefs, they do help us address worries as to whether AI systems are fully fledged epistemic agents. No matter whether they form beliefs or not, we have reason to think that they are fully-fledged epistemic agents, capable of acquiring knowledge, or the artificial version thereof.

§2. *Scepticism from safety*

Another attractive sceptical thought is that LLMs don't know (nor can they generate knowledge) because their methods of output formation are not safe. Many think that knowledge implies safety: if I know that p , then my belief is true in close worlds in which I base my belief by means of the same method as in the actual world.

Is the LLM output formation method safe? LLMs compress petabytes of text into weights. They do not store a database of facts; they store patterns, and match tokens based on statistical distribution, not empirical verification. In a nutshell, in the vast majority of cases – the good cases – Claude looks at a prompt and generates a highly probable sequence of words that makes for a true proposition. Is this method safe?

On the face of it is looking pretty good, one might think: the reasons why LLMs work to begin with is because the correlation between the most probable sequence of words (as per Claude's training dataset) and the state of the world is very robust: the world would have to be quite different from the way it is for the most probable token to be different. Crucially, that is not to say that there aren't problems: Highly

probable word sequences match the training dataset. If the training dataset contains biases, outdated information, or widespread human misconceptions, the most highly probable token will fail to be aligned with the world, even though it perfectly aligns with the model's internal math. Note, however, that this fails to create a problem for the claim that the correlation is robust: safety is not infallibility. Our belief formation processes too are vulnerable to biases, false assumptions, and so on, while remaining safe methods of belief formation.

But here is a sceptic's rejoinder: the safety claim focuses exclusively on upstream settings. However, Claude also comes with a downstream creativity setting – 'temperature' – that allows Claude to pick from a larger set (larger than one) of the most probable tokens. Temperature settings allow LLMs to sound more human-like, in that they get a degree of freedom of word choice. It's always a trade-off: too high a temperature setting, inaccuracy rates shoot through the roof. Too low, and Claude does not sound like a human interlocutor anymore.

Due to the temperature setting, using the exact same statistical method, the thought would go, Claude might confidently generate a plausible-sounding hallucination in a close world. The prompt is given, and Token A (true) has a 60% chance, while Token B (false) has a 40% chance. The random number generator picks Token A. You get a true statement. In a close possible world, everything is 100% identical—the prompt, the weights, the probabilities. The only difference is that the microsecond timing of the server results in a different random seed, causing the model to sample Token B. Because a literal coin toss can change the output from true to false under the exact same conditions, the output is unsafe.

We are not worried about all this at all. A few reasons for our optimism:

First, several people (one of us included) have offered convincing arguments that safety (in its current formulation) is not necessary for knowledge (Comesaña 2005, Kelp 2009, Neta and Rohrbaugh 2004). Consider for instance this counterexample: you are the proud owner of the world's most accurate clock. Despite being so very accurate, the clock might malfunction if a radioactive isotope decays in its vicinity. While you have taken every precaution for this not to happen, unbeknownst to you, today such an isotope was left in the vicinity of your clock. The isotope has a short half-life but, though this is by now improbable, hasn't decayed yet. When it decays, no matter the time, it will cause the clock to display 8:22. You look at the clock and form a belief that it's 8:22. The isotope still hasn't decayed and so the clock is working properly and displaying the time accurately. In this case, you know that it's 8:22 (you take a competent reading from a properly functioning and accurate clock). At the same time your belief isn't safe. It might easily have been that you arrived a minute earlier or later, and the isotope decayed in which case you would have had a false belief that it's 8:22 (Bogardus 2014, 299-303). A noteworthy feature of this case is that a central reason why your belief is unsafe is that it is random when the isotope will decay – just as in the temperature case it is random whether Token A or Taken B is picked. The fact that, in the clock case, you have knowledge even though you aren't safe from error provides us with reason to think that the same goes, *mutatis mutandis*, for temperature case.¹⁹

Second, it is less than clear why we should think that temperature settings bring a problematic level of luck into the process: recall that we said that the reason why LLMs work as well as they do to begin with is because the correlation between the most probable token and the state of the world is very robust: the world would have to be quite different from the way it is for these most probable tokens to be different. But surely this applies widely to the next 3 or 4 most probable tokens as well. Of course, the higher the temperature, the lower the safety of the method, since the closer the world where the correlation fails. But insofar as the temperature settings are optimised for as high as possible a level of accuracy, they are, *de facto*,

¹⁹ Cases of testimony also shed serious worries about the thought that safety is necessary for knowledge: Consider: as I walk down the street, I overhear you telling your friend that it's 2 o'clock, and I come to know that it's 2 o'clock. Had I walked slightly slower, I would have overheard someone else, who would have mistakenly said it was 1.30. But I didn't, and I'm fine.

also optimised for as high as possible level of safety – after all, both accuracy levels, and safety levels, depend on the strength of the correlation between plausible token and the world.

Understanding

Even if Claude has knowledge, we may wonder whether Claude can attain more sophisticated epistemic statuses, most notably understanding. As a first observation – once more, from behavioural data - note that we often say things like ‘Claude understands the geography of the Republic of France (RF) better than we do.’ That’s why, when you ask us, for instance, what the southernmost point of RF is, we recommend that you ask Claude. We are aware that RF has overseas territories south of Corsica, which is the southernmost point we can think of, but we don’t know more than that. Claude does. And the same goes for many other questions about the geography of RF. That’s why Claude understands the geography of RF better than we do. But of course, if we allow that Claude understands the geography of RF better than we do, since we have some positive degree of understanding of the geography of RF (better than someone who doesn’t even know that Corsica belongs to France), Claude must be capable of understanding the geography of RF also, at least to some degree.

One might worry that understanding geography doesn’t involve much in the way of explanation. Accordingly, one might think that even if Claude is capable of non-explanatory understanding, they may not be capable of explanatory understanding, such as understanding why or understanding of phenomena that involve explanations. Again, that can’t be right. Claude understands why Corsica isn’t the southernmost point of RF: because RF has all these overseas territories south of Corsica. In this way, Claude also has understanding of (the phenomenon of) the geography of RF that involves explanatory relations. What’s more, unlike us, Claude can name all the overseas territories of RF south of Corsica and place them on a map. As a result, their understanding of why Corsica isn’t the southernmost point of RF and their understanding of the geography of RF is broader and deeper than ours.

Let’s approach the question of understanding from a different angle. Most theories of understanding fall in one of two camps: there are knowledge-based theories, which analyse understanding in terms of knowledge (often knowledge of explanations (Hempel 1965, Khalifa 2017, Sliwa 2015)), and manipulationist theories, which analyse understanding in terms of the ability to manipulate representations (DeRegt 2017, Grimm 2016, Hills 2015).

It’s easy to see that simple knowledge-based accounts of understanding entail that LLMs have understanding, given our previous result that LLMs have knowledge. For instance, consider the most straightforward proposal: understanding why p just is knowledge why p (Sliwa 2015). On this view, Claude understands why Corsica isn’t the southernmost point of RF because Claude knows why Corsica isn’t the southernmost point of RF.

What’s more, LLMs can have understanding even on more sophisticated versions of knowledge-based accounts of understanding. On the view that one of us has defended elsewhere, what matters to understanding the geography of RF is not only knowledge about the geography of RF, but also the basing relations that obtain between these items of knowledge (Kelp 2015, 2021). Crucially, one of the central features that make LLMs into full epistemic agents is that they are capable of instantiating the basing relation. So, LLMs can have understanding even on more demanding versions of knowledge-based accounts of understanding.

What about manipulationist views? These views place emphasis on being able to make inferences, including about similar cases. Again, it is hard to deny that LLMs can do this, insofar as they can base. Claude has absolutely no problem in inferring what the southernmost point of RF would be if RF lost its current southernmost point. They also don’t have problem with many other relevant inferences such as what the size of its territory would be if it hadn’t lost Algeria, and even complex questions such as how many cities over 1m inhabitants there would be if the population of Paris were distributed equally among all cities in RF with more than 500k inhabitants, and so on.

Let's look at one specific, quite demanding account of understanding in the manipulationist camp, due to Alison Hills. Here it goes:

“[I]f you understand why p (and q is why p), then you believe that p and that q is why p and in the right sort of circumstances you can successfully: (i) follow some explanation of why p given by someone else; (ii) explain why p in your own words, (iii) draw the conclusion that p (or that probably p) from the information that q, (iv) draw the conclusion that p' (or that probably p') from the information that q' (where p' and q' are similar to but not identical to p and q); (v) given the information that p, give the right explanation, q; (vi) given the information that p', give the right explanation, q.” (Hills 2016, 663) Suppose, we were chatting to Claude about where to go on holiday. After considering a range of options, we conclude that we'd really like to see the Alhambra and the Acropolis and will have to decide which one we want to see more. We tell Claude that we eventually decided to go to Athens because we want to see the Acropolis more than the Alhambra. Assuming that Claude has beliefs (or that we replace belief by commitment in Hills's account), Claude now believes that we decided to go to Athens because we want to see the Acropolis more than we want to see the Alhambra. It is easy to see that Claude understands why we decided to go to Athens even by Hills' very demanding view. Claude did follow our explanation (and so can follow an explanation) of why we decided to go to Athens (i) and can explain why we decided to go to Athens in their own words (ii). Claude can draw the conclusion that we decided to go to Athens from the information that we wanted to see the Acropolis more than the Alhambra (iii) and, should we have told them that we decided to go to Granada because we preferred to see the Alhambra, they can also draw the conclusion that (p' =) we decided to go to Granada from the information (q' =) that we wanted to see the Alhambra more than the Acropolis (iv). Given the information that we decided to go to Athens, Claude can give the right explanation, i.e. that we wanted to see the Acropolis more than we wanted to see the Alhambra (v), and, finally, should we have given Claude the information (p') that we decided to go to Granada, Claude would have been able to give the right explanation, i.e. that we wanted to see the Alhambra more than we wanted to see the Acropolis. Even on a quite demanding manipulationist account, like Hills's, Claude comes out as understanding why we decided to go to Athens.

