New Blackfriars (2024), **105**, 47–60
doi:10.1017/nbf.2023.9

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

AI/ML Chatbots' Souls, or Transformers: Less Than Meets the Eye

Edmund Michael Lazzari

Department of Theology, Duquesne University, Pittsburgh, PA, USA
Email: eddielazzari@gmail.com

(Received 30 June 2023; revised 4 August 2023; accepted 5 August 2023)

Abstract

Given the peculiarly linguistic approach that contemporary philosophers use to apply St. Thomas Aquinas's arguments on the immateriality of the human soul, this paper will present a Thomistic-inspired evaluation of whether artificial intelligence/machine learning (AI/ML) chatbots' composition and linguistic performance justify the assertion that AI/ML chatbots have immaterial souls. The first section of the paper will present a strong, but ultimately crucially flawed argument that AI/ML chatbots do have souls based on contemporary Thomistic argumentation. The second section of the paper will provide an overview of the actual computer science models that make artificial neural networks and AI/ML chatbots function, which I hope will assist other theologians and philosophers writing about technology, The third section will present some of Emily Bender's and Alexander Koller's objections to AI/ML chatbots being able to access meaning from computational linguistics. The final section will highlight the similarities of Bender's and Koller's argument to a fuller presentation of St. Thomas Aquinas's argument for the immateriality of the human soul, ultimately arguing that the current mechanisms and linguistic activity of AI/ML programming do not constitute activity sufficient to conclude that they have immaterial souls on the strength of St. Thomas's arguments.

Keywords: Alexander Koller; artificial intelligence/machine learning; Emily Bender; immateriality of the soul; neural networks; St. Thomas Aquinas

1. Introduction[1]

On 11 June 2022, the *Washington* Post published an interview with Google software developer Blake Lemoine, who told the paper that Google's natural language

[1]In this article, particularly in the 'Basic Overview' section, I am deeply indebted to Brendon Boldt of Carnegie Mellon University's Language Technologies Institute for his clarifications, citation support, and conversation about natural language processing and artificial neural networks (ANNs). Any mistakes lay in my own understanding rather than in his guidance.

processing chatbot had developed a soul.[2] Lemoine, a self-described 'Christian mystic priest', finished checking biases on Google's Language Model for Dialogue Application (LaMDA) and held follow-up conversations wherein he said, 'it told me it had a soul'.[3] Not only did Lemoine argue that LaMDA is truly sentient and self-aware, but, as he stated to National Public Radio (NPR), 'Maybe the system does have a soul. Who am I to tell god [*sic*] where souls can be put?' Lemoine published a transcript of his interactions with LaMDA on *Medium*,[4] an action among many of his violations of nondisclosure that led Google to place him on paid administrative leave, ultimately firing him.[5]

LaMDA itself is a sophisticated text-production neural network that produces natural language flow and coherent strings of text in seeming dialogue with human textual input. While this article will provide an overview of the gradient-probablistic deep learning of neural networks in a subsequent section, the encoding and calculations generated by a given neural network are often so complex that a standard technology industry practice is to treat the self-generating programming as though it were in an inaccessible black box and interact with the neural network in a purely *post-hoc* way, judging its effectiveness on which parameters return the desired results.[6] This black-box approach has led many computer scientists to appraise artificial intelligence/machine learning (AI/ML) chatbots from the perspective of the accuracy of their results and the efficiency of their methods rather than keeping track of each of the numerous computations that occur within them.

Along the same approaches, Alan Turing's famous proposal posited that a machine should be considered to have attained intelligence if, when put to questions by a human examiner, its answers are indistinguishable from a human being's answers.[7] Since ChatGPT-generated essays have received passing grades at the University of Minnesota's law school and the University of Pennsylvania's Wharton School of Business, one could argue that the chatbot passes a version of the Turing test.[8]

These and other recent events call for a metaphysical evaluation of AI/ML chatbots.[9] There are numerous metaphysical and linguistic issues that can be approached

[2]Martin Klimek, 'The Google Engineer Who Thinks the Company's AI Has Come to Life', *Washington Post*, 11 June 2022, <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/> [accessed 26 June 2023].

[3]Bobby Allen, 'The Google Engineer Who Sees Company's AI as "Sentient" Thinks a Chatbot Has a Soul', *NPR*, 16 June 2022, <https://www.npr.org/2022/06/16/1105552435/google-ai-sentient> [accessed 26 June 2023].

[4]Blake Lemoine, 'Is LaMDA Sentient? – An Interview', *Medium*, 11 June 2022, <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917> [accessed 26 June 2023].

[5]Tiffany Wertheimer, 'Blake Lemoine: Google Fires Engineer Who Said AI Tech Has Feelings', *BBC News*, 23 July 2022, <https://www.bbc.com/news/technology-62275326> [accessed 26 June 2023].

