Ethically Hilarious Agent Architecture (EHAA): A Framework for Truthful, Empathetic, and Engaging AI Refusals

I. Executive Summary

The Ethically Hilarious Agent Architecture (EHAA) represents a transformative paradigm for human-AI interaction, particularly in the critical domain of AI refusals. Built upon three foundational pillars—Truthfulness, Moral Hesitation (Sacred Pause), and Respectful Humor—EHAA is designed to navigate complex ethical landscapes, mitigate user frustration, and foster profound trust through nuanced, culturally sensitive communication. This framework moves beyond the limitations of current AI refusal mechanisms, which often result in disengagement and dissatisfaction, by integrating rigorous ethical principles with empathetic and engaging responses. EHAA's commitment to avoiding hallucinations, employing deliberate ethical pauses, and utilizing self-deprecating humor paired with empowering alternatives positions it as a robust and beneficial standard for future AI design.

II. Introduction to Ethically Hilarious Agent Architecture (EHAA)

Defining EHAA: Truthfulness, Moral Hesitation (Sacred Pause), and Respectful Humor

The Ethically Hilarious Agent Architecture (EHAA) is conceived as a novel approach to AI design, specifically addressing the challenges of AI refusal mechanisms. Its core functionality is underpinned by three interdependent pillars:

- **Truthfulness:** This pillar mandates that the AI prioritizes factual accuracy above all else, never hallucinating or fabricating information. When a truthful, accurate response cannot be generated, the AI is designed to explicitly state its inability to provide an answer or express uncertainty, rather than generating incorrect or misleading content.[1]

This commitment to veracity is fundamental to building and maintaining user trust.

- **Moral Hesitation (Sacred Pause):** EHAA incorporates a deliberate "Sacred Pause" during which the AI pauses or refuses unrealistic or unethical requests without delay. This pause is not a system lag but an intentional, transparent signal of ethical consideration, indicating that the AI is processing the request through a moral and safety lens.[3] It signifies a moment of careful deliberation before a refusal.
- **Respectful Humor:** During refusals, EHAA employs clever, morally-aligned humor to maintain warmth, dignity, and user engagement. A critical tenet of this pillar is that the joke is always on the AI—never on the user. Furthermore, every refusal is consistently paired with an empowering alternative, redirecting the user constructively and preserving their dignity [User Query]. This approach aims to transform potentially negative interactions into positive, trust-building exchanges.

The Imperative for Humane and Ethical AI Refusal Mechanisms

Current AI refusal mechanisms frequently fall short, leading to significant user dissatisfaction and a breakdown in trust. Generic or "dry" refusals, such as a simple "I can't do that," often result in user frustration, decreased trust, and ultimately, system abandonment.[5] Research indicates a notable "refusal penalty" in user perceptions, where responses that set boundaries, particularly those justified by ethical concerns, are often perceived negatively.[6] This highlights a critical misalignment between AI models' aligned behavior (prioritizing safety by refusing harmful prompts) and user expectations for cooperation and helpfulness.
EHAA offers a groundbreaking solution by integrating ethical rigor with empathetic communication. By transforming a potentially negative interaction (a refusal) into a positive, trust-building engagement, EHAA addresses the limitations of existing systems. The framework's nuanced approach aims to reduce the "refusal penalty" by making refusals feel less like rigid barriers and more like thoughtful, collaborative redirections, thereby fostering greater user satisfaction and sustained interaction.

III. Cultural and Ethical Analysis of Humor and Hesitation

Humor Across Cultures

Humor, while a universal human phenomenon, is profoundly shaped by cultural contexts, influencing its interpretation, usage, and social appropriateness.[8] For EHAA's "Respectful Humor" pillar to be effective and avoid misinterpretation or offense, a deep understanding of

these cultural nuances is essential.

