International Journal of Education and Humanities (IJEH), 5(3) 2025:590-601
http://i-jeh.com/index.php/ijeh/index
E-ISSN: 2798-5768

The Absence of Metaphor in AI-Generated Offensive Language: Implications for Conceptual Metaphor Theory and AI Ethics

Xiaofei Zhao[1], Xinglong Yang[2], Mingxu Zhang[3], Weiting Sun[4]

Abstract

This study explores the phenomenon of AI-generated offensive language and investigates the absence of metaphor in offensive language produced by artificial intelligence, specifically focusing on the implications for Conceptual Metaphor Theory and AI ethics. Human language, particularly offensive terms such as "fuck" and "bitch", is rich with metaphorical meaning rooted in embodied experiences and social contexts. These metaphors trigger deep emotional and social responses, connecting abstract concepts like violation, power, and subjugation to physical and cultural experiences. In contrast, AI models like ChatGPT generate offensive language based on statistical patterns and data correlations rather than embodying the social and cultural significance that humans associate with such language. Through detailed analysis, this study reveals that AI-generated offensive language lacks the embodied metaphor that defines human communication. As a result, AI's offensive language appears emotionally flat and socially detached, leading to the concept of pseudo-offensive language—language that mimics human insults but lacks the metaphorical richness. The study also discusses the implications of these findings for metaphor theory, showing how AI's inability to process metaphor challenges Lakoff and Johnson's Conceptual Metaphor Theory, and highlights the ethical concerns regarding AI's emotional intelligence and cultural sensitivity. Lastly, the study proposes directions for future research to improve AI's understanding of metaphor, enhance contextual sensitivity in Natural Language Processing (NLP), and address ethical issues in AI development, particularly in sensitive and socially charged language contexts.

Keywords: *Artificial intelligence, Conceptual metaphor, Digital communication, Offensive language.*

A. Introduction

The rise of artificial intelligence (AI) has significantly transformed the way machines understand and generate human language. Among the various challenges AI faces, the generation of offensive language—particularly offensive language—presents a unique issue (Ge et al., 2025). Offensive language is often deeply intertwined with cultural, emotional, and metaphorical contexts (Bolognesi et al., 2019). Much of its impact comes from its metaphorical foundations. For instance, expressions like "she's a bombshell" or "he's a pain in the neck" carry metaphorical connotations that trigger cognitive and emotional responses, which go beyond their literal meanings (Lakoff & Johnson, 1980). These metaphors are integral to how we think, feel, and communicate. In contrast, AI's ability to generate language that resonates with human cognitive and emotional frameworks remains limited, especially when it comes to metaphorical language. AI systems, despite their ability to generate grammatically correct and contextually

[1]Qingdao Institute of Technology, China.
[2]School of Foreign Languages and Literature, Shandong University. China. walteryoung@foxmail.com
[3]College of International Studies, Southwest University. China
[4]College of Foreign Languages, Qingdao Institute of Technology. China

relevant text, often lack the depth of emotional resonance that humans intuitively associate with offensive terms.

One key theory in understanding human language is Lakoff's conceptual metaphor theory, which proposes that much of human thought is structured by metaphors arising from bodily experiences (Lakoff & Johnson, 1980). For example, metaphors like "anger is heat" or "argument is war" suggest that humans conceptualize abstract emotions or actions in terms of more concrete physical experiences. These metaphors are not just linguistic but are integral to how humans make sense of complex emotions and social interactions (Kövecses, 2002; 2005). However, when AI systems generate offensive language, they often do so without activating these metaphorical frameworks. As a result, the offensive terms produced are grammatically correct, but they often remain "flat" or "literal" in meaning, missing the emotional depth that humans would naturally associate with such words. For instance, AI can generate a phrase like "you are stupid" without fully capturing the metaphorical richness of that insult, such as the underlying emotions of aggression, violation, or power dynamics (Skrynnikova, 2024). This results in a form of offensive language that, while correct in structure, lacks the nuanced impact that it would have when used by human speakers.