Besides the fact that we attribute understanding to LLMs, and do so with good reason, the case that LLMs have understanding is further supported by theories of understanding. Theories, and even quite demanding ones, on both sides of the central divide between knowledge-based accounts and manipulationist accounts entail that LLMs can have understanding, both of phenomena and understanding why. In this way, there is reason not only for optimism that LLMs can have knowledge, but also that they can have understanding.

Conclusion

We have explored multiple reasons for thinking that complex Large Language Models, such as Claude and Chat GPT, are genuine epistemic agents. We have argued that these AI systems' behaviour constitutes strong, undefeated evidence that AI systems are evidence responsive inquirers and informants, and that this provides us with excellent reason for thinking that they are epistemic agents. What's more, we have argued that LLMs have what it takes to be knowers and to have understanding. In this way, we think there is reason for optimism: LLMs are epistemic agents capable of achieving positive epistemic statuses, conducting successful and unsuccessful inquiries, and satisfying substantive epistemic norms. If we are right, this has significant consequences for industry: a significant research programme in AI aims to build truth-tracking AIs, and a lot of research resources are spent by both industry and academia on figuring out how to do so. If we are right, complex LLMs like Claude and ChatGPT are epistemically aligned: they already are truth-trackers.

Part §3. Should We Trust AIs?

Introduction

In the previous section, we have argued that AI systems such as Claude and Chat-GPT are genuine epistemic agents: they respond to evidence, they inquire, they assert, and can even have knowledge and understanding. In this section, we will focus on the question as to whether we can acquire positive epistemic statuses such as knowledge and justification from such AI systems. In other words, we will be focusing on the epistemology of AI testimony and the epistemology of human-AI interaction more generally.

In the epistemology of testimony, the most central division is between anti-reductionists and reductionists. Very roughly, reductionists hold that testimony yields justification only in the presence of independent reason for believing what you are told (e.g. Faulkner 2011, Fricker 1994). For instance, the simple fact that we told you that there is milk in the fridge isn't enough to yield testimonial justification. It is only if you have reason for thinking that we are telling the truth that you acquire testimonial justification for believing. In contrast, anti-reductionists deny that testimony yields justification only if there is independent reason for believing what you are told (e.g. Burge 1993, Goldberg 2010, Simion 2021a). For instance, when we tell you that there is milk in the fridge that may be enough for you to acquire justification for believing that there is milk in the fridge, provided that defeaters are absent. In any case, you don't need independent reason to think that we are telling you the truth.

Another key divide is between those who think that testimony can only transmit positive epistemic status and those who think that it can also generate it. According to the transmission view, a hearer can acquire testimonial justification for believing that *p* only from sources that have justification for *p* to give (e.g. Goldberg 2018, Wright 2018). For instance, when we tell you that there is milk in the fridge, you get testimonial justification for believing that there is only if we have justification for believing that *p*. If we don't have justification to give, there is no justification for you to receive. In contrast, the generation view allows for testimony to generate justification for believing (e.g. Graham 2000, Lackey 2008). When we tell you that there is milk in the fridge, you may acquire justification for believing that there is milk in the fridge even if we don't know this/have no justification for believing it to be true, perhaps because we are engaging in wishful thinking or are simply lying about the presence of milk in our fridge.

The most liberal view in the epistemology of testimony is the view that a testimony is a source of justification by default:

The Default View. If a speaker, *S*, testifies that *p*, then the hearer, *H*, has *prima facie* testimonial justification for believing that *p* (in virtue of *S*'s testimony that *p*). (e.g. Burge 1993, Simion 2021a)

The Default View sides with both anti-reductionism and the generation view of testimony. This is because, pace reductionism, it doesn't require hearers to have independent reason for believing the testimony to be true, and, pace the transmission view, it doesn't matter whether the source has justification for the content of the testimony either.²⁰

We like the Default View, and we have defended it at length in the past²¹. Could the Default View be true of AI testimony? In what follows, we will argue that it is. To this effect, we will do two main things: first, we show that the main motivations that support the Default View for human testimony hold *mutatis*

²⁰ In her most recent book (2026), Annalisa Coliva argues for a hinge-epistemological version of the Default View: that routine reliance on AI is best understood as "hinge trust"—a non-moral, functional stance—rather than evidence-based or moral-psychological trust. Coliva frames this "e-trust" as a necessary precondition for using AI, distinguishing it from the evaluation of an AI's actual trustworthiness, which requires empirical assessment.

²¹ E.g. Kelp & Simion 2021, Simion 2021a.

mutandis for AI testimony as well. If so, the default view for AI testimony and the default view for human testimony stand and fall together. Second, we look at some relevant literature in the philosophy of AI that supports a negative answer, and we show that these arguments remain ultimately unsatisfactory.

3.1 The Default View

Is the Default View plausible for AI testimony? As a first observation, note that if our argument from the previous sections works and AI systems can be and often are genuine epistemic agents, capable of having a range of positive epistemic statuses including knowledge and justification, it should be much less worrisome to think that we can acquire justification from AI testimony in the way that the Default View would have us think we do.

One attractive feature of the Default View is that it accommodates the idea that testimonial justification can flow along chains of testimony of varying sizes. However, if only human testimony yields justification by default, whereas AI testimony doesn't, this explanation will no longer work. This is worrying especially as in our social epistemic lives human and AI testimony become increasingly intertwined. What's more, suppose you learn that a testimonial chain as a result of which you believe that *p* involves AI testimony. If human but not AI testimony yields justification by default, this should be epistemically significant information. In particular, you have at least some defeat that you wouldn't have had had you learned that the chain consisted only of human testifiers. Of course, the defeat you acquire may be quite minimal: while it may lower the degree of testimonial justification you end up with, you may still have enough justification to believe outright what you are told. But, of course, it is now easy to imagine a situation in which our social epistemic lives are sufficiently intertwined with AI testimony for it to be common and commonly known that multiple links in testimonial chains are AI testifiers, such that defeat agglomerates accordingly. If we thought that human but not AI testimony yielded justification by default, the result threatens to be a generalised scepticism about testimonial justification in a world like this. After all, we now have a standing defeater for all instances of testimony, no matter whether its immediate source is human or an AI system.

In good news, there is excellent reason to believe that, if human testimony yields justification by default, the same is true of AI testimony: the main motivations that support the Default View for human testimony hold *mutatis mutandis* for AI testimony as well. Here is why:

According to Tyler Burge (1993, 2013), the Default View holds on purely *a priori* grounds. In a nutshell, here is the view: Burge thinks that intelligible propositional expressions indicate a source endowed with rational abilities. If you hear an intelligible message coming your way, it is reasonable to believe that it was a rational source that produced it, rather than, say, that it was randomly generated by some plants in your garden when hit by a gust of wind. You don't need to first gather further positive reasons to attribute rationality to the generating source: that is the default, although defeasible assumption. Since intelligible propositional expressions presuppose rational abilities, intelligible presentations-as-true come *prima facie* backed by a rational source. Now, according to Burge, in virtue of having the function of generating true content, rational sources will reliably do so in normal conditions. If that is the case, and if intelligible propositional presentations-as-true indicate a rational source, it follows that intelligible presentations-as-true indicate a *prima facie* source of truth. In virtue of this, according to Burge, we are *a priori prima facie* entitled to take intelligible affirmation at face value (Burge 1993, 472).

Note that nothing in Burge's argument requires testimony to be human testimony for the Default View to hold: what is essential to default entitlement, on this account, is that that intelligible propositional expressions indicate a source endowed with rational abilities. Since LLMs undoubtedly produce intelligible propositional expressions, we get default entitlement to believe based on LLM testimony.

But isn't it true that our knowledge of what actually goes on under the hood may defeat this initial entitlement? Two things about this: first, even if defeat were to occur, it would not speak against the Default

View: after all, the Default View is a view about *prima facie* justification, and thus comes with an anti-defeat proviso: one is only justified by default to form beliefs based on testimony absent defeat. Second, Section 2 has argued extensively against the defeat claim: knowledge (or lack thereof) of what is going on under the hood of Claude tells us precious little about Claude’s epistemic status.

Say that you find yourself unconvinced by Burge’s claim to an a priori derivation of the Default View. Perhaps you think the account is, in a sense, reminiscent of Turing’s (1950) test, in that it takes behavioural evidence a bit too seriously for epistemic entitlement. Or perhaps you (like us) think that functions simply don’t guarantee reliability as Burge’s argument needs it (Simion 2021a). Not to worry: social versions of anti-reductionism will secure the same result. In past work, one of us has put forth the following social-norms-based defence of the Default View. Here it goes, in a nutshell: assertion is governed by an epistemic norm. Some epistemic norms are also social norms; plausibly, norms governing speech acts are among the main candidates for this double status. Whatever their preferred norm of assertion might be, most people in the debate agree upon one minimal, broadly Gricean thought: since it should be able to account for assertion’s epistemic function, it features a positive epistemic status – let’s call it +ES, and the corresponding norm of assertion +ESNA. Given that +ESNA is also a social norm, all else absent, in virtue of the social contract we are part of, conforming to it enjoys default rationality for the speakers. To see this, note that breaking a norm changes the normative landscape for the negative. If you don’t stand anything to gain from it, you should avoid it. All else absent, if there’s a norm that you should *phi*, *phi*-ing enjoys default rationality for you.