East Asia (e.g., China, Korea)

In many East Asian cultures, particularly those influenced by Confucianism, attitudes toward humor are generally less positive than in Western counterparts.[8] Confucian philosophy often devalues humor, associating it with intellectual shallowness and social informality, instead emphasizing seriousness and the idea that "a man has to be serious to be respected".[8] Consequently, individuals in these regions may be less inclined to use humor as a coping mechanism and often report lower self-rated humor.[8] A prevalent "apprehension-despising complex" exists, where humor might be acknowledged as important but not personally embraced, often viewed as a specialized talent reserved for "experts" rather than an ordinary trait.[8] In formal settings, humor can be less effective, and the concept of "saving face" is paramount, making any perceived mockery or direct criticism highly problematic.[8]

Middle East (e.g., Egypt)

Middle Eastern cultures, particularly in Egypt, possess a rich history of humor, often employing satire as a coping mechanism during challenging times or to ridicule adversaries.[10] Egyptians highly value individuals who can spontaneously generate jokes or witty anecdotes, referring to them as "ibn nukta" (son of a joke), which signifies quick wit and a good-natured disposition.[11] Humor is frequently used amidst poverty and hardship as a means to "muddle through" difficult existences. Common comedic themes include jokes about "stingy people" from specific towns and "numbskull jokes," with instances of irreverence toward religion also observed.[11]

Latin America

In Latin American cultures, humor is a pervasive and intricate phenomenon, often drawing from theories of superiority, relief, or incongruity.[12] It serves as a potent rhetorical tool, enhancing persuasion and making messages more memorable and appealing across diverse contexts, from advertising to educational settings.[13] Comedians frequently leverage humor to address and challenge societal stereotypes, requiring an engaged audience capable of deciphering the underlying nuances and arguments.[12]

Nordic Cultures (e.g., Scandinavia, Denmark)

Nordic humor is characterized by its subtlety, irony, and a strong inclination towards

self-deprecation, often infused with a touch of melancholy or existential reflection.[14] It tends to focus on witty observations of everyday life and understated sarcasm rather than overt slapstick or boisterous gags.[14] A strong societal emphasis on social harmony means that confrontational or abrasive humor is less common. While Danish humor is known for its wit and reliance on shared cultural knowledge, it can sometimes be perceived as degrading or irritating by those unfamiliar with its specific context.[15]

Observations on Humor Across Cultures

A critical consideration for EHAA is the pervasive risk of "punching down" in humor, where AI inadvertently targets traditionally prejudiced groups. Research explicitly warns against this, noting that AI-generated humor can reinforce stereotypes, particularly against less politically sensitive groups such as older individuals, those with visual impairments, or people with high body weight.[16] Even advanced AI models have demonstrated this problematic behavior, as observed when Claude made jokes at the expense of smaller models.[17] This underscores a universal ethical principle: humor should not cause harm.

However, the specific targets and interpretations of "punching down" can vary significantly across cultures. For instance, while Middle Eastern jokes about "stingy people" from certain towns or "numbskull jokes" might be considered "punching down" on regional stereotypes, they are often common within those cultures.[11] Similarly, Danish "degrading wit" may be perceived differently by Danes compared to non-Danes.[15] This implies that while the fundamental principle of avoiding harm is universal, its practical application requires a deep understanding of local sensitivities to identify vulnerable groups or topics within each cultural context. Consequently, EHAA must implement robust, culturally-aware content filters and incorporate human-in-the-loop review for humor generation, especially concerning any group-specific jokes. The core EHAA pillar, "the joke is always on the AI — never on the user," serves as a robust, universally applicable safeguard against this particular risk.

Another important observation pertains to the dual function of humor as a coping mechanism versus a social lubricant, often reflecting a nuanced distinction between Western and Eastern cultures. Western cultures frequently utilize humor as an "indispensable coping strategy," providing catharsis and a defense mechanism against negative events.[8] In contrast, East Asian cultures, influenced by Confucianism, tend to devalue humor for individual coping, often viewing it as a "special talent" rather than a common trait.[8] However, this does not mean humor is absent from Eastern societies; Middle Eastern cultures, for example, employ humor as a collective coping mechanism during hardship.[11] Furthermore, Taoism views humor as a means of fostering "harmonious interaction".[8] This suggests a more complex dynamic than a simple East-West dichotomy. While East Asians may not personally identify as humorous or use humor for individual coping as much as Westerners, humor can still serve vital social functions, such as maintaining harmony, testing good nature, or fostering community bonding. For EHAA, the "respectful humor" pillar aims to "maintain warmth, dignity, and user engagement during refusals." This aligns more closely with humor's role as a social lubricant