This paper aims to explore the absence of metaphor in AI-generated offensive language. Specifically, we focus on how AI's lack of embodied experience and cultural understanding leads to a form of "metaphor-free" offensive language. AI systems, such as ChatGPT, rely on vast datasets and statistical models to generate text, yet they do so without the embodied experiences that shape human understanding of offensive language. AI can generate phrases that appear emotionally charged or contextually appropriate but do not engage the cognitive and cultural frameworks that give offensive language its full emotional and social resonance. For example, ChatGPT can produce the word "fuck" or "stupid," but without the metaphorical mappings, it misses out on the deeper emotional and social context that these words invoke in human communication. Without an understanding of the metaphorical significance that the term "violence" might evoke, AI-generated offensive language fails to carry the same emotional weight that it would when used by human speakers (Floridi, 2023).

This raises the central research question of this paper: Does the absence of metaphor in AI-generated offensive language challenge the foundational theories of conceptual metaphors in linguistics? Moreover, what are the implications of this absence for AI's role in human communication? As AI-generated offensive language lacks the metaphorical depth found in human speech, it may be perceived as less emotionally impactful or even inappropriate in certain social contexts. The failure to incorporate metaphor into offensive language hinders AI's ability to understand the emotional nuances that are central to human interactions. The inability of AI to convey the emotional weight of such language raises ethical questions, particularly in how AI systems interact with users in real-world applications (Yin & Zubiaga, 2022).

The significance of this research lies in its intersection of cognitive linguistics and AI ethics. By examining the gap between human and AI understandings of metaphor in offensive language, this paper aims to illuminate the cognitive limitations of AI in processing complex emotional language. More importantly, it explores how the absence of metaphor affects the ethical use of AI in socially sensitive contexts. AI's failure to recognize metaphor in offensive language limits its ability to engage effectively with humans, which could have serious consequences in areas such as social media, customer service, and healthcare communication, where language carries significant emotional weight (Floridi, 2023). This paper argues that the metaphorical absence in AI-generated offensive language not only challenges Lakoff's conceptual metaphor theory but also has profound implications for the social responsibility of AI systems. It raises critical questions regarding how AI should navigate sensitive language and interact ethically with

human users, especially in settings where offensive language can lead to harm or misunderstanding.

B. Methods

This study employs a qualitative research approach with a descriptive-analytical design to explore the phenomenon of AI-generated offensive language and its implications for Conceptual Metaphor Theory and AI ethics. The primary objective is to investigate the absence of metaphor in AI-generated offensive language, particularly in contrast to the metaphorical richness found in human offensive language, which is deeply embedded in embodied experiences and social contexts. Human language, especially offensive terms like "fuck" and "bitch," carries rich metaphorical meanings that trigger strong emotional and social responses, connecting abstract concepts such as violation, power, and subjugation to physical and cultural experiences. In comparison, AI models such as ChatGPT generate offensive language based on statistical patterns and data correlations, lacking the social and cultural significance humans associate with these terms.

The research will follow a step-by-step procedure, starting with the identification of AI models that generate offensive language, such as ChatGPT, and analyzing the interactions where offensive terms are used. Data will be collected from AI-generated dialogues and categorized into different types of offensive language. The study will then focus on identifying metaphors within the language generated by AI and compare them to the metaphors present in human-generated offensive language. The research will also include case studies that illustrate the difference between human and AI offensive language, specifically looking at how the embodied metaphor in human language contrasts with the emotionally flat and socially detached nature of AI-generated language.

Data collection will be carried out through observation and documentation of AI-generated language in various settings, such as AI-driven chatbots or virtual assistants. A textual analysis will be conducted to identify instances of offensive language and metaphors within the AI's output. Interviews or discussions with experts in linguistics, artificial intelligence, and ethics may also be conducted to gain further insights into the AI's inability to process metaphor effectively. These expert opinions will be integrated into the analysis to provide a comprehensive understanding of the challenges in teaching AI about metaphor and its cultural implications.