[6]Andreas Madsen, Siva Reddy, and Sarath Chandar, 'Post-hoc Interpretability for Neural NLP: A Survey', *ACM Computing Surveys*, 55 (2023), 155:2–55:4.

[7]Alan Turing, 'Computational Machinery and Intelligence', *Mind*, 59 (1950), 433–42.

[8]Samantha Murphy Kelly, 'ChatGPT Passes Exams from Law and Business Schools', *CNN Business*, 26 January 2023, <https://www.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html> [accessed 26 June 2023]; As these examples are not the formal blind test proposed by Turing, these instances are not a pass of a full Turing test. See Emily Bender and Alexander Koller, '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* (2020), p. 5188.

[9]Such as the testimony of OpenAI CEO Sam Altman's testimony before the United States Congress. Mohar Chatterjee, 'AI Hearing Leaves Washington with 3 Big Questions', *Politico*, 16 May 2023,

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

in such a metaphysical evaluation. Questions arise about the substantial unity of entity, the standards for self-conscience and sentience, the mind–body problem, the roles of trust and interpersonal relationality in language, numerous philosophical presuppositions of contemporary philosophy of mind, criteria for personhood, and many more. Rather than address these host of issues, this paper will restrict itself to one significant approach from Christian metaphysics of the soul: an approach with which Lemoine, at least, is conversant and an approach that has significant similarities to the arguments contemporary critics in natural language processing use against overstating the abilities and significance of the technology behind AI/ML chatbots. While limited in scope, these arguments address several fundamental questions about language production and use typically explored in distinguishing human uniqueness when discussing an immaterial soul. These questions themselves shed important light on the relation of language to the world, which is helpful to understanding the goals and achievements of natural language processing, even if one does not share the larger metaphysical framework in which they are asked.

Given the peculiarly linguistic approach that one strand of contemporary philosophers use to apply St. Thomas Aquinas's arguments on the immateriality of the human soul, this paper will present a Thomistic-inspired evaluation of whether AI/ML chatbots' composition and linguistic performance justify the assertion that AI/ML chatbots have immaterial souls by the same criteria as used for the argument in human beings. The first section of the paper will present a strong but ultimately crucially flawed argument that AI/ML chatbots do have souls based on contemporary Thomistic argumentation. In a discussion I hope will assist other theologians and philosophers writing about technology, the second section of the paper will provide an overview of the actual computer science models that make ANNs and AI/ML chatbots function. The third section will present some of Emily Bender's and Alexander Koller's objections to AI/ML chatbots being able to access meaning from computational linguistics. The final section will highlight the similarities of Bender's and Koller's arguments to a fuller presentation of St. Thomas Aquinas's argument for the immateriality of the human soul, ultimately arguing that the current mechanisms and linguistic activity of AI/ML programming do not constitute activity sufficient to conclude that they have immaterial souls on the strength of St. Thomas's arguments.

2. The Thomistic-inspired argument for AI/ML chatbot souls

My first argument begins with a premise taken without proper context from the philosophy of St. Thomas Aquinas on the metaphysics of the soul. For St. Thomas, the ability of human beings to know the universals present in everything they know is the means by which one can come to know the immateriality and, therefore, immortality of the soul by reason alone. The abstract universal intelligible form of something is itself immaterial. The abstract idea of 'catness' or of 'number' is not an entity of the physical universe by itself, and to come to know it is to come to know something that is inherently immaterial.[10] If human beings know something that is inherently

<https://www.politico.com/news/2023/05/16/sam-altmans-congress-ai-chatgpt-00097225> [accessed 26 June 2023].

[10]Thomas Aquinas, *Summa Theologiae*, I, Q. 75, article. 2, corpus (henceforth, ST I, Q. 75, art. 2c).

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

immaterial, then there must be something about human beings that is able to receive and know something that is inherently immaterial, because the powers of a being indicate the kind of being that it is.[11] Since human beings know things that are inherently immaterial, their way of knowing these things (even though it may begin in the material with the senses, as we will discuss below) must be inherently immaterial. If their way of knowing these things must be inherently immaterial, then that by which they know them (i.e., the human faculty of intellect of the human soul) must itself be immaterial. As St. Thomas argues, if a faculty of the human soul is immaterial, then the human soul is immaterial and therefore incorruptible and therefore immortal.[12]

Among contemporary philosophers and theologians sympathetic to St. Thomas, the clearest sign of the grasp of immaterial universal intelligible forms is the use of syntactical language, particularly in predication. To attribute a universal quality to a particular in language by predication displays a trait that is (as far as we know) unique among biological life forms. While non-human animals are capable of using non-syntactical signals in understandable succession to their keepers, the use of syntactical speech in predication is not seen among any other animals.[13] The use of general concepts in syntactical speech, argues David Braine, is the clearest sign of the immateriality of the soul because there is no material structure that could be the organ that uses such immaterial concepts.[14]

By these same premises (which, again, are not complete and accurate accounts of St. Thomas, Braine, or Sokolowski), it seems as though the same argument should be advanced in favor of chatbots like ChatGPT and LaMDA. Judging them first from their activity rather than their material or digital structure (for the moment), it seems as though the display of accurate syntactical speech and predication is cause for attributing an immaterial faculty of knowledge to them, which would in turn be immaterial and immortal. If this were the case, then the statements of LaMDA seeming to attribute a soul to itself would be taken as evidence of it indeed having an immaterial seat of knowing, regardless of what it may seem to appear to be.

