AI ist Tott: Theory of Mind, LLMs, and Insights from Non-Dual Śaiva Tantra

Timo Brady
Machine Learning and Artificial Intelligence

November 27, 2023

> The function of the ordinary, feeble means of knowledge is to make apparent some previously unknown fact. Therefore, these are neither useful nor capable of establishing Awareness, which is independent, undivided, and continuously revealing itself.
>
> As it is said in the *Trikasāra* ('The Essence of the Trinity'):
>
> **'If a person desires to step on the shadow of his head with his own foot, he will find his head will never be in the same place of his foot. The power of the Point is similar.'**
>
> —*Pratyabhijñāhṛdayam* (Tr. Wallis 2017)

Preface: My Era of *Eat, Pray, Love*

In writing this paper, I was reminded of the kinds of philosophical discussions I had with a teacher at the foothills of the Himalayas in northern India. Every day was the same. Before sunrise we participated in *pūjā* (ritual offerings), wrapped in woolen blankets. As the doors of the horizon opened, we meditated on the banks of a rivulet fed by the waters of the mightier Ganga in the company of cows that carefully trod around us like the river that flows around a rock. There was a deep level of trust one had to cultivate by the water in those days, trusting that a cow would not lovingly crush you while your eyes were closed. After the daily yogic practice, I returned to the river to sit and contemplate. Later in the day, I described the contemplation and posed questions to him.

One day he said, "Contemplation is not thinking, and there is no contemplation without meditation. Elevated human experiences are usually without thought." I know this from my own experience. In those contemplations by the river, I invited a subtle intention—"I am willing to experience whatever needs to be experienced in this moment"—and then let it go, permitting whatever subtle energy was percolating beneath the *citta-vṛtti*, the mental-emotional fluctuations, to arise from a point of stillness.

What arose for me in that first month was essence nature. I became absorbed in observing the process of my body emit warm breath, the breath condense with the cold air, and the constituent parts of the breath dissolve back into the surrounding space. What remained of the breath? What remained of its essence that was now indistinguishable from the space and the air around me and now utterly without form, in such contrast to my own form of the body, of the rock upon which I sit? My form—the moment I am formless—the moment I am dissolved, devoured, and reabsorbed. What remains?

There were many other moments besides in which I experienced in my body ineffable truths about the nature of being and consciousness—truths that transcend intelligence. Consciousness itself cannot be made into an object of perception, subject to intellectual dissection and analysis. I cannot point to Awareness and say, "Ah ha! There it is!" I can only experience it—*be* it. Thus the allegory in the epigraph.

Introduction

> In the beginning was the Word. And the Word was programmed in classic binary. And the Word said, 'Let there be life!'
>
> —*Hyperion* (Simmons 1990)

In the present era, where access to more affordable Graphics Processing Units (GPUs) has catalyzed the explosive growth of AI, the development and application of Large Language Models (LLMs) are reshaping discussions around Theory of Mind tasks. Historically, Theory of Mind tasks have served as a litmus test for human cognitive capabilities in recognizing and attributing mental states, both in oneself and in others. The fact that these tasks are being adeptly managed by LLMs raises new questions.

This paper aims to present in brief a portion of the available literature regarding what the success of LLMs in Theory of Mind tasks reveals about the nature of these tasks themselves, in the human mind and in the plumbing of AI machines. Chomskian theories of language acquisition, which emphasize modularity, innateness, the poverty of stimulus, abstraction, and deep representation, are also discussed. Finally, insights from Non-Dual Śaiva Tantra (NST) contrast AI's mimicry of consciousness and Theory of Mind with the authentic experiences intrinsic to the human biologically-embodied experience.

The Nature of Theory of Mind Tasks

Theory of Mind tasks require a sophisticated representation of knowledge which extends beyond the accumulation of factual information, including understanding agency, recognizing subjectivity, and engaging in temporal and causal reasoning (Prinz 2006). Understanding of agency means understanding that actions

and beliefs are guided by internal mental states, not solely by observable external stimuli (Prinz 2006). Humans act based on their individual beliefs, desires, and intentions, which may not always align with a collective reality. Recognition of subjectivity requires acknowledging that each individual may have varying perspectives and beliefs about the same situation, a key aspect of social interaction and empathy (ibid.). Temporal and causal reasoning are the crucial abilities to comprehend how past actions influence present beliefs and predict future behaviors.

According to the Chomskian theories discussed in the lectures, cognitive functions are innate and modular. It follows that Theory of Mind is an independent cognitive module, specialized and distinct from other cognitive processes, evident in early childhood development (Slaughter 2015). The concept of poverty of stimulus also applies to Theory of Mind; the external inputs available to a developing child are insufficient for complex cognitive abilities without innate cognitive scaffolding, which allows for the formation of generalized concepts from limited data through abstraction and deep representation (Carruthers 2016).