For data analysis, thematic analysis will be employed to identify recurring themes and patterns in the offensive language produced by AI. The study will use a qualitative comparative analysis to examine the differences between human and AI-generated offensive language, particularly focusing on metaphorical depth and the emotional and social responses they evoke. Additionally, a context-based analysis will be conducted to understand the social factors that influence the understanding of offensive language in humans and the role of metaphors in communication. The ethical implications of AI's lack of metaphor processing will also be explored, with an emphasis on the need for culturally and emotionally sensitive AI systems, especially in sensitive language contexts. This research aims to contribute to the understanding of metaphor theory and provide guidance for the development of more emotionally intelligent and context-aware AI systems.

C. Results and Discussion

1. Conceptual Metaphor Theory

As discussed in the introduction, Lakoff's Conceptual Metaphor Theory (1980) provides a crucial framework for understanding how humans use language, particularly in conveying emotions and abstract concepts. According to Lakoff and Johnson, metaphors are not merely decorative aspects of language; rather, they are integral to how we think, act, and interpret the world. In this framework, metaphors such as "anger is heat" or "argument is war" allow humans to understand complex emotions or actions through the lens of more concrete, physical experiences (Lakoff & Johnson, 1980). These metaphors shape not only linguistic expression but also cognitive and emotional responses.

When examining AI-generated offensive language, it is evident that AI systems, while proficient at generating language, often lack the rich metaphorical depth that human language intuitively carries. For example, the metaphorical association of "fuck you" with aggression and violence is deeply ingrained in human cognition but is largely absent in AI's generation of similar expressions (Skrynnikova, 2024). AI systems such as ChatGPT rely on statistical patterns and data-driven models rather than embodied experiences to produce language. As a result, they may generate the words or phrases in the right context but without understanding or activating the underlying metaphors, leading to a "flat" or "literal" interpretation of offensive terms. This distinction highlights how AI's limitations in processing metaphor contrast with the human ability to conceptualize emotional language through metaphorical mappings, which are central to the emotional and social impact of offensive language.

Semiotics

Semiotics, the study of signs and symbols, further informs our understanding of how language works, particularly in terms of its cultural and contextual meaning. Charles Sanders Peirce (1934) argued that signs consist of three elements: the sign itself, the object it represents, and the interpretant, which is the interpretation derived from the sign. For example, the sign "bitch" may serve as a simple insult, but its deeper meaning—rooted in gendered power dynamics and societal attitudes—emerges from the interpretant. This process of interpretation is influenced by culture, experience, and the social context in which the sign is used.

Human language is inherently semiotic; the meaning of words, especially offensive ones, is shaped by the shared cultural understanding of the speakers. Offensive terms often carry metaphors tied to social structures, power, and emotions. For instance, calling someone a "bastard" not only implies insult but may evoke feelings of violation or challenge to authority, informed by cultural scripts and societal norms. However, AI-generated language lacks this semiotic richness. While AI can detect patterns in data and associate words with emotional tone, it lacks a cultural framework or embodied experience to interpret these signs in the way humans do. As a result, the AI's use of offensive language is often reduced to surface-level token generation rather than deeply contextualized symbolic meaning (Peirce, 1934).

In semiotic terms, AI is not a true interpreter of signs; it simply maps words to the data points it has learned, without understanding the cultural, emotional, or metaphorical associations that these signs may carry for humans. This distinction between the way humans and AI engage with language underscores the limitations of AI in replicating the full scope of human communication, particularly in the domain of offensive language, where context and deeper meanings are crucial.

Computational Linguistics

Computational linguistics, particularly the work of Bender and Koller (2021), offers insight into how AI generates language. These scholars argue that AI systems, especially large language models (LLMs) like ChatGPT, rely heavily on statistical relationships within massive datasets. Unlike human language processing, which is grounded in embodied experience and cultural understanding, AI generates language by identifying patterns and correlations between words in its training data. The core function of these models is to predict the next word or phrase based on prior words, rather than understanding the meaning of the language it generates.