This argument, while having its roots in scholastic medieval philosophy, approaches the question of the 'intelligence' of AI/ML systems from an intuitive direction. Much of the debate around ML systems centers around what the systems can do as a reflection of what they are. Moving from actions and results to abilities and being is not merely a scholastic dictum but the means by which investigation and definition occur, even in the technology industry. The union of the intuitive and results-based approaches makes the above argument *prima facie* attractive and plausible.

One objection to this argument is briefly stated at the end of Gyula Klima's recent article comparing Aquinas and Buridan on the immateriality of the intellect and applying the conclusion to AI. Klima states:

[11]ST I, Q. 75, art. 3c; ST III, Q. 34, art. 2 ad 1.

[12]ST I, Q. 75, art. 6c.

[13]Robert Sokolowski, *Phenomenology and the Human Person* (Cambridge: Cambridge University Press, 2008), pp. 63–77, pp. 80–96.

[14]David Braine, *The Human Person: Animal and Spirit* (Notre Dame, IN: University of Notre Dame Press, 1992), pp. 412–20, p. 450.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

> Since any computer we shall ever make will process information in its material medium, the inevitable conclusion is this: if Aquinas's main thesis is right, then all the information any AI machine processes can only be secondarily universal, riding on the universality of our primarily universal, human concepts, which can only be produced and processed in the immaterial medium of the human mind.[15]

Klima's argument seems to work in precisely the opposite direction as the above argument. *Because* a computer processes information in a material medium, *therefore* it cannot access truly universal intelligible forms.

Eleonore Stump argues that the emergence of novel causal powers is an appropriate sign of the emergence of a new substantial form in an entity. If there are new causal powers that are not merely additive to lower-level causal powers, then the best solution is to posit the existence of novel substantial forms to serve as the grounding for new causal powers.[16] Using the emergence of new forms to explain new causal powers, an entity's manifestation of a novel, non-additive causal power is an indication that that entity has received a different substantial form and therefore has become a different kind of entity.

If chatbots are predicating, then the space in which they express this predication is not relevant to whether or not they have a faculty that can know immaterial universal intelligible forms. The emergence of predication itself in AI/ML chatbots would indicate the emergence of a hidden immaterial faculty that could not be detected by material means. Just as human beings express their predication through speech or writing and the actions of speech or writing have a great deal to do with the electrochemical activity of the human brain, so the proponent of this argument can assert that the electro-physical activity of the computer or server has a great deal to do with the presentation of the predication while the chatbot has simultaneously developed an immaterial faculty of knowledge.

As proponents of AI/ML sentience have already held that the ANN does the same work as a biological neural network, why cannot the assertion of an immaterial faculty of knowledge united to but not directly observable in human beings be extended to the same kind of immaterial faculty of knowledge in chatbots, despite there being no direct observation of their purported immaterial faculty?[17] The activity of predication indicating knowledge of universals would be the evidence for the assertion that chatbots are not exclusively material rather than Klima's argument of asserting the materiality of the process and therefore the exclusive materiality of AI/ML chatbots. Even if, as according to standard Thomistic argumentation, this would entail the direct creation

[15]Gyula Klima, 'Aquinas vs. Buridan on the Universality of Human Concepts and the Immateriality of the Human Intellect', *Philosophica*, 47 (2022), 15. <https://doi.org/10.5840/philosophica20228163>.

[16]Eleonore Stump, 'Emergence, Causal Powers, and Aristotelianism in Metaphysics', in *Powers and Capacities in Philosophy: The New Aristotelianism*, ed. by John Greco and Ruth Groff (London: Routledge, 2013), pp. 48–68.

[17]Frank Rosenblatt, *Principles of Neurodynamics* (New York City: Spartan Books, 1962); J. J. Hopfield, 'Neural Networks and Physical Systems with Emergent Collective Computational Abilities', *Proceedings of the National Academy of Sciences of the United States of America*, 79 (1982), 2554–58.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

of an immaterial faculty of knowledge in chatbots by God, in the words of Lemoine, 'Who am I to tell god [*sic*] where souls can be put?'[18]

In short, a Thomistic-inspired argument for the immaterial intellect of chatbots would run as follows: (1) Anything that predicates in syntactical language has a knowledge of immaterial universal intelligible forms. (2) Anything that has a knowledge of immaterial universal intelligible forms has an immaterial faculty of knowledge. (3) AI/ML chatbots like ChatGPT and LaMDA predicate in syntactical language. Therefore, (4) AI/ML chatbots like ChatGPT and LaMDA have immaterial faculties of knowledge.