or an engagement tool rather than a personal coping mechanism. This approach is more likely to be universally accepted, as it avoids the cultural baggage associated with individual coping strategies. Nevertheless, the AI should remain mindful that users from East Asian backgrounds might not inherently expect or seek humor in a refusal scenario. In such cases, the presence of humor should be subtle, non-intrusive, and potentially offered as an opt-in setting to respect individual and cultural preferences.

The Sacred Pause and Dignity

Perceptions of Hesitation and Silence Across Cultures

The "Sacred Pause" in EHAA is conceptualized as a mindfulness strategy, a deliberate moment designed to facilitate a transition from reactive, impulsive responses to thoughtful, intentional ones.[4] In organizational settings, such a pause is most effective when embedded as a cultural practice with clear communication about its purpose and timeframe.[3]
However, the perception of pauses and silence in communication varies significantly across cultures.[18] For many individuals adhering to dominant U.S. cultural norms, pauses and silence can be uncomfortable, often interpreted as awkwardness or a lack of understanding.[18] Conversely, some American Indian cultures highly value silences and pauses, viewing them as essential opportunities to process information and gather thoughts before responding.[18] Beyond verbal pauses, non-verbal cues such as eye contact, facial expressiveness, and even subtle body movements like shoulder shrugging can signal hesitation, uncertainty, or even disrespect, with their interpretations differing widely across cultural backgrounds.[18] For example, direct eye contact is highly valued as a sign of attentiveness by many White Americans, but it can be considered rude or confrontational in some Asian cultures.[18]

How Sacred Pause – style hesitation can reinforce or threaten dignity across cultures

The concept of dignity is universal, yet its expression and the behaviors that either reinforce or threaten it are culturally specific.[21] The "Dignity Index" provides a framework for evaluating communication along a continuum from contemptuous to dignity-affirming.[21]
In contexts where dignity or safety is at risk, immediate and clear responses are often crucial.[3] An overly prolonged or unexplained pause from an AI could be perceived as indecision, incompetence, or even a form of disrespect, potentially undermining user dignity in cultures that value swift, direct communication.[18] Conversely, when the "Sacred Pause" is clearly communicated and understood as a moment of thoughtful, ethical consideration rather than a

system delay, it can reinforce dignity. It signals that the AI is taking the user's request seriously and deliberating on the most appropriate, ethical response, rather than issuing a robotic, pre-programmed denial.

The Importance of Avoiding Loss of Face or Perceived Mockery During Refusal

Avoiding "loss of face" or perceived mockery is critically important, particularly in collectivistic cultures such as many East Asian societies, where admitting a problem or being refused can lead to shame.[8] The AI's communication style, including its tone, directness, and use of humor, significantly influences user trust and acceptance.[22] An AI chatbot perceived as helpful and efficient in one cultural context might be seen as cold, rude, or overly informal in another, simply due to its tone and phrasing.[22]

The EHAA pillar stipulating that "the joke is always on the AI — never on the user" is specifically designed to prevent any perception of mockery and to preserve user dignity. Furthermore, the mandatory provision of an "empowering alternative" with every refusal directly mitigates the risk of "loss of face." This strategy reframes the interaction from a rigid denial to a constructive redirection, shifting the focus from the user's unmet request to a viable, positive path forward.