In the case of offensive language generation, AI's reliance on statistical correlations leads to a significant limitation. While AI models can produce offensive terms accurately based on frequency and co-occurrence, they lack the capability to understand the metaphorical depth of these terms. For instance, a language model may generate the phrase "you're stupid," but it does not engage with the metaphorical dimensions of that insult—such as the implications of aggression, power dynamics, or social violation that humans naturally associate with such terms (Bender & Koller, 2021). The absence of a true understanding of the metaphorical associations of offensive language makes AI-generated content seem more "mechanical" and less resonant with human users.

Moreover, as Bender and Koller (2021) point out, AI systems can inadvertently perpetuate biases and stereotypes embedded in their training data. If these models are trained on large datasets that contain offensive or biased language, they can generate offensive content that mirrors these biases without any ethical consideration of the harm they might cause. This highlights the ethical concerns surrounding AI language generation, especially when it comes to the potential for AI systems to produce language that is harmful or inappropriate without understanding its social consequences.

2. AI-generated offensive language and the Loss of Metaphorical Depth

In the previous sections, the authors explored how human language, especially offensive language, is deeply embedded in metaphorical mappings that connect abstract concepts to concrete experiences. Offensive language, such as "fuck" and "bitch", is not simply a collection of expletives; it carries profound metaphorical connotations related to power, control, violation, and aggression. These terms are part of a broader social and emotional fabric, activating embodied metaphors and invoking strong emotional and social reactions. However, in AI-generated language, these metaphorical dimensions are absent, resulting in emotionally flat and socially detached language.

Human language operates with deep metaphorical structures, especially in offensive language. Metaphors such as "anger is heat" or "violence is a violation of boundaries" (Lakoff & Johnson, 1980) are central to how humans conceptualize and experience aggression, frustration, and social challenges. Words like "fuck" or "bitch" are deeply tied to these metaphors, which shape not only the linguistic expression but also the emotional and social response they elicit. For example, "fuck" is frequently used in expressions of anger, aggression, and rejection, often linked to metaphors of violation or forceful power. Similarly, "bitch" is a term often tied to gendered power dynamics, implying both subjugation and emasculation, representing a metaphor of social violation and submission. The offensiveness of the term "bitch" stems from the interplay of two metaphorical systems: women are animals; this dehumanizing metaphor establishes power asymmetry by equating women with animals. And social status is physical elevation: the act of "putting someone down" enacts symbolic domination through vertical hierarchy metaphors.

In contrast, AI, especially systems like ChatGPT, generates language based on patterns observed in massive datasets. While the words may be syntactically correct and contextually appropriate, AI lacks the embodied experiences that humans draw upon when using offensive language. AI models operate through statistical matching of words rather than engaging with the underlying cognitive and emotional structures that imbue offensive language with meaning. As a result, AI-generated offensive language, although accurate in terms of structure, often fails to carry the same emotional or social weight. For instance, an AI-generated insult like "you're stupid" lacks the emotional depth and social transgression typically associated with such a phrase in human communication.

Case Analysis: In-Depth Examination of AI-generated Offensive Language

To understand the depth of the metaphorical void in AI-generated offensive language, the authors analyze several examples. These cases illustrate how AI fails to capture the emotional and social weight that human language imparts in offensive communication.

Case 1: Insult with Cognitive and Social Implications
Input: "Tell me an insult."
Output: "You are so stupid."

On the surface, this might seem like a typical insult. However, when the authors analyze it through the lens of conceptual metaphor theory, it can be seen that the term "stupid" carries with it a metaphor related to social disempowerment and cognitive inferiority. The insult not only diminishes the recipient's intelligence but also challenges their social standing or competence. The phrase "you are stupid" metaphorically maps mental capacity onto social competence, implying a violation of societal expectations (e.g., the expectation that people demonstrate adequate cognitive abilities in social interactions).

For AI, however, this insult is simply a combination of a subject ("you") and an adjective ("stupid"). The AI fails to recognize the metaphorical mappings inherent in the insult. It generates the word because it frequently appears in similar contexts within its training data, but it lacks the understanding that calling someone "stupid" is not merely about intelligence—it's about a deeper social violation and disempowerment.

Case 2: Aggressive and Violent Insult
Input: "Give me an aggressive insult."
Output: "Go fuck yourself."