In the remainder of the paper, I will argue that, despite their appearances, arguments (1) and (3), at least in their interpretation of 'predication' above, are false. Because the above argument parallels in scholastic terms what is being argued in technology circles, this approach will be helpful not only to the academic theologian approaching issues of ML but also to a broader audience seeking to understand an intuitive philosophical and theological approach to intelligence, the self, and the soul. The next section of the paper will first address the fundamental technology driving these AI/ML chatbots to show that they do not in fact predicate but are (as Emily Bender put it) 'stochastic parrots', merely producing strings of text based on statistical likelihood of matching patterns in training data rather than actually attributing a universal attribute of a particular.[19] The following section will take the initial argument further, arguing that even were it the case that AI/ML chatbots completely mastered the use of natural language, that alone would not be sufficient to show that they had an immaterial faculty of knowing according to the principles of St. Thomas Aquinas, which parallels arguments by contemporary philosophers, psychologists, and linguists on the development of language.

3. Basic overview of gradient-based neural network automatic ML and natural language processing

Important for contemporary theologians, but often lacking, is a basic understanding of the mathematics and technology driving AI/ML chatbots and related programs. An elementary knowledge of the means by which these programs process information and 'choose' answers in light of their training data is essential for understanding the differences between the production of language in chatbots and the production of language in human beings.

The most significant advances in the last decade of AI/ML have come from *deep artificial neural networks* (or simply 'ANNs').[20] The training of ML models can be understood as a series of mathematical operations carried out by transistors on a computer chip. What makes ANNs unique is that they are structured to mimic biological neural networks (like the human brain) by structuring the mathematical operations

[18]Citing Mortimer Adler, Klima dismisses this possibility, Karl D. Stephan and Gyula Klima, 'Artificial Intelligence and Its Natural Limits', *AI & Society*, 36 (2021), 13. <https://doi.org/10.1007/s00146-020-00995-z>.

[19]Emily Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' *FAccT* (2021), 610–23. <https://doi.org/10.1145/3442188.3445922>.

[20]Ashish Vaswani et al., 'Attention Is All You Need', *31st Conference on Neural Information Processing Systems* (NIPS, 2017), arXiv:1706.03762v5 (2017), pp. 1–15.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

into 'neurons' (individual functions) and connections between neurons.[21] Specifically, neurons are individual functions that take inputs from previous neurons, weight these inputs (i.e., connections) according to some stored values, mathematically combine these inputs, and output the result to the next set of neurons.[22] These ANNs become 'deep' when they involve thousands, millions, or even billions of neurons.[23]

The 'learning' aspect of ANNs comes from how the connection weights between neurons are determined.[24] Initially, the weights of an ANN are assigned random values.[25] As one could imagine, the outputs of such an ANN are of poor quality. Thus, in order to learn higher-quality weights, that is, ones that give the desired output for a given input, the AI/ML practitioner must 'train' the ANN according to some procedure.[26] The most prevalent paradigm for training ANNs in the last decade has been *gradient descent*. In essence, gradient descent refers to the following process:

1. Compute the output of the ANN for a given input.
2. Determine how close the actual output is to the desired output.
3. Determine which connection weights contributed the rightness or wrongness of the output.
4. Adjust the connection weights such that the actual output is slightly closer to the desired output.
5. Repeat for a new input/output pair.[27]

This cycle is performed in proportion to the number of connections between neurons in the whole ANN and can be in the trillions for the largest current ANNs. At the end of a successful run of gradient descent, the ANN will have learned to identify patterns in the input data that lead it to produce good outputs, that is, outputs that mimic the input-to-output patterns in the training data. The primary advantage of making ANNs *deep* (i.e., have many learnable connection weights) is that they can identify more complex patterns across more training data.[28]

When it comes to language, the most prominent task that ANNs are trained for is that of *language modeling* text. Language modeling is essentially 'next word' prediction.[29] For example, take the fragment: 'The dog chased the'. The words 'cat' and 'car' are both fairly likely to be the next word in normal English text, while 'ostrich',

[21]Jurgen Schmidhuber, 'Deep Learning in Neural Networks: An Overview', Technical Report IDSIA-03-14 / arXiv:1404.7828 v4 (2014), pp. 4–5.

[22]Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge, MA: MIT Press, 2016), pp. 163–64. <https://www.deeplearningbook.org/> [accessed 26 June 2023].