In turn, the normative default for speakers has direct impact on the normative default for hearers: since stopping at the red light is the default for drivers, entitlement to cross the street on a green light is the default for pedestrians. Similarly, if conforming to +ESNA enjoys default rationality for the speakers, hearers are by default entitled to believe what they are being told.

Will this argument translate nicely to AIs? We think that it does. LLMs inhabit our norm-governed epistemic community just as much as human testifiers do: we have already argued at length that their behaviour is governed by norms of proper function dropping out of both their design functions and their etiological functions. If so, no-epistemic-norm-violation is their default, all else absent, and entitlement to believe based on their outputs is ours, all else absent.

If so, our work here is done for now: both main motivations in the literature in favour of the Default View for human testimony hold *mutatis mutandis* for AI testimony as well.

In what follows, we look at some relevant literature in the philosophy of AI that foes against the Default View, and we show that these arguments remain ultimately unsatisfactory.

3.2 Transparency and the Explainability View

There has been some discussion of testimonial justification in the literature on AI, and we will get to it in due course. However, this literature is small in comparison with the literature on trust and trustworthiness in AI. Crucially, the literature on AI trust and trustworthiness has bearing on our question about the epistemology of AI testimony. To see this, consider the following highly plausible claim:

Justification-Rational Trust. H has *prima facie* testimonial justification for believing that p in virtue of S’s testimony that p iff it is *prima facie* epistemically rational for H to trust S’s testimony that p to be true.

And recall the Default View:

The Default View. If a speaker, S, testifies that p, then the hearer, H, has *prima facie* testimonial justification for believing that p (in virtue of S’s testimony that p).

It is easy to see that the Default View and Justification-Rational Trust together entail:

Testimony-Rational Trust. If S testifies that p, then it is *prima facie* epistemically rational for H to trust S's testimony that p to be true.

Note, importantly, that all principles above are indexed to *epistemically* rational trust. The reason for this is simply that practical considerations can make trust too easy (epistemically): if I know that whatever happens (whether you lie or not) the practical consequences are good for me, it might be practically rational for me to trust your testimony, in spite of heavy defeat. Epistemic rationality, however, will not follow suit.

Is Testimony-Rational Trust plausible for AI? One reason one might think the answer to this question is 'no' involves the *black box problem*. The black box problem, in short, is that AI systems are often able to generate a verdict but unable to generate an explanation of how the system came to that verdict. The black box problem is nicely illustrated in the case of Lucie from Cappelen and Dever (2021): Lucie is confused after her request for a loan is denied due to the AI system ('SmartCredit') determining that she has a very low credit score. Lucie asks the bank to explain *why* she has been assigned a low credit score. They can't give an answer: SmartCredit makes reliable predictions (we are to suppose) about who is likely to pay back their loans, but not even its creators can look under the hood and understand why Lucie's score, in particular, is low. Lucie pushes back: Perhaps she [Lucie] can at least be told what it was in her history that SmartCredit found objectionable. Was it her low annual income that was responsible? Was it those late credit card payments in her early twenties? Or was it the fact that she follows a number of fans of French film on Twitter...

Unfortunately, the bank doesn't have much helpful to say about this, either. It's easy enough to spot particular variables in the initial data set - the bank can show her where in the input her annual income is, and where her credit card payment history is, and where her Twitter follows are. But they don't have much to say about how SmartCredit then assesses these different factors. All they can do is point again to the cascading sequence of calculations - there are the initial numbers, and then there are millions upon millions of mathematical operations on those initial numbers, eventually dropping out a final output number. The bank explains that that huge sequence of mathematical operations is just too long and complicated to be humanly understood—there's just no point in trying to follow the details of what's going on. No one could hold all of those numbers in their head, and even if they could, it's not clear that doing so would lead to any real insight into what features of the case led to the final credit score. (Cappelen and Dever 2021: 9–10)

Lucie's frustration is understandable. Likewise, it's easy to imagine her (or anyone else's) inability to understand why SmartCredit assigned her a low credit score as partly responsible for her felt difficulty in *trusting* SmartCredit. As von Eschenbach (2021) puts things:

Trust requires that we have reason to believe both that AI is reliable and acting in our behalf, and so transparency into how AI operates and reaches its outcomes and predictions is needed in order for us to be able to judge AI to be trustworthy (2021: 1620).

And indeed, a growing number of theorists have been calling for the development of *explainable* AI (XAI), explicitly as a means to building trust in AI. The underlying assumption of this major research programme is that, if we want people to trust (predictive) AI, the AI should be transparent: we should be able, at least in general, to understand why the AI makes the predictions it does.

Cases like Lucie's motivate the following uniformity assumption for trusting AI:

Uniformity Assumption: User rational trust in AI requires user understanding. And user understanding requires an explanation of why the AI has produced the output it has.

In turn, the Uniformity Assumption motivates the following very popular view on what user rational trust in AI requires:

The Explanation View of AI Trust: Prima facie rational trust by users in outputs of AI systems requires an explanation of why the AI produced the output it has.

Now, if the Explanation View of AI Trust is correct, it is hard to see how Testimony-Rational Trust could be true. After all, Testimony-Rational Trust entails that if an AI tells you that *p*, then it is prima facie rational for you to trust the testimony, no matter whether there is an explanation of why the AI testified that *p*. or not In contrast, the Explanation View of AI Trust entails that if there is no explanation of why the AI testified that *p*, it is not rational for you to trust the AI's testimony that *p*. Since Testimony-Rational Trust is entailed by Justification-Rational Trust and the Default View, The Explanation View of AI Trust is incompatible with the conjunction of these two claims. Since it is hard to see how Justification-Rational Trust could be at fault here, there is reason to think that if The Explanation View of AI is true, then the Default View is false.

We may be inclined to put this down as a first substantive result in the epistemology of AI testimony. However, this would be premature. This is because there is reason to think that the Uniformity Assumption and the Explanation View of AI are not fully satisfactory at the end of the day. We are sceptical that there is any deep connection between trust and explainability. Consider an ordinary case of testimony:

A: Which way is the nearest subway stop?

B: Take a left at Great Western.

A: Ok, thanks!

At no point does B ask A for an explanation of either why or how they know that the nearest subway is to the left on Great Western. Nevertheless, it seems perfectly appropriate (in normal circumstances) for A to *trust* B that the nearest subway is where they say it is—to take them at their word. It's a perfectly ordinary case of trust, and explanation just doesn't come into it. Furthermore, all theories of testimonial justification give (or strive to give) this prediction.

Of course, this exchange is not very much like testimony from AI. A can easily imagine how B could in principle explain why [they believe that] the subway is to the left on Great Western ('I saw it on a map,' 'I pass it every day on my way to work'). The modest point here is that it's not as though actual explanation needs to be a part of trust in general. In fact, we often *don't* ask for explanation precisely *because* we trust someone enough to take them at their word. If we say that we trust our children to walk to school by themselves, but in fact we constantly track their phones every time they leave the front door, it's not clear that we *really* trust them to walk to school by themselves: a robust enough trust wouldn't require independent verification.²² Similarly, if we say that we trust someone to tell us the truth, but in fact we always check what they say by seeking an explanation of why, it's not clear that we *really* trust them to tell us the truth: a robust enough trust wouldn't require independent verification.

One might object that, while it's true that there is no *actual* explanation why offered in the example of trust above, B's assertion is explainable in principle. What's distinctive of AI testimony in cases like Lucie's credit score is that the answer is not even explainable in principle.

Nevertheless, one can trust testimony even when explanation is not even available in principle to one due to the e.g. complexity of the issue. This often happens when we trust expert testimony. Maybe we will never be able to understand quantum mechanics. For us (given our training), the field of quantum mechanics is effectively a black box. Our situation with respect to quantum mechanics is roughly the same as Lucie's with respect to SmartCredit. We can see that physicists put some values into some equations that

²² E.g. Hieronymi 2008b.

generate certain outputs, but we can't follow the calculations to their end, nor do we really understand why physicists treat the input values the way that they do. Even so, we can rationally trust scientific experts when they make a prediction on the basis of quantum mechanics.

One might pursue the above line of objection further: perhaps we are not able to understand quantum mechanics. However, some human beings are able to understand it. In contrast, in the black box AI cases we have been considering, no human is able to understand how the relevant outputs are arrived at. And, the objector may ask, isn't that a crucial difference between the two cases, one that allows us to say that we can rationally trust expert testimony but not black box AI testimony?

Two things about this: first, note that our brain and cognitive processes are no less a black box than AIs. Indeed, it is not clear at all why we should think that fully understanding our cognitive processes is more within our reach than understanding AIs. If so, this argument threatens to overgeneralise to scepticism about testimonial justification *tout court*.