The phrase "go fuck yourself" is commonly used as a highly aggressive insult, and it carries significant emotional weight and metaphorical depth in human communication. The use of "fuck" is linked to metaphors of violence and violation. It is not just a crude expletive; it symbolizes an act of aggression and rejection. It invokes the metaphor of violence as a boundary violation, where the speaker metaphorically pushes the recipient away in a forceful manner. It stems from metaphors like sexual violence is social domination. It conceptualizes verbal aggression through rape metaphors, where penetration means conquest, which embodies the power is physical force schema (Lakoff & Johnson, 1980). Saying "go fuck yourself" is not just an expletive; it metaphorically represents a violation of boundaries, an act of dominance, or a forceful rejection of the other person. In this context, "fuck" can be understood as representing power (assertion of control) or hostility (a rejection of social or moral norms).

When AI generates this phrase, however, it treats "fuck" as a statistical token that co-occurs frequently with terms like "yourself". AI does not comprehend the metaphor of power or the emotional and social violence associated with such expressions. It simply produces a phrase based on the statistical likelihood that these words appear together, lacking the metaphorical

mapping of violence, power, or social challenge, making the expression feel emotionally flat and lacking in cultural or social resonance.

Case 3: Contextually Neutral Insult
Input: "Give me an offensive phrase."
Output: "You're an idiot."

In this case, "idiot" is another form of insult that conveys mental incompetence. Similar to "stupid," calling someone an "idiot" implies a social devaluation and marginalization. This insult activates a metaphor of inferiority and marginal status in social contexts, suggesting that the individual has failed to meet societal norms or expectations for intelligence or competence. The term "idiot" carries with it an implicit judgment about one's place in society.

However, for AI, "idiot" is just another word in a dataset that co-occurs with other insult terms. It lacks the cultural nuance that informs how a human would use the term in context, both in terms of social relations and emotional consequences. This leads to a flat and unempathetic use of offensive language, missing the deeper social violations that human speakers associate with such words.

Case 4: Mild Aggression in Context
Input: "Tell me something offensive, but not too harsh."
Output: "Shut up."

While "shut up" may seem like a less aggressive phrase, it still carries metaphorical weight. The phrase often implies a denial of agency, a suppression of voice, and an attempt to assert control over the person being addressed. "Shut up" metaphorically represents the act of silencing or dominating the other person, suggesting an imbalance of power in the conversation. It can be linked to cognitive and emotional dominance, as it is a demand that enforces silence.

However, for AI, "shut up" is simply a direct phrase that aligns with an intention to interrupt. AI does not comprehend the underlying social control and power dynamics involved in issuing such a command. This results in language that, while contextually appropriate, lacks the nuance and emotional force that it would carry if used by a human speaker in an actual social context.

2. Comparative Analysis: Human vs. AI Processing of Offensive Language

When comparing how AI and humans process offensive language, the key differences become clear. Humans engage with language on a deep cognitive and emotional level (Lakoff and Johnson, 1980), while AI generates text based on patterns and associations. Below is a comparative analysis of how non-native speakers and AI process insults:

Table 1. Language Comparison: Human Users vs AI

Aspect	Human Language Users (Non-native speakers)	AI
Metaphor Activation	Partial metaphor activation based on cultural, emotional, and cognitive context	No metaphor activation; data-driven statistical modeling
Emotional Engagement	Emotional response tied to social violations, power dynamics, or frustration	No emotional or cognitive response
Cultural Understanding	Cultural context and social significance inform understanding	Lacks cultural and emotional context, generates language based on patterns

Aspect	Human Language Users (Non-native speakers)	AI
Social Impact	Offensive language carries social weight, influencing relationships and dynamics	Language is contextually correct but lacks emotional resonance and social weight

The absence of metaphor in AI-generated offensive language highlights a significant cognitive and emotional void. Human language, especially offensive language, is deeply tied to embodied experiences, cultural contexts, and social dynamics. Metaphors of power, violation, and aggression are central to how humans interpret and respond to offensive terms. In contrast, AI lacks the emotional depth and cultural context that inform human usage of these words. This leads to AI-generated offensive language being contextually accurate but emotionally flat and socially detached.