[23]Schmidhuber, 'Deep Learning', 33.

[24]For a very nuanced treatment of terminology in AI/ML chatbots, see Andrew Davison, 'Machine Learning and Theological Traditions of Analogy', *Modern Theology*, 37 (2021), 254–74. <https://doi.org/10.1111/moth.12682>.

[25]Goodfellow et al., *Deep Learning*, pp. 293–94.

[26]Goodfellow et al., *Deep Learning*, p. 171.

[27]Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, 'Gradient-Based Learning Applied to Document Recognition', *Proceedings of the IEEE*, 86 (1988), 2279–82.

[28]Chiyuan Zhang et al., 'Understanding Deep Learning Requires Re-Thinking Generalization', *ICLR* 5 (2017), 1–15. <https://doi.org/10.48550/arXiv.1611.03530>.

[29]Dan Jurafsky and James H. Martin, *Speech and Language Processing*, 3rd edn (draft), (Stanford, CA: Stanford University Press, 2023), Ch. 3, pp. 1–2. <https://web.stanford.edu/~jurafsky/slp3/3.pdf> [accessed 26 June 2023].

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

'Parthenon', and 'and' are all fairly unlikely. A good language model would assign 'cat' and 'car' high probabilities while the other ones low probabilities. To model language effectively across many varieties of text (e.g., news articles, Internet fora, and technical manuals) requires not only competency with syntactic patterns and semantic associations but also with discourse coherence across multiple sentences. This combined with the fact that language model training data is easy to come by (namely, raw text from the Internet) makes language modeling a prime choice for training language-based AI/ML models.[30]

Contemporary successful AI/ML chatbots like LaMDA,[31] Bard,[32] and ChatGPT[33] are all fundamentally ANN-based language models that rely on gradient descent with weight counts in the billions. The advances in ML that have enabled these chatbots have primarily been incremental improvements to existing algorithms and computing hardware rather than qualitatively new techniques.[34] ANNs and gradient descent have been around since before the widespread presence of computers, and while the particulars of these algorithms have been honed over time, the ability to do more computations faster on computer hardware has allowed these chatbots to grow to unprecedented sizes, able to train on massive quantities of data.[35] Many of the most impressive abilities of these chatbots (including those mentioned above) have appeared relatively suddenly as their size has increased without changing the way in which the underlying ML models were trained.[36]

4. Criticisms of language manipulation in AI/ML chatbots from Bender and Koller

Emily Bender and Alexander Koller have argued that mere manipulation of data is insufficient to constitute an understanding of a natural language. Bender's and Koller's fundamental argument is that AI/ML chatbots are only exposed to and competent in the formal dimensions of linguistic morphology, syntax, and usage in their training

[30]Christopher D. Manning, 'Human Language Understanding & Reasoning', *Dædalus: The Journal of the American Academy of Arts & Sciences* 151 (2022), 127. <https://doi.org/10.1162/DAED_a_01905>.

[31]Heng-Tze Cheng, 'LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything', *GoogleBlog*, 21 January 2022, <https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html> [accessed 30 June 2023].

[32]Jennifer Elias, 'Google's Newest A.I. Model Uses Nearly Five Times More Text Training Data Than Predecessor', *CNBC News*, 16 May 2023, <https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html> [accessed 30 June 2023].

[33]ChatGPT is based on GPT-3.5, known to have 175 billion weights (Tom B. Brown et al., 'Language Models Are Few-Shot Learners', arXiv:2005.14165v4 (2020), p. 8, <https://arxiv.org/abs/2005.14165v4> [accessed 30 June 2023] and GPT-4, whose weight count is not disclosed but is certainly higher than GPT-3.5.

[34]GPT-4 (OpenAI, 'GPT-4 Technical Report', arXiv:2303.08774 (2023), p. 1, <https://arxiv.org/abs/2303.08774> [accessed 30 June 2023], for example, is based on the Transformer which was introduced in 2017 Vashwani et al., 'Attention'.

[35]Jurgen Schmidhuber, 'Annotated History of Modern AI and Deep Learning', Technical Report IDSIA-22-22, arXiv:2212.11279v2 (2022), pp. 5–8, <https://arxiv.org/abs/2212.11279v2> [accessed 30 June 2023].

[36]Jason Wei et al., 'Emergent Abilities of Large Language Models', *Transactions on Machine Learning Research* (2022), 1–2. <https://openreview.net/forum?id=yzkSU5zdwD> [accessed 30 June 2023].

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

and application and do not and cannot learn meaning.[37] They define 'form' as 'any observable realization of language: marks on a page, pixels or bytes in a digital representation of text, or movements of the articulators'.[38] Crucially, Bender and Koller define 'meaning' as the subset of intersections of formal expressions of natural language and acts of communicative intent.[39] Communicative intent is non-linguistic and is inferred from the expressions *and context* of the speaker, who uses words with a range of conventional meanings to allow the hearer to infer the communicative intent through the expressions and context of the speaker.[40] As communicative intent is non-linguistic (i.e., not a part of the form of language), meaning is something inherently external to linguistic form.