Second, there are clear cases of AI testimony such that the explanation for the testimony is not understandable by humans but, very plausibly, we can rationally trust the AI testimony. Consider the Google DeepMind's Alpha Zero. Now Alpha Zero is famous for achieving true mastery a range of board games including chess and Go (Silver et al. 2017). Alpha Zero is stronger than any human chess player and any other chess engine at the time of its release. When Alpha Zero tells you that a certain move is an excellent move, it is hard to deny that it is rational for us to trust Alpha Zero's testimony. Crucially, however, Alpha Zero is a black-box AI in the relevant sense (Carter 2024b). While Alpha Zero uses Monte Carlo Tree Search, an interpretable search algorithm, which allows us to see the trees of positions it is exploring, and while we can also see things such as which moves it is considering and how it evaluates positions numerically, there is no explanation of it that is understandable by humans. What we may get is that a certain move is an excellent move because it is assigned a certain numerical score by Alpha Zero. But that does not allow us humans to understand why the move is an excellent move, not if we don't understand why Alpha Zero assigns scores in the way it does. And this is precisely what we don't get. In this way, there is reason to believe that Alpha Zero is black-boxy in just the same way as Smart Credit.²³

Given that this is so, it is easy enough to see that the Explanation View of AI Trust entails that it would not be rational for us to trust the testimony of Alpha Zero. For instance, if Alpha Zero were to tell us that a certain move is an excellent move, it would not be rational for us to trust its testimony; alternatively, we would not acquire testimonial justification for believing that the move is indeed an excellent one. That strikes us as manifestly mistaken. Of course it is rational to trust the testimony of the strongest chess player in the world when they offer testimony on whether a certain move is excellent! Indeed, the greatest human players (players whom it is rational for us to trust) themselves give serious weight to what AlphaZero says, even when they themselves don't understand it. If anything, we submit, it would be irrational not to trust the testimony of Alpha Zero here. The fact that we cannot understand how Alpha Zero arrived at its claim is of little consequence.

Before moving on, we'd like to make one important observation. We don't mean to suggest that explainability isn't valuable in AI. In fact, we are happy to grant that explainability is preferable all else equal. Rather, our point here is merely that, pace the Explanation View of AI Trust, it is not required for rational trust in the outputs of AI and that, as a result, we cannot derive an explanation condition on AI testimonial justification from the Explanation View of AI Trust.

3.3 Computational Reliabilism

²³ What's more, the situation wouldn't be any different if Alpha Zero were fully opaque and didn't use an interpretable search algorithm. Even so, it would undeniably be rational for us to trust Alpha Zero's testimony that a certain move is a good move.

Let's move on to the most prominent treatment of justification in the AI literature. The central view here is computational reliabilism. Champions of computational reliabilism are also concerned with conditions under which we can trust AI systems, including black-box ones. The central idea is that we can trust the outputs of AI systems if we have reason for thinking that they are reliable. And reason for thinking that they are reliable is unpacked as a range of reliability indicators which essentially feature in the design and maintenance of the algorithm at issue in the AI system. Here is a recent statement of the view:

Computational Reliabilism (CR). A human agent S is justified in believing the algorithm's output o if and only if o was rendered by a reliable algorithm. A reliable algorithm is one that produces true outputs o most of the time. To this end, the algorithm must have been specified, coded, and maintained through diverse reliability indicators. (Durán 2026, 59-60)

CR is a version of reliabilism. Durán and collaborators acknowledge process reliabilism as an inspiration for their view. Process reliabilism is, roughly, the view that one's belief that p is justified if and only if it is produced by a reliable belief forming process. For a belief forming process to be reliable is for it to tend to produce true beliefs (e.g. Goldman 1979, Kornblith 2002). It is easy to see a kinship between CR and process reliabilism.

Before moving on, it may be worth noting that process reliabilism, as characterised above, is a view of *prima facie* justification. Process reliabilists acknowledge the possibility of defeat of justification and have offered different accounts of this phenomenon (e.g. Goldman 1979, Graham and Lyons 2021). Likewise, we take it that CR is an account of *prima facie* justification. After all, there is every reason to think that the justification afforded by the outputs of even the most reliable algorithms is defeasible. Accordingly, going forward, we will take CR to be an account of *prima facie* justification.

Despite the kinship between process reliabilism and CR, there are differences. The perhaps important one, for our purposes, surfaces at the end of the official definition of CR, when Durán claims that '[t]o this end, the algorithm must have been specified, coded, and maintained through diverse reliability indicators.' It is not entirely obvious what the status of this claim is in the definition. Fortunately, this unclarity is dispelled later on when Durán clearly states that: "To be reliable here means that the algorithm is specified, coded, used, and maintained utilizing specific, tailor-made reliability indicators designed for the purpose at hand" (Durán 2026, 74).

Standard versions of reliabilism analyse reliability only in terms of a tendency to produce true beliefs. In contrast, CR imposes these additional conditions. To count as reliable, an algorithm must be designed and maintained via reliability indicators. Examples of reliability indicators are verification, validation, robustness analysis, and expert knowledge, among others. The idea here is that these reliability indicators provide reason for trusting AI systems. Crucially, they can do so without appeal to explainability. In fact, one of the key selling points of CR is that it explains how we can rationally trust AI systems even when they are genuine black-box AI systems that remain forever opaque to human cognisers. The vast majority of AI systems we use, including black-box ones, are designed and maintained via reliability indicators. So, we can rationally trust them. In this way, CR scores points against the rival Explanation View of AI Trust.

At the same time, it is easy to see that CR is incompatible with the Default View. Suppose an AI system that was not specified, coded, used and maintained using reliability indicators produces a piece of testimony. In that case, CR entails that you do not have *prima facie* justification for believing what the AI system told you. In contrast, the Default View entails that you do have *prima facie* justification for believing what the AI system told you. Thus, the two views are incompatible.

Once more, the Default View comes under pressure, perhaps more so than before. After all, CR promises to be a rendering of an attractive general view of justification, to wit, process reliabilism. At the same time, CR doesn't require explainability to allow for rational trust in AI systems, and it can allow for black-box AI systems to be sources of justification. In this way, CR improves on the Explainability View

of AI Trust in an important respect. In light of this, we may think that we now have very strong reason to worry that the Default View cannot be correct.

While we agree that CR improves on the Explainability View of AI Trust, especially when it comes to its predictions about black-box AI, there is reason to think that CR isn't satisfactory at the end of the day. To see this, consider a case of an AI system that satisfies the conditions of CR and that you have used many times to acquire testimonial justification in the past. Today, the system is hacked, and the hackers are attempting to sabotage the system by deleting the algorithm and replacing it by random strings of symbols. Unbeknownst to everyone, thanks to an incredible stroke of luck, the random string of symbols constitutes an algorithm that, while different from the original one, does the job of the original one just as well. You are fully unaware of all this. At some point after the hack is completed, you ask the algorithm a question, and it provides you with an answer. In this case, intuitively, you acquire justification for believing the answer the AI system provided after the hack, just as you did before the hack. CR cannot accommodate these observations. In fact, CR entails that there is an important epistemic difference between the epistemic status of the AI system before and after the hack. Before the hack, testimony by the AI system does provide you with justification for believing what you are told, whereas after the hack it doesn't.

Here is another problem for CR. Suppose researchers find a monkey at their computers looking at a screen on which a computer is playing itself in chess. Video surveillance indicates that the monkey was typing randomly on the keyboard. However, as luck would have it, the result was an algorithm similar to Alpha Zero. With astonishment, the researchers discover the AI system created is incredibly good at playing chess. Matches against the best human players and chess engines are arranged, and it comes to light that the AI system wins every single game against them. The AI system can also evaluate moves and categorise them as excellent, good, inaccurate, mistake, and blunder. Suppose the AI system tells you that a certain move is excellent. In this case, you acquire justification for believing that it is excellent. However, once again, CR cannot accommodate this observation. Since the underlying algorithm wasn't specified, coded, used, and maintained via reliability indicators, CR entails that the algorithm isn't reliable and that its outputs do not provide justification for humans.

Once more, we'd like to emphasise that we don't mean to suggest that reliability indicators aren't valuable in AI. In fact, we are happy to grant specifying, coding, using, and maintaining algorithms through reliability indicators is preferable all else equal. Rather, our point here is merely that, pace CR, it is not *required* for the acquisition of justification from AI outputs.

3.4 Back to the Default View

It comes to light that the Default View isn't all that easy to take down. In particular, two of the most relevant rival views, the Explainability View of AI Trust and Computational Reliabilism, turn out to be too demanding to offer a viable way forward in the epistemology of AI testimony.

What about the considerations that motivated researchers to adopt views that are incompatible with the Default View? In particular, what about the black box problem? If we adopt the Default View and we grant Justification-Rational Trust, we do get the result that it is *prima facie* rational to trust what we are told. But is that really plausible if the testifiers are black box AI system? For instance, is it really plausible to think that Lucie should trust SmartCredit's testimony that she shouldn't be given the loan? That might seem implausible. But if so, doesn't this mean that the Default View is in trouble after all?

Fortunately, the answer to this question is 'no'. The Default View does indeed entail that SmartCredit's testimony provides *prima facie* justification for believing that she shouldn't be given the loan. What's more, the testimony makes it *prima facie* rational for her to respond with trust. Crucially, though, *prima facie* justification is defeasible justification. Lucie may have justification for believing a vast range of things that have evidential bearing on SmartCredit's claim or the trustworthiness of SmartCredit as a source of information. For instance, Lucie may know that she has been a good actor financially, that many of her

friends with similar profiles got loans similar to the one she requested etc. In that case, Lucie has a range of defeaters for SmartCredit's claim that she shouldn't be given the loan. As a result, SmartCredit's testimony that Lucie shouldn't be given the loan is defeated for her: while the testimony provides *prima facie* justification, *ultima facie* Lucie may not be justified in believing what she is told. In fact, the mere fact that she is declined an explanation may generate a defeater for Lucie in this particular context. She might justifiably find it suspicious that no explanation can be provided for a decision that affects her life as substantially as this one. She may justifiably wonder how it can even be that a decision like this can be made without any explanation backing it and whether the entire process is above board. All of these considerations may serve to provide defeat for the trustworthiness of SmartCredit's testimony. If so, the Default View will predict that Lucie does not acquire *ultima facie* justification for believing that she has a low credit score.