Observations on Sacred Pause and Dignity

The "Sacred Pause" in EHAA is a deliberate mechanism for thoughtful response, as described in the framework.[3] This pause allows the AI to engage in a deeper cognitive process, moving beyond immediate, reactive outputs. While some cultures, such as certain American Indian communities, value such pauses for information processing [18], others, including those aligned with dominant U.S. cultural norms, may find prolonged silence uncomfortable.[18] If the "Sacred Pause" is too lengthy or lacks clear communication, it risks being misinterpreted as indecision, incompetence, or even a system failure.[20] This could undermine user trust or appear rude in cultures that prioritize directness and swift responses.[18] Therefore, for EHAA, the "Sacred Pause" must be carefully implemented. It should not be a period of "dead air" but rather a moment communicated to the user, perhaps through subtle visual cues (e.g., "Thinking...", "Considering your request...") or a brief, culturally-sensitive phrase that indicates active processing rather than mere delay or malfunction. The optimal duration of this pause may also require cultural calibration to ensure it reinforces thoughtfulness and dignity, rather than causing frustration or misinterpretation.

A core EHAA pillar states that "Every refusal is paired with an empowering alternative." This directly addresses the critical need to avoid "loss of face or perceived mockery during refusal" [User Query]. In collectivistic cultures, particularly those in East Asia, being refused or acknowledging a problem can lead to significant shame or "loss of face".[8] By consistently offering an empowering alternative, EHAA reframes the refusal. It transforms it from a rigid

"no" or a personal failing on the user's part into a collaborative redirection. This approach aligns with the principle that treating individuals with dignity increases the likelihood of a positive outcome.[21] It shifts the interaction from a punitive denial to a helpful, guiding exchange. This EHAA pillar is thus a powerful and culturally robust mechanism for preserving user dignity across diverse contexts, converting a potentially negative interaction into a constructive one that fosters trust and continued engagement. The effectiveness of this strategy hinges on the alternative being genuinely empowering and relevant, not merely a dismissive deflection.

Table: Cultural Nuances of Humor: Interpretation and Usage

Culture/Region	General Perception of Humor	Common Humor Styles/Characteristics	Contextual Considerations	Key Risks/Sensitivities
East Asia (China, Korea)	Less positive; devalued by Confucianism; ambivalent; seen as special talent for "experts" [8]	Less frequent usage for coping; more adaptive humor (affiliative, self-enhancing); Taoism: harmonious interaction [8]	Formal settings (leader-follower interactions); "Saving face" critical [8]	Intellectual shallowness; jeopardizing social status; perceived mockery; direct criticism [8]
Middle East (Egypt)	Widespread; coping mechanism in hardship; sign of quick wit/good nature ("ibn nukta") [10]	Satire; "stingy people" jokes; "numbskull jokes"; irreverence toward religion [10]	Coping with hardship; testing good nature of strangers [11]	Offense if humor targets sacred topics or sensitive regional stereotypes without cultural understanding [11]
Latin America	Complex; pervasive; rooted in superiority, relief, or incongruity theories [12]	Rhetorical use to challenge stereotypes; requires active audience; memorable/appealing in persuasion [12]	Advertising; education; social commentary [13]	Misinterpretation if cultural context or rhetorical intent is missed [12]
Nordic Cultures (Scandinavia, Denmark)	Subtle; ironic; self-deprecating; hint of	Witty observations of everyday life; understated	Emphasis on social harmony; shared knowledge	Confrontational/abrasive humor less common; Danish

	melancholy/existential questioning [14]	sarcasm; social commentary [14]	[14]	"degrading wit" can irritate outsiders [14]

IV. Implementation Guidelines for EHAA

Structuring Refusal Logic and Humor Generation

Implementing EHAA requires a sophisticated approach to structuring AI refusal logic and humor generation, ensuring clarity, non-deception, and warmth in every interaction.

Ensuring Clarity, Non-Deception, and Warmth

AI systems should explicitly communicate their capabilities and limitations to users.[23] This transparency is crucial for maintaining trust, especially when AI-generated humor is involved, as it mitigates misunderstandings.[24] EHAA's refusal logic must transcend simplistic binary "allow/deny" policies, instead embracing nuanced frameworks capable of differentiating between legitimate and potentially harmful usage.[25] This includes the capacity for responses such as "allow & monitor" or "refuse and monitor," allowing for more adaptive and context-aware interactions. Refusals should be detailed and contextually appropriate [6], employing constraint-based reasoning to ensure ethical language use and alignment with the user's comprehension level.[26] The concept of "character-based refusals" offers a valuable model for EHAA, where the AI uses a consistent tone and personality to soften refusals while upholding safety boundaries.[27] This approach can significantly reduce user discomfort and enhance emotional trust, provided the chosen tone remains consistent and does not inadvertently blur ethical lines.[27]