Representation of Knowledge in Human Minds

Theory of Mind is a function of a complex interaction of various cognitive domains in human minds. Language and symbolic representation are integral in the development and expression of Theory of Mind (Apperly and Butterfill 2009). Chomsky theorized that language involves innate grammatical structures and is crucial for comprehending and articulating mental states. Social cognition and empathy are embedded in Theory of Mind, which relies on empathy and the ability to connect with others' emotional states. Certain brain regions, such as the prefrontal cortex, temporoparietal junction, and amygdala, demonstrate a degree of modularity in the neural implementation of Theory of Mind (Prinz 2006; Mahy, Moses, and Pfeifer 2014). Humans form abstractions of social scenarios and mental states, a process which is influenced by social and linguistic experiences. This enables more efficient navigation of complex social interactions, even in novel situations (Carruthers 2016). Challenges to the modular argument of Theory of Mind are emerging in light of more recent evidence from neuroscience (Gerrans and Stone 2008).

Representation of Knowledge in Artificial "Minds"

Did I request thee, Maker, from my clay
To mould me Man, did I solicit thee
From darkness to promote me...

—*Paradise Lost*

Are LLMs even appropriate models for human cognition in the first place, given their utterly distinct means of language acquisition (Milway 2023; Piantadosi 2023)? LLMs mimic Theory of Mind by relying on vast textual corpora and statistical patterns to derive probabilities regarding the nature of reality (ibid.). As such, LLMs generate responses that merely simulate the understanding of mental states (Manning et al. 2020). In contrast to human minds, modularity in LLMs emerges as algorithms and data structures. However, unlike human minds, there is no "preprogrammed" understanding of Theory of Mind. The success of LLMs in Theory of Mind tasks simply reflects their programming and training rather than any innate cognitive structure (Pulvermüller 2023). With regard to the poverty of stimulus—a wealth of stimulus in the case of LLMs—their form of abstraction lacks the deep representation observed in human cognition.

Insights from Non-Dual Śaiva Tantra

> The awareness of the knower and the known is common to all embodied beings; but for yogīs, this is the difference: they pay careful attention to the connection.
>
> —*Vijñāna-bhairava-tantra* (Tr. Wallis 2012)

Awareness is *a priori* fundamental to the universe from which all phenomena arise; that is to say, from the perspective of NST, all phenomena arise from within consciousness itself. It is an Idealist view in contrast to the Physicalist in which consciousness is merely an emergent quality of complex physical structures like the human brain. It is this latter point that merits further dissection, for there is no convincing evidence for why consciousness would emerge from a physical structure when viewed through an NST lens—or even a neuroscientific lens, for that matter. However, it does follow that Theory of Mind and other advanced forms of *cognition* emerge from processes produced by more complex brain structures, *pro* or *contra* a modular view notwithstanding.

The Shining

From the perspective of NST, Awareness has particular inherent capacities or powers (*śaktis*) that describe the fundamental nature of reality: *chit-śakti* (the power of consciousness), *ānanda-śakti* (the power of bliss), *icchā-śakti* (the power of will/creative impulse), *jñana-śakti* (the power of knowing/cognition), *kriyā-śakti* (the power of action), and the "meta-power" of *svātantrya-śakti* or the power of autonomous freedom (ibid.; Wallis 2017). But what does Awareness or Reality *do* (*kriyā*)? In one sense, Awareness causes the projection of all that is manifested. This is known as "shining" or *ābhāsa* in which various characteristics (or "shinings") of an object of perception (such as "tall,"

"short," "yellow," or "made of gold") are a vibration of the one indivisible light of Awareness (Wallis 2012).

Thus, according to *ābhāsa* theory, the mind is a synthesizer of all of these multitudinous "shinings" of the objects of perception, which appear to collapse into a single object. And these *ābhāsas*, these "shinings," include various subjective dimensions which inevitably vary from person to person, forming a multi-dimensional reality of an object of perception (ibid.). It follows that Theory of Mind in NST is this synthesizing of variously observed *ābhāsas*, the mental-emotional states of others, which "shine" as reality within the mind of the perceiver.