Now, the take home lesson here is not that there are always defeaters for AI testimony. Rather, our point is that it is most natural to think that, in this particular context, Lucie has defeaters for SmartCredit's testimony with the result that Lucie entirely rational in not trusting the AI testimony in this case.

Note also that defeat is not the only resource champions of the Default View have at their disposal to deal with cases in which it may seem rational not to trust the testimony of an AI system. To see this, consider another kind of example that has received a considerable amount of attention in the literature: black-box AI systems in medical practice. Black-box AI systems are very reliable in diagnosing certain conditions, such as skin cancer, and calculating doses for treatment. Consider a black box medical AI is designed to be able to give its diagnosis in the form of testimony. Even so, we might have reservations in trusting the testimony in these cases when they come from black-box AI systems. Crucially, note that in these cases, the stakes are usually extremely high. What's more, what's at issue here is not only forming a belief in the testimony of a black-box AI but also acting on it. While the conditions for rational trust for believing and rational trust for acting may coincide in many cases, they can come apart. In fact, one paradigm type of case in which they do come apart is when the stakes for acting are high. If your friend tells you that they have milk in the fridge, you do have justification for believing what you are told. It is rational for you to trust their testimony. But that doesn't mean that it is rational for you to accept a bet on it that offers a return of £1 if there is milk in the fridge and will cost you your life if there isn't. Crucially, the fact that it isn't rational for you to trust the testimony of your friend in contexts in which acting on it comes with so high a risk doesn't mean that your friend's testimony doesn't provide you with justification for believing what they told you. Rather, what is going on here is that the question as to whether testimony gives you justification for believing what you are told comes apart from the question as to whether it is rational for you to act on what you are told (Simion 2021b). The situation is similar in the case of the medical black-box AI that we are considering, at least in the sense that the question as to whether black-box AI testimony provides doctors, say, with justification for believing what they are told comes apart from the question as to whether it is rational for doctors to act on the testimony when a life is at stake and when they, the doctors, are ultimately responsible for the course of action they choose.

3.4 Rational Trust and Trustworthiness

The epistemology of AI testimony is a key topic in the epistemology of AI. Even so, questions about testimony are not the only questions that are epistemically relevant in our interactions with AI systems. AI systems could be more directly involved in our cognition, for instance via brain computer interfaces. What's more, AI systems may deliver outputs that are epistemically assessable but don't have propositional form. For instance, AI systems may output probability distributions, maps, images, and the like. Here, too, the question as to when it is rational to trust AI systems arises. For that reason, we will now address the question under which conditions it is rational for us to trust AI systems more generally.

Recall that we already looked at one prominent answer to this question, in our discussion of AI testimony: the Explanation View of AI Trust. According to this view, *prima facie* rational trust in the outputs of AI systems requires an explanation of why the AI produced the output it has. We have already seen why this view remains ultimately unsatisfactory: rational trust doesn't in general require explanation in the way envisaged. What's more, we have even seen cases of AI systems that could be rationally trusted even though no explanation of the sort required by the Explanation View of AI Trust is forthcoming. How can we improve on this view?

In our quest to understand when it is rational to trust AI systems, we suggest that, instead of focusing on whether AI is explainable, we focus on whether AI is *trustworthy* (Simion and Willard-Kyle Forthcoming).

When we place our trust in someone, they can either come through, or they can fail to come through. When they do come through, things go well, and when they don't, they don't. Crucially, the question as to whether it is rational for us to trust someone isn't answered by whether they come through. Suppose you trust a friend whom you know to be extremely untrustworthy to bring desert to your dinner party. In that case, even if they happen to bring desert on this occasion, it was not rational for you to trust them to do so.

In light of this, we might think that what matters to whether it is rational to trust someone is not whether they will do a certain thing – e.g. bring desert –, but whether you have justification for believing that they will do it. However, having justification for believing that someone will do something and rationally trusting you to do it come apart. We can justifiably believe that you will do something merely in virtue of knowing that we will force you to do it. For instance, we may force you to bring desert to the dinner party at gunpoint. While we have justification for believing that you will bring desert to the dinner party, this has little to with it being rational for us to trust you to bring desert; it is certainly not what makes it rational for us to trust you to bring desert.

Perhaps, then, what matters to rational trust isn't understood in terms of whether the entrusted party comes through, but rather in terms of whether they are trustworthy. Even that isn't enough. Your roofer friend may be extremely trustworthy when it tasked with preparing Croquembouche (an extremely difficult-to-make type of desert and your favourite). Even so, since you have no reason at all to believe that your friend is thus trustworthy, it would be rational for you not to trust them to prepare Croquembouche for the dinner party, which is why you don't ask them for it and request simple lemon cake instead.

While we agree that trustworthiness matters to understanding rational trust, we want to suggest that the simple fact of the matter as to whether you are trustworthy isn't quite it. Rather, we want to suggest that what is key is evidence of trustworthiness. We suggest the following principle:

Rational Trust: It is *prima facie* epistemically rational for S to trust R to φ iff S does not have *prima facie* justification for believing that R is not trustworthy with respect to φ -ing.

Rational Trust is, of course, a strong anti-reductionism about rational trust: absent defeat (evidence of lack of trustworthiness), it is rational to trust. Once more, note, importantly, that Rational Trust is restricted to *epistemically* rational trust. The reason for this, again, is simply that practical considerations can make trust too easy (epistemically): if I know that whatever happens (whether you ϕ or not) the practical consequences are good for me, it might be practically rational for me to trust you to ϕ , in spite of heavy defeat. Epistemic rationality, however, will not follow suit.

Here's one way that we think focusing on trustworthiness helps in the case of AI. Being trustworthy is widely taken to involve more than mere reliability.²⁴ We think this gap between reliability and trustworthiness is an important place to look when trying to understand failures to trust AI. Sophisticated AI can be very reliable—but that, by itself, isn't enough to make AI trustworthy, if one has evidence of lack of trustworthiness. Or, put another way, even if Lucie *knows* SmartCredit is super reliable, that won't settle

²⁴ Hawley 2019.

the question of whether it's rational for her to *trust* SmartCredit, since trustworthiness and reliability come apart.

That isn't to say reliability doesn't matter for trust. It does. But the specialist literature is in broad agreement (e.g. Baier 1896, Hawley 2019, Potter 2002) that trustworthiness requires some extra condition X.

But introducing condition X creates an immediate problem for the very concept of trustworthy AI. Extant accounts of condition X are very often anthropocentric - they are categories that don't seem to apply to (a wide range of) AI. Insofar as one of those theories is right, it just *can't* be rational to trust AI. AI wouldn't be the kind of thing that can be rationally entrusted. Call this the *anthropocentric challenge* for trusting AI.

We think the anthropocentric challenge can be overcome. The right account of trustworthiness is *not* anthropocentric and can countenance trustworthy AI (and trustworthy artefacts in general). Indeed, we think this is one of the virtues of our preferred account. But to fully appreciate the anthropocentric challenge, we begin with some competitors to our account, on which it is difficult to see how AI could be trustworthy.

First, consider Annette Baier's (1986) influential account of trustworthiness. On her view, trustworthiness is reliability due to *good will*: the trustworthy person not only reliably fulfils their commitments but does so in virtue of their goodwill toward the trustor.

This account makes it difficult to see how AI could be trustworthy. SmartCredit doesn't have goodwill toward Lucie, for instance. SmartCredit is completely indifferent toward Lucie. Perhaps a *very* sophisticated AI could have goodwill toward Lucie – it might develop the right sort of affective capacities to host a good or (in the latter case) bad will toward people. But most of our current AI is, we presume, not like that. So, on Baier's account, most AIs are not candidates for trustworthiness.

Crucially, we have argued in previous work (Kelp and Simion 2023) there are problems for Baier's account even when setting the anthropocentric challenge aside. Suppose that George has promised Ann that he will wash the dishes every weeknight. He washes the dishes out of a sense of duty, but he hates doing so and comes to deeply resent Ann for (as he sees it) extracting this promise from him. George does not have goodwill toward Ann. His only feeling for Ann is deep resentment. Nevertheless, given that he always *does* wash the dishes just like he promised, and because he promised, it seems that he *is* trustworthy with respect to washing the dishes. Ann could justly criticize George for his unjust sense of resentment, but she can't fault him for being untrustworthy with respect to washing the dishes. And, indeed, it's rational for Ann to continue to trust George to wash the dishes, despite his antagonism.

Substituting a good will for condition X, as Baier does, makes trustworthy AI impossible. But a good will is the wrong substitute anyway. It's time to turn to other accounts.

According to Nancy Potter (2002: 25, emphases ours), being trustworthy amounts to being 'one who can be counted on, *as a matter of the sort of person he or she is*, to take care of those things that others entrust to one'. Being the right sort of person involves having a good or virtuous character. A virtuous *character* is a robust thing. It's not clear at all that (most contemporary) AI systems like SmartCredit have anything like virtuous characters, or any characters whatsoever. Accordingly, on Potter's account, AI systems are not candidates for trustworthiness.

But once again, there are independent reasons to be sceptical of Potter's proposal. Recall Ann and George. It's not clear that George has a good character - he seems like a jerk. But in the original version, we could at least say this for him: he washed the dishes out of a sense of duty. And perhaps that's enough to attribute to him a kind of (thin) good character, however unfair his attitude toward Ann. But we can tweak the example further to ensure he does not meet Potter's condition. Imagine that George doesn't do the dishes out of a sense of duty at all, but out of a sense of fear. He believes that God will punish him if he doesn't fulfil his promises to Ann; otherwise, he wouldn't wash the dishes. Even so, if George really does reliably and stably fulfil his promise to Ann to wash the dishes, it seems that he *is* trustworthy with respect to washing the dishes. Ann could justly criticize George for his bad character, but she can't fault

him for being untrustworthy with respect to washing the dishes. And, indeed, it's rational for Ann to continue to trust George to wash the dishes, despite his flawed character.