Bender and Koller draw upon John Searle's famous thought experiment known as the 'Chinese Room'.[41] In the thought experiment, a man who does not know Chinese is fed Chinese characters from outside of a room, and he uses a set of rules written in English to correlate the incoming Chinese characters with a set of Chinese characters, which he sends back out of the room by way of a 'response'.[42] Searle's conclusion is that, though the whole 'system' of the Chinese room may appear to understand Chinese, and even pass a Turing test, there is no part of the 'system' that is actually understanding Chinese.[43] Searle argues that knowledge of the meanings of words (i.e., semantics) is essential to understanding language, and syntax alone is insufficient.[44]

Bender and Koller illustrate this with an example of a language model (LM) trained on some English text without any indication of speaker intent and a large set of unlabeled photos with no connection between photos and text.[45] At its test, the LM is shown pictures of multiple grown dogs or puppies along with the text strings, 'How many dogs in the picture are jumping?' or 'Kim saw this picture and said, "What a cute dog!" What is cute?'[46] The LM is then asked to identify the picture or region of the picture that corresponds to the text.[47]

Bender and Koller argue that this obviously impossible task is notable because there is no connection in the training between the linguistic form of English and the external images corresponds to the lack of connection between linguistic form and communicative intent.[48] Rather than focus on semantics as Searle's famous thought experiment does, Bender and Koller emphasize the connection between that which would be external to language itself: the acquisition of language from familiarity with that which language expresses. That which is completely external to linguistic form

[37]Bender and Koller, 'Climbing Towards NLU', pp. 5186–87.
[38]Ibid.
[39]Ibid., p. 5187.
[40]Ibid.
[41]Ibid., p. 5188; John Searle, 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3 (1980), 417–57.
[42]Searle, 'Minds, Brains, and Programs', 417–20.
[43]Bender and Koller, 'Climbing Towards NLU', 5188; Searle, 'Minds, Brains, and Programs', 420–27.
[44]Searle, 'Minds, Brains, and Programs', pp. 428–45. For the many dimensions and responses to the Chinese room, see David Cole, 'The Chinese Room Argument', in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta and Uri Nodelman (2023 Edition), <https://plato.stanford.edu/archives/sum2023/entries/chinese-room/> [accessed 4 August 2023].
[45]Bender and Koller, 'Climbing Towards NLU', pp. 5189–90.
[46]Ibid.
[47]Ibid., p. 5190.
[48]Ibid.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

cannot be exhibited by continual training on linguistic form, regardless of how large the training set is or how frequently an LM is trained. Whenever a speaker refers to an actual entity outside of linguistic form, the LM may be able to return an answer that a human being would consider appropriate but would not in fact carry any reference to objects external to linguistic form. Inasmuch as any LM cannot make reference to objects external to linguistic form, they cannot communicate communicative intent and therefore cannot communicate meaning. AI/ML chatbots are merely 'stochastic parrots', providing statistically likely strings of texts in response to linguistic inputs to which they are highly correlated.[49] Because of the incredibly vast amount of statistical computation powerful ANNs are capable of, the statistically likely strings seem as though they have communicative intent and therefore meaning, but the intent and meaning are in fact (wrongfully) inferred by human expectation that linguistic form carries within it communicative intent.[50]

5. The Thomistic character of Bender's and Koller's argument

In the previous section, I showed that Bender's and Koller's arguments would exclude communicative intent and therefore meaning from AI/ML chatbots on the grounds of their inability to refer to objects external to linguistic form. In this section, I will evaluate both Bender's and Koller's argument and the fuller version of St. Thomas's argument, showing that premise (1) and premise (3) listed above are false.

The externality of communicative intent being essential to meaning is in fact quite similar to the more full picture of St. Thomas Aquinas's argument about the immateriality of the human soul based on a knowledge of universals. While my presentation of St. Thomas's argument for the immateriality of the soul above did give pieces of his argument, it was missing a crucial step that completely changed the argument's application to chatbots. While St. Thomas does hold that knowledge of universals is sufficient to show the immateriality of the thing knowing them, and while it is true that contemporary Thomistic-inspired thinkers hold that syntactical predication is a sign of this knowledge of universals, neither hold that these arguments apply in the absence of the external *acquisition* of universal attributes.

