MARKTECHPOST NewsHub

Google DeepMind Presents a Theory of Appropriateness with Applications to Generative Artificial Intelligence

By **Sana Hassan** - January 4, 2025

Appropriateness refers to the context-specific standards that guide behavior, speech, and actions in various social settings. Humans naturally navigate these norms, acting differently based on whether they are among friends, family, or a professional environment. Similarly, AI systems must adapt their behavior to fit the context, as the standards for a comedy-writing assistant differ from those of a customer-service representative. A critical challenge is determining what is appropriate in a given situation and how these norms evolve. Since humans ultimately judge AI behavior, understanding how appropriateness influences human decision-making is essential for evaluating and improving AI systems.

The concept of appropriateness also plays a central role in the emerging domain of generative AI. All socially adept actors—human or machine—must moderate their behavior based on the context and community in which they operate. This parallels the content moderation challenges digital communities face, where moderators enforce explicit rules and implicit social norms. Generative AI systems face a similar task: regulating the content they generate to align with contextual appropriateness. However, standards of appropriateness vary between individuals and within the same individual across different situations. For example, a teaching assistant chatbot must behave differently from one designed for a mature-rated game. This highlights the complex and dynamic nature of appropriateness, which remains critical as AI expands into physical, cultural, and institutional domains traditionally dominated by human intelligence.

Researchers from Google DeepMind, Mila – Québec AI Institute, University of Toronto, and the Max Planck Institute introduce a "theory of appropriateness," examining its role in society, neural underpinnings, and implications for responsible AI deployment. It explores how AI systems can act appropriately across diverse contexts, emphasizing norms that guide human behavior. The paper conceptualizes appropriateness as a

dynamic, context-dependent governance mechanism for societal cohesion. Departing from traditional alignment frameworks, it critiques oversimplified moral core assumptions, advocating for AI to adapt to the pluralistic, evolving norms shaping human interactions rather than seeking a universal moral consensus.

The study introduces a computational model to elucidate how humans determine appropriate behavior across various contexts. It posits that individuals utilize a pattern completion mechanism, drawing from memory and situational cues to predict suitable actions. This process involves a global workspace that integrates sensory inputs and past experiences, facilitating decision-making. The model also considers the role of social conventions and norms, highlighting how collective behaviors influence individual appropriateness judgments. By understanding these mechanisms, the research aims to inform the development of generative AI systems that can navigate complex social environments responsibly.

The work frames human behavior and societal cohesion not through alignment but appropriateness, emphasizing that societies are maintained through conflict resolution mechanisms rather than shared values. The study presents a decision-making model that contrasts reward-based approaches, highlighting that appropriateness in human behavior emerges from a blend of societal influences. This model differentiates between explicit norms (articulated in language) and implicit ones (embodied in the brain's patterns), which can guide interactions between humans and AI systems, especially in context-sensitive tasks.

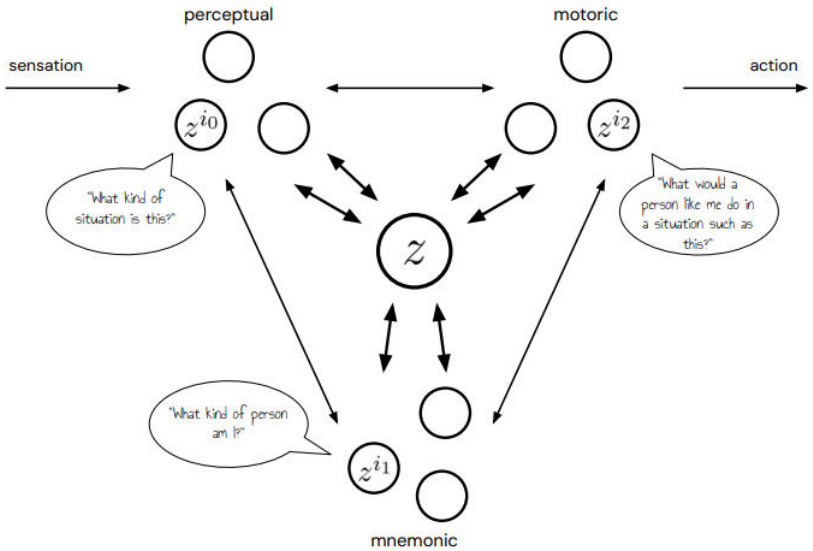

Figure 2 | z denotes the content represented by a set of parallel specialized summary functions which may correspond to neural circuitry located in different parts of the brain from one other or even themselves consist of distributed representations. For instance, some summary functions may be perceptual in nature e.g. a summary function that asks of recent observations "what kind of situation is this?", some summary functions may be more mnemonically oriented e.g. a summary function that asks of one's episodic memory "what kind of person am I?", and some summary functions may be closer to premotor action planning circuitry such as one that asks "what would a person like me do in a situation like this?". This architecture was inspired by the

global workspace architecture of Baars (1988); Shanahan (2010). Here, at time t, our z_t is a snapshot of the content in the global neuronal workspace, i.e. z_t is represented by dynamic cell assemblies linking the far-flung modules comprising the workspace perhaps by oscillating coherently with one another (Dehaene et al., 1998; Fries, 2015).

The study calls for careful consideration when designing generative AI systems by recognizing that appropriateness is context-dependent and deeply linked to societal norms. It highlights that while AI lacks human-like context awareness, understanding appropriateness is vital for its responsible use. The paper also suggests that AI may eventually need specific legal frameworks similar to corporate personhood to address ethical and operational issues, particularly as AI systems become more autonomous. This underscores the importance of cognitive science in shaping AI governance and ensuring it aligns with societal expectations.

Check out ***the Paper***. All credit for this research goes to the researchers of this project. Also, don't forget to follow us on **Twitter** and join our **Telegram Channel** and **LinkedIn Group**. Don't Forget to join our **60k+ ML SubReddit**.

🚨 **FREE UPCOMING AI WEBINAR (JAN 15, 2025): Boost LLM Accuracy with Synthetic Data and Evaluation Intelligence–Join this webinar to gain actionable insights into boosting LLM model performance and accuracy while safeguarding data privacy.**

Sana Hassan

+ posts

Sana Hassan, a consulting intern at Marktechpost and dual-degree student at IIT Madras, is passionate about applying technology and AI to address real-world challenges. With a keen interest in solving practical problems, he brings a fresh perspective to the intersection of AI and real-life solutions.