Focusing on the will or character of the entrusted has led us into problems. That suggests a more externalist approach. We think Hawley's (2019) work is near the mark. According to Hawley, trustworthiness is simply reliability in avoiding unfulfilled commitments: this requires both caution in incurring new commitments and diligence in fulfilling existing commitments.

Can AI be trustworthy in Hawley's sense? It seems so. After all, AI can avoid unfulfilled commitments. In fact, if anything, one might worry that Hawley's view makes it rather *too* easy for AI to be trustworthy. That's because Hawley's account of trustworthiness is *negative*, and trust is achieved by being reliable at *not* violating commitments. Unsurprisingly, the negative character of Hawley's proposal is its most controversial element. Suppose George meticulously avoids ever making any promises or commitments to Ann, even implicitly, even though Ann is his friend and in urgent need of help. Is this, by itself, enough to make George a trustworthy friend to Ann? Not obviously. Suppose, alternatively, that George only commits to bad things – to lie to Ann all the time, and never help her when she needs it. Does this make George more trustworthy? Again, we don't think so.

For this reason, we favour (and have defended)²⁵ a more positive externalist account of trustworthiness. We now turn to our preferred approach.

3.5 Trustworthy AI: An Ought-Based Account

Several of the definitions of trustworthiness in the last section made 'trustworthy AI' something of a category mistake. AIs *can't* be trustworthy because they (typically) don't have the right kind of inner lives or moral characters. Methodologically, we prefer to go the other direction. We think it's intuitive that AI *can* be trustworthy. In this way, we agree with, e.g., Cappelen and Dever (2021) that accounts of trustworthiness ought to be de-anthropocentrized.

On our preferred account (Kelp and Simion 2023), then, here is what it is to be maximally trustworthy:

Maximal Trustworthiness One is maximally trustworthy with regard to φ -ing if and only if one has a maximally strong disposition to fulfil one's obligations to φ .

One important clarification: in some natural languages (English included, and in contrast with e.g. most Romance languages) 'obligation' is often used in a normatively 'oomphy' fashion. Here, we use obligation in a technical, non-normatively-thick sense, to simply map on to *ought*: being under an obligation, then, does not imply, in our sense, anything more normatively 'oompmhy' than simply being the subject to a corresponding ought.

Consider, again, George and his level of trustworthiness with respect to washing the dishes. If George has a maximally strong disposition to fulfil his obligations to wash the dishes, then he is maximally trustworthy to do so. It doesn't matter *why* he is disposed to fulfil those obligations. It *could* be feelings of goodwill (as for Baier), a good character (as for Potter), but none of these *must* be the reason George is disposed to fulfil his obligations. Having the relevant disposition is enough, regardless of what inner reason (if any) motivates George to have that disposition.

Trustworthiness, on this account too, comes apart from reliability: I can be perfectly reliable with regard to *phi*-ing but not *trustworthy*. This can be because, for instance, in *phi*-ing, I do not fulfil any relevant obligation, since I am not under any relevant obligation to *phi*.

²⁵ See here for an article symposium on our account and our responses: <https://link.springer.com/collections/jgefachiaa>.

Let's clarify a few details before pressing on. Dispositions have both stimulus conditions and manifestation conditions. The stimulus condition for trustworthiness is having an obligation to φ : obligations occasion the opportunity for trustworthiness. The manifestation condition for trustworthiness to φ is simply φ -ing, or fulfilling one's obligation to φ .

Dispositions should be evaluated relative to suitable conditions. One can be maximally trustworthy with respect to φ -ing even if one fails to fulfil one's obligations to φ in unsuitable situations. For example, it wouldn't count against George's trustworthiness to wash the dishes if he failed to manifest his disposition to do so when the water stopped working.

So far, we've given an account of maximal trustworthiness. But trustworthiness is a matter of degree. How trustworthy one is, we think, is just a function of how close someone is to being maximally trustworthy. Suppose George reliably does all his homework (when obligated to do so) unless his favourite football team have a match whereas Ann reliably does all her homework (when obligated to do so) even if her favourite football team are playing. But neither finishes all their homework when the Eurovision finals are on. Neither is maximally trustworthy with respect to doing their homework. Nevertheless, Ann is more trustworthy than George at doing all her homework because Ann is *closer* to being maximally trustworthy.

Someone can be trustworthy without being maximally trustworthy. That's because 'trustworthy,' like 'tall,' is a context-sensitive term. Whether one is outright trustworthy to φ (as assessed in a given context) depends on whether one is trustworthy *enough*:

Outright Trustworthiness to φ : ' S is trustworthy to φ ' is true in context c if and only if S approximates maximal trustworthiness to φ closely enough to surpass a threshold on degrees of trustworthiness determined by c .

It shouldn't be surprising, on reflection, that trustworthiness works this way. Someone might be trustworthy to watch your kids in one context (when you're in the room next door), but not in another (when you're out of the country for a week).

We defend this account of trustworthiness in more detail elsewhere.²⁶ But here are a few quick reasons we like it. First, as already noted, it nicely explains why and how trust and reliability come apart. We can only be trusted to do things that fulfil obligations. Second, our account is like Hawley's in that it doesn't refer to anything about the inner life, motivations, or character of the relevant agent. This de-anthropocentrized approach rightly leaves open the possibility of AI trustworthiness, and it also avoids the troublesome cases of trustworthy but motivationally atypical human agents.

But third, unlike Hawley's view, our account is not negative. People (or AIs) can be untrustworthy with respect to φ -ing even if they fail to actively incur a commitment to φ . That's because we are subject to some obligations whether we commit to them or not. Again, this seems right. I don't get to be 'off the hook' for being untrustworthy with respect to not pushing people down the stairs just because I fail to *commit* to not pushing people down the stairs.

But wait - does our account really predict that AI can be (non-trivially) trustworthy? After all, our view says that AI can only manifest trustworthiness to φ when they have an *obligation* to φ . But are AI the sorts of things that can be subject to obligations?

Two things about this: first, recall that we use obligation in a technical, non-normatively-thick sense, to simply map on to *ought*: being under an obligation to φ , then, does not imply, in our sense, anything more normatively 'oomphy' than simply being the subject of an ought to φ .

Second: Are AIs subject to oughts? Yes, they are. Once more, artefacts, just like traits and activities, can be governed by norms sourced in their *functions*. There are, for instance, properly and improperly

²⁶ So far, we've just been talking about attributions of being trustworthy to φ . But sometimes, we attribute trustworthiness *simpliciter*. We think trustworthiness to φ is the fundamental notion, which is why we've focused on it here. But see Kelp & Simion (2023) for our preferred treatment of trustworthiness *simpliciter*.

functioning hearts, democracies, and washing machines. Just so, there are properly and improperly functioning AI systems.

Once more, two kinds of functions that are relevant for us here— the design function and the etiological function. AI systems can have both kinds of functions. As we have seen, it's also possible for artefacts to develop etiological functions that differ from their design functions: Suppose a programmer created an AI system to do X, but it turns out to do Y very well, and that's why people keep using the relevant AI. When this happens, the etiological function (typically) overrides the normative significance of the design function.

So AI systems *are* subject to oughts. A diagnostic cancer AI *ought* to identify cancer - that's what its function is! And so, our attempt to de-anthropocize an account of trust that is applicable to AI has been successful. But we can clarify things by tweaking the definitions to make it explicit that the relevant kind of obligations are function-generated:²⁷

Maximal Trustworthiness for AI For all x where x is an AI, x is maximally trustworthy with regard to φ -ing if and only if x has a maximally strong disposition to meet its functional norms-sourced obligations to φ .

Attributions of Outright Trustworthiness to φ for AI For all x where x is an AI, ' x is trustworthy to φ ' is true in context c if and only if x approximates maximal trustworthiness to φ closely enough to surpass a threshold on degrees of trustworthiness determined by c .

Rational Trust in AI: For all x where x is an AI, it is *prima facie* epistemically rational for S to trust x to φ iff S does not have *prima facie* justification for believing that x is not trustworthy with respect to φ -ing.

Much of our work here has been theoretical. We've given an account of what trustworthiness for AI is and when it is rational to trust an AI. However, especially in the case of AI, we encounter an important practical problem. People often *don't* trust AI, even when it's very reliable and, by the lights of our view, trustworthy. Does our theoretical work yield any practical hints for how we might properly increase trust in AI?

Before answering, we want to emphasize that our conclusion here is *not* that people should trust AI more readily than they currently do. People should trust a particular AI only if they do not have *prima facie* justification to believe that it is not trustworthy. And while we're optimistic about possibility of trustworthy AI, we also aren't out to blame the layperson (like, perhaps, Lucie) who doesn't immediately trust a particular AI system. Whether Lucie should trust SmartCredit depends on whether Lucie has justification for believing SmartCredit to be untrustworthy.

Nevertheless, we don't want people to *needlessly* fail to trust AI. When AI is trustworthy, does our account say anything about how to increase such rational trust?