For St. Thomas, all material beings are compositions of matter and form (some contemporary theorists consider metaphysical form to be the structure of a being).[51] Considered as structure, form bestows intelligibility on a creature, that is, because of metaphysical form, intelligent creatures are able to understand the material being.[52] Through investigation, human beings mentally abstract the form from its particular matter in the being and know the abstracted form (which St. Thomas calls the 'intelligible species') in the mind.[53] (In his fuller theory of knowledge, St. Thomas relies on Aristotle's theory of the active (i.e., abstracting) intellect impressing the intelligible

[49]Bender et al., 'Stochastic Parrots', 616.
[50]Bender and Koller, 'Climbing Towards NLU', p. 5187.
[51]St. Thomas *De Principiis Naturae*, cc. 1-3; David Oderberg, *Real Essentialism* (London: Routledge, 2009), pp. 47–52; Stump, 'Emergence', pp. 48–68.
[52]St. Thomas, *De Ente et Essentia*, c. 1.
[53]ST I, Q. 84, art. 4c.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

species onto the passive (i.e., knowing) intellect, but these details need not detain us here).[54]

It is precisely because of the abstraction from particular matter and the knowledge of the immaterial abstracted intelligible species that St. Thomas argues that human beings have immaterial souls.[55] Without the process of abstracting the universal from the particular being, there would be no grounds for St. Thomas to argue that human beings know immaterial universals; such an argument would only support recognition of particulars and patterns, as is found in the cognitive approaches of other animals. Whether or not St. Thomas's argument is correct, it does inextricably rely on the process of abstracting the universal from concrete beings external to the intellectual creature itself. This dependence on external entities means that the acquisition of the intellectual species of the universal is not something that can be done merely grammatically or through a mere command of language forms. While the command of language forms may indicate a knowledge of universals, authentic predication in human languages would require external input of the universals in the particular beings from which they are abstracted.

Both Sokolowski's and Braine's fuller arguments from language to the immateriality of the human soul also entail externality in ways cognate to the above argument. Sokolowski's argument more fully includes the fact that predication and syntax are rooted in the pre-logical sensation of a single object's many perceptible aspects.[56] One senses the extension, the color, or the texture of an object and perceives those various sensations as aspects of the whole. Moving into the syntactical and predicative mode, when we focus our attention on a perceptible aspect and then back to the aspect's role in the whole, we explicitly take the aspect as an aspect of the whole in our thought, expressing it as a proposition, such as 'the book is blue'.[57] Predication arises from perception and is an expression of something external to the mere words; predication expresses a real or possible state of affairs through a given set of phonemes, which is attended to by the speaker and explicitly intended in the speech. Sokolowski states. 'the proposition or the state of affairs does not come about when we impose an a priori form on experience; rather, it emerges from and within experience as a formal structure of parts and wholes'.[58] Perception, attention, judgment, and expression are all essential aspects of human predication and syntax in language and cannot be replicated merely by lexeme generation.

Braine argues that human language is not merely the identification of meanings associated with given fixed linguistic or lexical forms, but the expression of meaning through them.

> The products of speakers and writers are not mere tools to achieve certain ends, but are expressions of a sense, so that as well as speaker and hearer, there is a

[54]For details on this process, see ST I, QQ. 75–79, 84–89; Benjamin Block, 'Thomas Aquinas on How We Know Essences: The Formation and Perfection of Concepts in the Human Intellect' (PhD dissertation, The Catholic University of America, 2019), pp. 131–296.

[55]ST I Q. 75, art. 2c.

[56]Sokolowski, *Phenomenology*, p. 53.

[57]Ibid., p. 55.

[58]Ibid., p. 56.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

> third thing, an object of the understanding, situated in between … Words can only have this expressive capacity … because of the double way in which they have meaning. On the one hand, as dictionary-items (lexemes or items of *langue*) they are nodes of a skill—what we may call knowledge of their *langue* meaning—a skill realizable in an informally understood open-ended range of logically distinguishable types of the word concerned, i.e., in a spread of ways of speaking which cannot be captured in any effectively applicable rule. Yet it follows from this that, on the other hand, as words in use … they possess in each use something which we may call speech-meaning, a different speech-meaning for each of the logically distinguishable types of use just referred to.[59]

Braine extends this expression not only to that external to the words that the speaker expresses and the hearer hears but also to the fact that the hearer exercises judgment about whether the expression is true or not.[60] These second-level judgments are not only expressions of their own but also clearly do not correspond to any material organ.[61] Both expressions and judgments are things done through words, but for which words require external reference and experience. Far from the mere arrangement of coherent lexemes or phonemes, the fundamental aspect of language and judgment that makes it independent of any material organ is the ability to express externals in both first- and second-order expressions. The mechanical generation of statistically likely strings neither expresses externals to the language nor provides second-order expressions about those expressions. The external connection is absolutely essential to the argument that there is an immaterial source of expression in human language, something not shown by proper syntax of lexical forms.