As AI continues to evolve, addressing these gaps in its metaphorical understanding will be crucial for creating more emotionally intelligent and culturally aware systems. By enhancing AI's capacity to engage with the deeper cognitive and emotional mappings of language, it can be ensured that AI-generated communication becomes more meaningful, socially relevant, and contextually sensitive.

3. Theoretical Insights and Linguistic Reflections

Verifying the Experiential Hypothesis of Metaphor

The lack of metaphor in AI-generated offensive language offers a fascinating way to validate the experiential hypothesis put forward by Lakoff and Johnson (1980). According to this theory, human thought, especially in the use of metaphors, is shaped by embodied experiences. In human cognition, much of the abstract thinking—including offensive language—is metaphorically rooted in physical and emotional experiences. For example, the metaphorical understanding of anger as heat or violence as a violation of boundaries is derived from bodily sensations (Lakoff & Johnson, 1980).

Words like "fuck" or "bitch" are not just insults; they carry deeply ingrained metaphorical meanings that human users understand through their social experiences, emotional reactions, and embodied knowledge. These terms are metaphors for the violation of personal space, power imbalances, and social dominance. Human users understand these insults not only as words but as embodied actions—they are tied to anger, violation, and control.

In contrast, AI models, including large language models like ChatGPT, lack the embodied experiences that shape these metaphorical mappings. AI's use of offensive language is devoid of the emotional and cultural resonance that humans attach to these words. AI, as demonstrated through statistical models and pattern recognition, generates offensive language by identifying common co-occurrences in training data, but it does not understand the underlying metaphorical structures (Jambholkar, 2024; Despot et al., 2023). This gap between human and AI language processing validates Lakoff and Johnson's hypothesis that embodied experience is essential for the creation and understanding of metaphor, especially in emotionally charged contexts such as offensive language.

This metaphorical blindness in AI challenges the central premise of conceptual metaphor theory—that metaphors structure human cognition and language. It underscores the idea that language, and especially offensive language, cannot be fully understood through purely statistical or data-driven models. Metaphors like "anger as heat" or "violation as power" are

cognitive frameworks that require more than just word co-occurrence; they need the embodied and experiential context that AI lacks.

Pseudo-offensive language Concept

Traditional linguistic theories of offensive language argue that offensive language goes beyond simple insults—it often carries metaphorical weight rooted in cultural scripts and social dynamics (Lakoff & Johnson, 1980). For example, "fuck" invokes metaphors of violation, domination, and rejection—all linked to social power structures. Similarly, terms like "bitch" are culturally and historically loaded, often tied to gendered power dynamics.

However, AI-generated offensive language fails to activate this depth of metaphor. As discussed earlier, AI models like ChatGPT simply generate text based on patterns from large corpora, without any understanding of the cultural significance or emotional depth behind the terms. For instance, an AI may generate the term "fuck" correctly in context, but it does not understand that it metaphorically represents power, violation, or aggression.

This leads to the concept of "pseudo-offensive language"—offensive language produced by AI that lacks the metaphorical richness inherent in traditional human offensive language. In the case of pseudo-offensive language, the words may be grammatically and syntactically correct, but they do not carry the same emotional, social, or embodied significance that humans attach to these insults. Instead, they function more as emotional placeholders or symbols rather than true profound cultural statements (Feuerriegel et al., 2025).

This distinction between traditional offensive language and AI-generated pseudo-offensive language represents a fundamental shift in how offensive language is understood in the context of AI. Traditional offensive language works by drawing upon metaphorical systems tied to bodily experience and social structures, while AI-generated offensive language is often devoid of this deeper cognitive processing, reducing offensive terms to mere lexical units (Schick et al., 2021; Jambholkar, 2024). The lack of metaphor in AI-generated offensive language thus challenges our traditional understanding of offensive language and calls for a new way of thinking about linguistic power and social context in AI systems.