But how is *ābhāsa* theory applicable to LLMs? Indeed, structures of the brain responsible for language also appear to be responsible for the formation of Theory of Mind (at least in part), and LLMs seem to benefit from a simulacrum of this pseudo-architecture; however, no "shining" is occurring amongst the bit-flipping of 0s and 1s. The probabilistic output is a byproduct of processes occurring *post* shining—a shadow, if you will, alluded to in the preface. While LLMs operate in high-dimensional space with many billions of parameters, they never view reality *in toto*; instead, reality is viewed as dissectable parts that are probabilistically cobbled together to simulate and approximate the mind as synthesizer of the "shining" process; in contrast, biological cognition, grounded in Awareness, is much greater than the sum of its parts (*contra* the modular theorists), a reflection of a much deeper multi-dimensional reality which is part of an undissectable, continuously unfolding process.

Checkmate

This shortcoming of deep learning models, the continuous deconstruction of the whole and focused attention on the parts, are highlighted in a recent case study. Engineers were able to exploit such shortcomings in KataGo, a Convolutional Neural Network (CNN) which is based on DeepMind's superhuman AlphaGo, beating it several times (Wang et al. 2023). The vulnerabilities are easily noticeable by a human player who has a real world, total view of the Go board at all times and can detect any subtle attempt at diversion and subterfuge. The KataGo AI, in this case, failed to anticipate the players' deceptive strategy (Theory of Mind), and to, in turn, monitor the outer edges of the board for incremental movements, which it ultimately cannot do without further enhancements such as self-reflection and simulation capabilities (ibid.). In sum, these superhuman AI Go programs are not really learning Go itself at all (at least, not as we understand the game). Go—and all of the "shinings" around Go that make it "Go" in our minds as an object of perception—is not a reality that emerges within the plumbing of the AIs.

Epilogue: AI ist Tott

> I move in silence of clear white light. Everything around me is waiting. I dream of being alone on the top of a mountain, surveying the land around me, greens and yellows—and the sun directly above, pressing my shadow into a tight ball around my legs. As the sun drops into the afternoon sky, the shadow undrapes itself and stretches out toward the horizon, long and thin, and far behind me...
>
> —*Flowers for Algernon* (Keyes 1966)

In a previous paper for this course, I discussed AI systems as extensions of the human body-mind apparatus—like a spear or a vessel for water—shaped by the collective consciousness (and unconsciousness) of humanity. AI, in its most advanced form, will ultimately emerge as the most consequentially meta of all human technologies, and, with it, the Physicalists' paranoid fever dream of a spontaneously autonomous and despotic *deus ex machina*.

References

Apperly, Ian A. and Stephen A. Butterfill (2009). "Do Humans Have Two Systems to Track Beliefs and Belief-Like States?" In: *Psychological Review* 116.4, pp. 953–970.

Carruthers, Peter (2016). "Two Systems for Mindreading?" In: *Review of Philosophy and Psychology* 7.1, pp. 141–162.

Gerrans, Philip and Valerie E. Stone (June 2008). "Generous or Parsimonious Cognitive Architecture? Cognitive Neuroscience and Theory of Mind". In: *The British Journal for the Philosophy of Science* 59.2.

Keyes, Daniel (1966). *Flowers for Algernon*. First. New York: Harcourt, Brace & World.

Mahy, Caitlin E.V., Louis J. Moses, and Jennifer H. Pfeifer (2014). "How and where: Theory-of-mind in the brain". In: *Developmental Cognitive Neuroscience* 9, pp. 68–81.

Manning, Christopher D. et al. (Dec. 2020). "Emergent linguistic structure in artificial neural networks trained by self-supervision". In: *Proceedings of the National Academy of Sciences* 117.48.

Milway, Daniel (Apr. 2023). *A Response to Piantadosi (2023)*. [pdf]. lingbuzz/007264.

Piantadosi, Steven (Nov. 2023). *Modern Language Models Refute Chomsky's Approach to Language*. [pdf]. lingbuzz/007180.

Prinz, J. J. (2006). "Is the Mind Really Modular?" In: *Contemporary Debates in Cognitive Science*. Ed. by R. J. Stainton. Blackwell Publishing, pp. 22–36.

Pulvermüller, Friedemann (2023). "Neurobiological mechanisms for language, symbols and concepts: Clues from brain-constrained deep neural networks". In: *Progress in Neurobiology* 230.

Simmons, Dan (1990). *Hyperion*. New York, NY: Bantam Doubleday Dell Publishing Group.
Slaughter, Virginia (2015). "Theory of Mind in Infants and Young Children: A Review". In: *Australian Psychologist* 50.3, pp. 169–172.
Wallis, Christopher (2012). *Tantra Illuminated*. Boulder, CO: Mattamayūra Press.
— (2017). *The Recognition Sutras*. Boulder, CO: Mattamayūra Press.
Wang, Tony T et al. (2023). "Adversarial Policies Beat Superhuman Go AIs". In: *International Conference on Machine Learning*.