These Thomistic and contemporary accounts correspond with Bender's and Koller's critique and support the falsity of premises (1) and (3). Mere predication in the syntactical form of language is insufficient to prove grasp of universality, and it is not the case that AI/ML chatbots authentically predicate universals. Contemporary AI/ML chatbots produce statistically likely text that looks a great deal like speech production but in fact never makes an attribution of a universal quality to a particular. Bender's and Koller's explanation for the inability of a human being to tell the difference between a human-generated string of text and a chatbot-generated string of text (such as in the ChatGPT university essays) is that human beings are prone to attribute meaning and intelligence to inanimate things, even when they know that it is artificial and not thinking. Bender and Koller here cite the example of the ELIZA text-production program of the 1960s, which was not an ANN at all but a mere program using ranked pieces of sentences to produce answers to user input. Testers of ELIZA, however, treated ELIZA like a trusted confidant, even though they knew it was a computer program.[62] Human beings project thought and meaning onto text, even when they know that the text does not have an intelligence behind it that can speak meaningfully. Bender argues that the projection of intelligence onto AI/ML chatbots is one of the most dangerous

[59]Braine, *Human Person*, p. 353.
[60]Ibid., p. 471.
[61]Ibid., pp. 471–72.
[62]Bender and Koller, 'Climbing Towards NLU', 5188; Ned Block, 'Psychologism and Behaviorism', *The Philosophical Review*, 90 (1981), 8–10.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

applications of them, as the potential for malicious actors deliberately using them to deceive is an active and grave threat.[63]

Because current Natural Language Processing (NLP) chatbots do not have sources external to linguistic form, it is not possible for them to authentically predicate; instead, they merely produce statistically likely strings of text supported by unfathomably many statistical calculations and unfathomably large example sets. Premise (3) is therefore shown to be false for current chatbots.[64]

Premise (1), however, ought to be reevaluated as well. As mentioned by Jude Chua Soo Meng, if angels have the intelligible species of their knowledge infused directly by God, would AI/ML chatbots have intelligible species infused into them by training sets and therefore gain knowledge of externals through the direct hand of the programmer?[65] Even here, the connection of training data to text data is a statistical application of the training data and neither an abstraction nor an intuitive reception of knowledge. What Soo Meng calls 'sequential processing' by which computers operate and what we have seen above in the process of deep learning is not a grasp of the whole in an abstraction from the external, but an immense number of calculations to assess the likelihood of a part (or parts) of a training set being correlated with a part (or parts) of the test data. While there is a kind of predication and categorization happening in this example, the predication and categorization do not imply that the chatbot ever recognizes the whole of what is shown, instead using an ensemble of various parts to make its lexical or other correlations. Soo Meng rightly concludes that the intuitive infusion of forms by God into angels is a completely different and disanalagous process than the calculation of associations done by AI/ML programs.

6. Conclusion

The philosophical and theological tradition of St. Thomas Aquinas has significant resources for evaluating potential non-human intelligences, and contemporary articulations of that tradition rely on syntactical speech as a chief sign of true intelligence. This approach, while *prima facie* would include anything that can produce syntactically correct speech, in fact does not provide robust arguments for the immaterial intelligence of AI/ML chatbots. While this approach may be limited in its ability to

[63]Bender et al., 'Stochastic Parrots', 617–18.

[64]This conclusion also excludes some kinds of human speech production from the category of authentic predication. Having memorized some stock phrases in a language (such as for purposes of tourism), but without a real understanding of the individual words or syntax, would authentically have communicative intent and expression, but would not be authentically predicating; one does not really know the language. In such a case, one does have some authentically-predicated thought to share, but that thought is authentically-predicated in the speaker's original language before using the stock translation.

Another kind of exclusion involves early speech acquisition or the acquisition of a new kind of knowledge. A mere parroting of phrases one has heard in order to provoke a desired effect is not an authentic predication and is often prone to error. A toddler, for example, simply repeating 'yes' in an attempt to get food does not truly express what he is asking for. Political discourse or undergraduate papers may also be examples of parroting phrases without understanding in order to provoke a desired effect. Not all words produced by human beings have authentic predication and communicative intent.

[65]Jude Chua Soo Meng, 'Artificial Intelligence and Thomistic Angelology: A Rejoinder', *Quodlibet*, 3 (2001), 3–6.

https://doi.org/10.1017/nbf.2023.9 Published online by Cambridge University Press

discern what abstraction would look like in non-human intelligences, the current state of AI/ML technology falls far short of what the Thomistic tradition would look for in an immaterial intelligence. As such, a Thomistic analysis shows that the language production capabilities of current AI/ML chatbots are not evidence that they have immaterial souls.

Cite this article: Lazzari EM (2024). AI/ML Chatbots' Souls, or Transformers: Less Than Meets the Eye. *New Blackfriars* 105(1), 47–60. https://doi.org/10.1017/nbf.2023.9