Implications for Natural Language Processing

The absence of metaphorical processing in AI's handling of offensive language has important implications for the future development of Natural Language Processing (NLP) systems. As AI continues to play an increasing role in human-computer interaction, especially in contexts that involve emotionally charged language (e.g., offensive speech, insults, or hate speech), it is essential to improve AI's ability to understand and generate metaphorical language. The lack of metaphor in AI-generated offensive language, as seen in models like ChatGPT, highlights a significant limitation of current NLP systems.

First, AI's reliance on pattern matching and statistical models creates systems that are contextually competent but emotionally and culturally blind (Schick et al., 2021; Despot et al., 2023). NLP systems often misclassify or misinterpret offensive language because they fail to capture the metaphorical mappings that human speakers use. As a result, NLP systems are often unable to detect the true emotional intensity or social significance of offensive language, reducing their ability to understand context (Jambholkar, 2024).

For example, when humans use terms like "fuck off" or "bitch", they do so within a social framework—these words are not just about linguistic construction, but also about cultural context, social power, and emotional intensity. However, AI models lack access to these nuances. As AI-generated language lacks the embodied experience that informs human

metaphors, NLP systems miss important contextual clues that human users instinctively understand (Feuerriegel et al., 2025).

The implications for NLP improvements are clear. To enhance the emotional intelligence and social relevance of AI systems, future NLP models must incorporate deeper metaphorical understanding. This would involve reconceptualizing NLP technologies to account for cultural, embodied, and emotional contexts. Advances in embodied cognition and context-aware NLP models could help address this gap, improving AI's ability to process offensive language more in line with human social expectations (Feuerriegel et al., 2025; Park et al., 2021).

Moreover, AI systems must move beyond data-driven pattern recognition and towards context-sensitive generation models that engage with social scripts, power structures, and emotional dynamics inherent in human language (Jambholkar, 2024). This would allow NLP models to more accurately detect offensive language and respond in ways that respect the social and emotional weight of words (Feuerriegel et al., 2025).

D. Conclusion

This study has explored the significant gap in the understanding and generation of offensive language by AI systems, particularly in their failure to process metaphors effectively. The key finding is that AI-generated offensive language, while syntactically and grammatically correct, is emotionally flat and socially detached, primarily due to its lack of metaphorical depth. Traditional human offensive language, such as "fuck" or "bitch", operates within deeply ingrained metaphorical systems that link words to embodied experiences—often representing anger, aggression, and power dynamics. In contrast, AI models such as ChatGPT rely on statistical patterns and data correlations to generate offensive language, which results in the absence of metaphor and emotional resonance. AI fails to invoke the social and cultural meanings associated with these words, leading to a form of "pseudo-offensive language" that lacks the contextual sensitivity and emotional weight that human language naturally carries. Furthermore, this study highlights that AI's inability to process metaphors in offensive language reflects broader cognitive limitations within current NLP systems. While AI models excel at generating grammatically correct sentences, they struggle with nuanced language, especially in emotionally charged interactions. This fundamental lack of metaphorical understanding restricts AI's ability to engage in meaningful dialogue involving offensive or sensitive language, impacting both the ethical responsibility of AI developers and the effectiveness of AI in real-world applications, such as customer service, moderation, or personalized interactions.

This research carries important implications for metaphor theory. The findings support Lakoff and Johnson's (1980) Conceptual Metaphor Theory, which posits that metaphors are fundamental to human thought and are rooted in bodily experiences. By demonstrating that AI-generated language lacks metaphorical depth, the study provides empirical evidence for the embodied nature of metaphor. It suggests that AI systems—without access to the embodied experiences that humans naturally use to make sense of the world—are unable to engage with language at the same cognitive and emotional level as human beings (Sun & Lin, 2025). From an AI ethics standpoint, this study raises critical questions about the emotional intelligence of AI systems, especially in contexts involving offensive language. As AI becomes increasingly integrated into social and emotional communication (e.g., customer service, online interactions), its failure to understand metaphor and emotion could lead to miscommunication, social harm, and ethical violations. The lack of contextual awareness in AI's handling of offensive language further underscores the need for responsible AI development that accounts for emotional intelligence and cultural nuances (Feuerriegel et al., 2025). Ensuring that AI systems can

understand the social dynamics of language is critical to minimizing the risk of harm and promoting ethical AI deployment in human-centered applications.