We think the movement to make AI more explainable is a good thing: after all, understanding why an AI has delivered a particular result may be evidence for its trustworthiness, and thus it might defeat extant evidence of lack of trustworthiness at the context, should it be present. The same will hold for the sorts of reliability indicators that CR emphasises. Having reason for thinking that the AI systems we use are specified, coded, and maintained through suitable reliability indicators, may also be evidence for their trustworthiness and thus may also contribute to making it rational to trust them by defeating evidence of lack of trustworthiness, should such evidence be present. Crucially, to say explainability and reliability indicators provide evidence for trustworthiness is not to say that either one is necessary for rationally trusting AI systems. It is just to say that, for cases in which defeat is present, they can provide defeater defeaters.

²⁷ See Simion and Kelp 2023.

This discussion also suggests two further, even more straightforward ways to build rational trust in AI: via induction and testimony, i.e. via *exposure* and *instruction*. For the latter, someone can come to rationally trust AI in spite of misleading defeat on the basis that they are *told* the relevant AI is trustworthy. This can be an important tool in *spreading* or *transmitting* trust in AI, but in some ways, it only pushes the problem back a step. How do we get testifiers to trust AI in the first place? Exposure to AI means giving people the opportunity to witness (or, better, interact with) AI in environments where (trustworthy) AI can consistently manifest their obligation-meeting dispositions, and thus acquire inductive evidence. People often learn about the disposition of an artefact to fulfil its function by seeing it at work (or, if possible, by trying it out for themselves). Our suggestion is that this applies to AI artefacts as much as it does to kitchen knives: Give AI opportunities to display their trustworthiness in a wide-range of low-stake situations to allow observers (or users) to form a stable inductive basis for defeat defeat.

Conclusion

We have explored whether it is rational for us to trust AI. We have first focused on AI testimony. We found the most prominent views according to which rational trust in the testimony of AI requires explanation or specification, coding, and maintenance via reliability indicators are too strong. And we have argued that our preferred view, the Default View, is preferable on both extensional and prior plausibility grounds. We then looked at rational trust in AI more generally. We argued that explainability is once again not required and argued for a view according to which rational trust in an AI requires that one lacks justification for believing the AI not to be trustworthy. We provided a view of trustworthiness in terms of dispositions to fulfil one's functions according to which AIs can be trustworthy and which allows for rationally trust in AI. At the same time, the view also leaves room for rational distrust in AI. Once more, if we are right, this has significant consequences for industry: the XAI research programme should be replaced with a programme developing trustworthy AI: AI with a strong disposition for ought compliance.

References

- Alston, William P. 1988. The Deontological Conception of Epistemic Justification, *Philosophical Perspectives*, 2: 257-299.
- Anscombe, G. E. M. 1957. *Intention*. Blackwell.
- Askill, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., Dassarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T.B., Clark, J., McCandlish, S., Olah, C., & Kaplan, J. (2021). A General Language Assistant as a Laboratory for Alignment. *ArXiv, abs/2112.00861*.
- Baier, Annette 1986. Trust and antitrust. *Ethics* 96:231-260.
- Bender, Emily M. and Alexander Koller. 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, Online. Association for Computational Linguistics.
- Bender, Emily M.; Timnit Gebru; Angelina McMillan-Major; Shmargaret Shmitchell (2021), *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *Association for Computing Machinery*, pp. 610–623
- Bogardus, Tomas 2014. Knowledge Under Threat. *Philosophy and Phenomenological Research* 88:289-313.
- Boghossian, Paul, and Williamson, Timothy (2020). *Debating the A Priori*. Oxford University Press.
- Bondy, Patrick & Carter, J. Adam (2018). The basing relation and the impossibility of the debasing demon. *American Philosophical Quarterly* 55 (3):203.
- Bonjour, Laurence 1980. Externalist Theories of Empirical Knowledge. *Midwest Studies in Philosophy* 5:53-73.
- Boult, Cameron 2021. Epistemic blame. *Philosophy Compass* 16:e12762.
- Boyle, Matthew 2009. Active belief. *Canadian Journal of Philosophy* Supplementary 35:119-147.
- Brown, Jessica 2020. What is Epistemic Blame? *Noûs*, 54:389-407.
- Buller, David J. 1998. Etiological theories of function: A geographical survey. *Biology and Philosophy* 13:505-527.
- Burge, Tyler 1993. Content preservation. *Philosophical Review* 102:457-488.
- Burge, Tyler (2013). Cognition through understanding: Self-knowledge, interlocution, reasoning, reflection: *Philosophical essays*, 3. Oxford: Oxford University Press.
- Butlin, Patrick and Viebahn, Emanuel 2023. AI Assertion. *Ergo: An Open Access Journal of Philosophy*.
- Cappelen, Herman and Dever, Josh 2021. *Making AI Intelligent: Philosophical Foundations*. New York, USA: Oxford University Press.
- Cappelen, Herman and Dever, Josh 2026. *Recognizing Artificial Minds: A Philosophical Defense of AI Cognition*. Oxford: Oxford University Press.
- Carter, J. Adam 2024a. *A Telic Theory of Trust*. Oxford University Press.
- Carter, J. Adam 2024b. *Digital Knowledge: A Philosophical Investigation*. London: Routledge.
- Carter, J. Adam & Titus, Lisa Miracchi (2024). What the tortoise should do: A knowledge-first virtue approach to the basing relation. *Noûs* 58 (2):456-481.
- Carter, Ronald (2012). *Vocabulary: Applied Linguistic Perspectives* (1st ed.). Routledge.
- Chalmers, David J. 2023. Could a large language model be conscious? *Boston Review* 1.
- Coley, John D. and Tanner, Kimberly D. (2012). Common Origins of Diverse Misconceptions: Cognitive Principles and the Development of Biology Thinking. *CBE: Life Sciences Education*, 11 (3): 209-215.
- Coliva, Annalisa. 2012. Varieties of failure (of warrant transmission: What else?!). *Synthese*, 189:335-354.
- Coliva, Annalisa 2026. *Social and Applied Hinge Epistemology*. Oxford University Press.
- Colombatto C, Fleming SM (2024). Folk psychological attributions of consciousness to large language models. *Neurosci Conscious*.
- Comesaña, Juan 2005. Unsafe Knowledge. *Synthese* 146:395-404.
- Craig, Edward. 1990. *Knowledge and the State of Nature. An Essay in Conceptual Synthesis*. Clarendon Press.

- Cummins, Robert E. 1975. Functional analysis. *Journal of Philosophy* 72:741-64.
- Davis, Emmalon 2016. Typecasts, Tokens, and Spokespersons: A Case for Credibility Excess as Testimonial Injustice. *Hypatia* 31:485-501.
- Dinges, Alexander & Zakkou, Julia (2020). Much at stake in knowledge. *Mind and Language* 36 (5):729-749.
- de Regt, Henk W. 2017. *Understanding Scientific Understanding*. New York: Oxford University Press.
- Dotson, Kristie 2012. A Cautionary Tale: On Limiting Epistemic Oppression. *Frontiers: A Journal of Women Studies* 33:24-47.
- Douven, Igor 2006. Assertion, knowledge, and rational credibility. *Philosophical Review*, 115:449-485.
- Dowty, David R. (1979). *Word Meaning and Montague Grammar*. Springer.
- Dubber, Markus D., Pasquale, Frank and Das, Sunit eds. 2020. *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.
- Durán, Juan M. 2026. Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms. In Juan M. Durán and Giorgia Pozzi, *Philosophy of science for machine learning: Core issues and new perspectives*. Springer. pp. 55-79.
- Evans, Ian (2013). The problem of the basing relation. *Synthese* 190 (14):2943-2957.
- Faulkner, Paul 2011. *Knowledge on Trust*. Oxford, GB: Oxford University Press.
- Fischer, John M. and Ravizza, Mark 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.
- Fricker, Elizabeth 1994. Against Gullibility. In Anjan Chakrabarti and Bimal K. Matilal, *Knowing from Words*. Kluwer Academic Publishers.
- Fricker, Elizabeth Forthcoming. On the metaphysical and epistemic contrasts between human and AI testimony. *Inquiry: An Interdisciplinary Journal of Philosophy*.
- Fricker, Miranda 2007. *Epistemic injustice: power and the ethics of knowing*. New York: Oxford University Press.
- Fricker, Miranda. 2012. Group Testimony? The Making of A Collective Good Informant. *Philosophy and Phenomenological Research* 84:249-276.
- Gettier, Edmund. 1963. Is justified true belief knowledge? *Analysis* 23:121-123.
- Goldberg, Sanford C. 2010. *Relying on others: an essay in epistemology*. Oxford: Oxford University Press.
- Goldberg, Sanford C. 2015. *Assertion: On the Philosophical Significance of Assertoric Speech*. New York: Oxford University Press.
- Goldberg, Sanford C. 2018. *To the Best of Our Knowledge: Social Expectations and Epistemic Normativity*. Oxford, United Kingdom: Oxford University Press.
- Goldman, Alvin I. 1979. What is Justified Belief? In George Pappas, *Justification and Knowledge: New Studies in Epistemology*. Boston: D. Reidel. pp. 1-25.
- Goldstein, Simon and Levinstein, Benjamin Anders forthcoming. Does ChatGPT Have a Mind? *Philosophy of AI*.
- Graham, Peter J. 2000. Conveying information. *Synthese* 123:365-392.
- Graham, Peter J. and Lyons, Jack C. 2021. The Structure of Defeat: Pollock's Evidentialism, Lackey's Framework, and Prospects for Reliabilism. In Jessica Brown and Mona Simion, *Reasons, Justification, and Defeat*. Oxford Oxford: Oxford University Press. pp. 39-68.
- Grimm, Stephen R. 2016. Understanding and Transparency. In Stephen R. Grimm, Christoph Baumberger and Sabine Ammon, *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. London: Routledge.
- Hattiangadi, Annandi (Forthcoming). *What Is Artificial General Intelligence?* Cambridge University Press.
- Hawley, Katherine 2019. *How to Be Trustworthy*. New York, NY: Oxford University Press.
- Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.
- Hicks, Michael Townsen ; Humphries, James & Slater, Joe (2024). ChatGPT is bullshit. *Ethics and Information Technology* 26 (2):1-10.
- Hieronymi, Pamela 2008a. Responsibility for Believing. *Synthese*. 161:357-73.