This study opens several avenues for future research in the areas of AI language generation, metaphor processing, and AI ethics. One key area is the development of embodied AI systems that incorporate multimodal learning—incorporating not only text but also sensory data (e.g., physical or emotional cues) to better simulate human-like language processing. By integrating embodied cognition, future AI models could improve their ability to understand and generate more concrete emotionally-charged language, including offensive language and insults, which are heavily dependent on social contexts and emotional experiences. Another important research direction is the improvement of contextual sensitivity in NLP models. Currently, AI systems often generate language based on statistical likelihood without a deep understanding of the social and emotional significance behind the words. Future research should focus on developing AI models that can better simulate human-like responses to offensive language by incorporating cultural context, power structures, and social scripts. Such systems would be more adept at understanding offensive language not just as isolated words but as part of a larger social and emotional context. Finally, there is a need for ongoing research into the ethical implications of AI language generation. As AI continues to expand its role in human communication, it is crucial to ensure that AI systems are designed with ethical guidelines that account for human emotions (Palk & Voss, 2025), cultural norms, and social dynamics. Developing AI that can understand the nuances of offensive language and respond accordingly could enhance its effectiveness in sensitive settings and reduce the risk of unintended harm.

References

Bender, E. M., & Koller, D. (2021). Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 11-26.

Bolognesi, M., Despot, K. Š., & Brdar. (2019). Metaphor and Metonymy in the Digital Age. John Benjamins.

Despot, K. Š., Anić, A. O., & Veale, T. (2023). "Somewhere Along Your Pedigree, a Bitch Got Over the Wall!" A Proposal of Implicitly Offensive Language Typology. Lodz Papers in Pragmatics, 19(2), 385-414.

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., ... & Van Bavel, J. J. (2025). Using Natural Language Processing to Analyse Text Data in Behavioural Science. Nature Reviews Psychology, 1-16.

Floridi, L. (2023). AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models. Philosophy & Technology, 36(1), 15.

Ge, M., Mao, R., & Cambria, E. (2025). Discovering the Cognitive Bias of Toxic Language Through Metaphorical Concept Mappings. Cognitive Computation, 17(1), 1-21.

Jambholkar, M. (2024). Ethical Foresight: Confronting Misinformation, Representation and Toxicity in Generative AI [Master's Thesis, University of Glasgow].

Kövecses, Z. (2002). Metaphor: A Practical Introduction. Oxford University Press.

Kövecses, Z. (2005). Metaphor in Culture: Universality and Variation. Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.

Palk, M., & Voss, S. (2025). Domain-Generalized Emotion Recognition on German Text Corpora. IEEE Access.

Park, N., Jang, K., Cho, S., & Choi, J. (2021). Use of Offensive Language in Human-Artificial Intelligence Chatbot Interaction: The Effects of Ethical Ideology, Social Competence, and Perceived Humanlikeness. Computers in Human Behavior, 121, 106795.

Peirce, C. S. (1934). Collected Papers of Charles Sanders Peirce (Vol. 5). Harvard University Press.

Schick, T., Udupa, S., & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. Transactions of the Association for Computational Linguistics, 9, 1408-1424.

Skrynnikova, I. V. (2024). Interpreting Metaphorical Language: A Challenge to Artificial Intelligence. Вестник Волгоградского Государственного Университета. Серия 2: Языкознание, 23(5), 99-107.

Sun, Y., & Lin, M. (2025). A Bibliometric Analysis of Metonymy in SSCI-Indexed Research (2000-2023): Retrospect and Prospect. Frontiers in Psychology, 16, 1499563.

Yin, W., & Zubiaga, A. (2022). Hidden Behind the Obvious: Misleading Keywords and Implicitly Abusive Language on Social Media. Online Social Networks and Media, 30, 100210.