- Hieronymi, Pamela 2008b. The Reasons of Trust. *Australasian J Philos* 86(2):213–236
- Hills, Alison 2015. Understanding Why. *Noûs* 49:661–688.
- Honoré, A.M., 1964. Can and Can't, *Mind* 73:463–479.
- Hutson, Matthew 2012. *The 7 Laws of Magical Thinking: How Irrational Beliefs Keep Us Happy, Healthy, and Sane*. New York: Hudson Street Press.
- Kauppinen, Antti. 2018. Epistemic Norms and Epistemic Accountability, *Philosophers' Imprint* 18.
- Kelp, Christoph 2015. Understanding phenomena. *Synthese* 192:3799–3816.
- Kelp, Christoph 2016. Justified belief: Knowledge first-style. *Philosophy and Phenomenological Research*, 93:79–100.
- Kelp, Christoph 2018. Assertion: A Function First Account. *Noûs* 52:411–442.
- Kelp, Christoph 2021. *Inquiry, Knowledge, and Understanding*. Oxford University Press.
- Kelp, Christoph, and Simion, Mona 2017a. Commodious Knowledge. *Synthese*, vol 194(5): 1487–1502.
- Kelp, Christoph and Simion, Mona 2017b. Criticism and Blame in Action and Assertion. *Journal of Philosophy* 114:76–93.
- Kelp, Christoph, and Simion, Mona 2018. Assertion: The Constitutive Norms View (with C. Kelp). In Goldberg, S. (ed), *Oxford Handbook of Assertion*. Oxford University Press.
- Kelp, Christoph and Simion, Mona 2021. *Sharing Knowledge: A Functionalist Account of Assertion*. Cambridge: Cambridge University Press.
- Kelp, Christoph 2009. Knowledge and Safety. *Journal of Philosophical Research* 34:21–31.
- Kelp, Christoph and Simion, Mona 2023. What is trustworthiness? *Noûs* 57:667–683.
- Khalifa, Kareem 2017. *Understanding, Explanation, and Scientific Knowledge*. Cambridge, UK: Cambridge University Press.
- Kornblith, Hilary 2002. *Knowledge and its place in nature*. New York: Oxford University Press.
- Kvanvig, Jonathan. 2011. Norms of assertion. In J. Brown and H. Cappelen Eds., *Assertion: New Philosophical Essays*. Oxford University Press.
- Lackey, Jennifer 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.
- Lackey, Jennifer 2007. Norms of Assertion. *Noûs* 41:594–626.
- Lackey, Jennifer 2023. *Criminal Epistemic Injustice*. Oxford: Oxford University Press.
- Lehrer, Keith 2000. *Theory of knowledge*. Boulder, Colo.: Westview Press.
- Marcus, Gary, & Davis, Ernest (2020). GPT-3, Bloviator: Open AI's language generator has no idea what it's talking about. *MIT Technology Review*.
- McHugh, Conor. 2013. Epistemic Responsibility and Doxastic Agency. *Philosophical Issues* 23:132–57.
- Mele, Alfred. R. 2003, *Motivation and Agency*, Oxford: Oxford University Press.
- Millikan, Ruth G. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.
- Moore, G. E. 1939. Proof of an external world. In *Philosophical Papers*. Allen and Unwin, 1959.
- Nagel, Jennifer (2010). Epistemic anxiety and adaptive invariantism. *Philosophical Perspectives* 24 (1):407–435.
- Nagel, Jennifer (2017). Factive and nonfactive mental state attribution. *Synthese*, 194(11), 4225–4246.
- Neander, Karen 1991. The teleological notion of 'function'. *Australasian Journal of Philosophy* 69:454 - 468.
- Neta, Ram and Rohrbaugh, Guy 2004. Luminosity and the safety of knowledge. *Pacific Philosophical Quarterly* 85:396–406.
- Nickel, Philip J. 2013. Artificial Speech and Its Authors. *Minds and Machines* 23 4:489–502.
- O'Connor, Timothy 2021. Emergent Properties, *The Stanford Encyclopedia of Philosophy* Winter 2021 Edition, Edward N. Zalta ed., URL = <<https://plato.stanford.edu/archives/win2021/entries/properties-emergent/>>.
- Potter, Nancy N. 2002. *How Can I Be Trusted?: A Virtue Theory of Trustworthiness*. Rowman & Littlefield Publishers.
- Pritchard, Duncan 2005. *Epistemic Luck*. Oxford University Press.
- Pryor, James 2000. The skeptic and the dogmatist. *Noûs* 34:517–549.

- Pryor, James 2004. What's wrong with Moore's argument? *Philosophical Issues* 14:349-378.
- Putnam, Hilary 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7, 131-193.
- Quattrocio, Walter, Capraro, Valerio, Perc, Matjaz 2020. Epistemological Fault Lines Between Human and Artificial Intelligence. [10.31234/osf.io/c5gh8](https://doi.org/10.31234/osf.io/c5gh8) v1
- Saba, Walid S. (2023) Stochastic LLMs do not understand language: Towards symbolic, explainable and ontologically based LLMs. in Almeida JPA, Borbinha J, Guizzardi G, Link S, Zdravkovic J, (Eds.) *Concept. Model*, Springer Nature Switzerland, Cham, 3–19.
- Schlosser, Markus. 2019. Agency, The Stanford Encyclopedia of Philosophy Winter 2019 Edition, Edward N. Zalta ed., URL = <<https://plato.stanford.edu/archives/win2019/entries/agency/>>.
- Shah, Nishi 2003, How Truth Governs Belief, *The Philosophical Review* 112:447-482.
- Shah, Nishi and Velleman, J. David. 2005, Doxastic Deliberation, *The Philosophical Review* 114:497-534.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T.P., Simonyan, K., & Hassabis, D. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv*, [abs/1712.01815](https://arxiv.org/abs/1712.01815).
- Simion, Mona Forthcoming. Exercising Epistemic Agency. *Criminal Law and Philosophy*.
- Simion, Mona 2016. Perception, History, and Benefit. *Episteme*, 13(1): 61-76.
- Simion, Mona 2020. A priori perceptual entitlement, knowledge-first. *Philosophical Issues* 30:311-323.
- Simion, Mona 2021a. Testimonial contractarianism: A knowledge-first social epistemology. *Noûs* 55:891-916.
- Simion Mona 2019. Knowledge-first functionalism. *Philosophical Issues* 29: 254–267.
- Simion, Mona 2021b. Shifty Speech and Independent Thought: Epistemic Normativity in Context. Oxford: Oxford University Press.
- Simion, Mona 2022. Closure, Warrant Transmission, and Defeat. In *New Perspectives on Epistemic Closure*, Eds.: Jope M. and Pritchard, D. Routledge.
- Simion, Mona 2024a. Resistance to Evidence. Cambridge: Cambridge University Press.
- Simion, Mona 2024b. Resistance to evidence and the duty to believe. *Philosophy and Phenomenological Research*, 108, 203-216.
- Simion, Mona. 2025. Knowledge-First Epistemology: A Defence. Cambridge: Cambridge University Press.
- Simion, Mona and Kelp, Christoph 2020a. How to be an anti-reductionist. *Synthese* 197 7:2849-2866.
- Simion, Mona and Kelp, Christoph 2020b. Trustworthy artificial intelligence. *Asian Journal of Philosophy* 2:1-12.
- Simion, Mona and Willard-Kyle, Christopher 2026. Trusting AI: Explainability vs. Trustworthiness. *Communication with AI: Philosophical Perspectives*, (Herman Cappelen and Rachel Sterken, eds.), Oxford University Press.
- Sliwa, Paulina 2015. Understanding and Knowing. *Proceedings of the Aristotelian Society* 115:57-74.
- Slote, Michael. 1979. Assertion and belief. In J. Dancy Ed., *Papers on Language and Logic*. Keele University Press.
- Sosa, Ernest 2015. *Judgment and Agency*. Oxford: Oxford University Press.
- Sterken, Rachel K., and James R. Kirkpatrick. 2024. Conversational Alignment With Artificial Intelligence in Context. *Philosophical Perspectives* 38, no. 1: 38, 89–102.
- Turri, John (2011). Believing for a Reason, *Erkenntnis*, 74(3): 383–397.
- Unger, Peter 1975. *Ignorance: A Case for Scepticism*. Clarendon.
- van Woudenberg, René, Ranalli, Chris and Bracker, Daniel 2024. Authorship and ChatGPT: a Conservative View. *Philosophy and Technology* 37:1-26.
- Wedgwood, R. (2006). The normative force of reasoning. *Noûs*, 40(4):660–686.
- Williams, Bernard 1973, Problems of the Self: *Philosophical Papers* 1956-1972, Cambridge: Cambridge University Press.
- Williamson, Timothy 2000. *Knowledge and Its Limits*. Oxford University Press.

Wright, Crispin 2004. Warrant for nothing (and foundations for free)? *Proceedings of the Aristotelian Society*, Suppl. Vol. 78:167-212.

Wright, Stephen 2018. *Knowledge Transmission*. London: Routledge.

s