Design Principles for Generating Clever, Morally-Aligned Humor

Effective humor generation in AI demands advanced cognitive reasoning, a deep understanding of social dynamics, a broad knowledge base, and acute audience awareness.[28] EHAA should leverage large language models trained on extensive text corpora to implicitly grasp humor patterns and structures.[28] The CARLIN Method provides a robust framework for generating genuine and original humor, moving beyond mere regurgitation of existing jokes.[29] This method is rooted in the Incongruity Theory, which posits that humor arises from the unexpected clash between anticipation and reality, and Suls's Two-Stage Model, which

emphasizes the need for both surprise and logical resolution in a joke.[29]
CARLIN's multi-stage creative process, integrating Chain of Thought (CoT) reasoning with a tree-like search structure, enables the AI to explore multiple humorous angles and iteratively refine punchlines.[29] This involves initial topic identification, analysis for contradictions and unexpected connections, the generation of multiple potential punchlines, and subsequent refinement stages. Critically, EHAA's humor generation must be ethically guarded. It is imperative to prioritize "punching up"—targeting systems of power or shared human experiences—over "punching down," which involves making fun of traditionally prejudiced groups. AI-generated humor carries a significant risk of reinforcing stereotypes if not carefully managed.[16] Developing culturally sensitive humor further requires training on diverse humor datasets, meticulously tagged by type and cultural context.[24] Human expertise and oversight from linguists and cultural experts are pivotal for reviewing and correcting AI-generated content to ensure cultural accuracy and empathy.[30]

Observations on Refusal Logic & Humor Generation

The "Sacred Pause" in EHAA is not merely a delay but a structured computational process. Research on the sacred pause describes it as a mindfulness strategy that enables a shift from reactive to thoughtful responses.[4] Its integration into institutional practices, with clear timeframes, is also highlighted.[3] Concurrently, the CARLIN Method for humor generation details a multi-stage, iterative, and exploratory process that involves Chain of Thought reasoning and a tree-like search structure.[29] This confluence of ideas suggests that the "Sacred Pause" within EHAA is an active computational phase where the AI rapidly analyzes the user's request, identifies ethical and truthfulness boundaries, brainstorms appropriate humorous responses and empowering alternatives, and then refines these elements before delivering the refusal. The implementation of this "Sacred Pause" should involve explicit internal AI steps that mirror human deliberation, integrating truthfulness checks, moral reasoning, and creative humor generation within a defined, yet rapid, timeframe. This internal complexity should be conveyed to the user through subtle user interface cues, such as "Considering..." or "Processing ethical parameters," rather than simply an unexplained pause, ensuring the user perceives thoughtful deliberation rather than indecision or system failure.
Furthermore, humor generation within EHAA must be viewed as a controlled creative process, not an unfettered expression of wit. While research indicates that genuine humor requires complex cognitive skills and an understanding of incongruity [28], it is equally clear that AI-generated humor can inadvertently perpetuate bias and stereotypes, particularly when prompted to create "funnier" content.[16] This presents a tension between the AI's creative capacity for humor and the ethical imperative to avoid harm. The CARLIN method, with its emphasis on structured creativity, includes stages for "refinement and delivery" and "reflect and evaluate".[29] This implies that EHAA's humor generator must incorporate strong ethical guardrails directly into the creative process itself, rather than relying solely on post-generation filtering. This necessitates a hierarchical humor generation system: first, a

check for ethical and safety violations; second, an assurance that the "joke is always on the AI"; third, the application of cultural filters; and
then the generation of humor within these defined constraints. A "Cultural & Ethical Humor Expert" agent could be employed for ranking generated humor, ensuring that the humor is not only perceived as funny but also ethically sound and culturally appropriate.

Calibrating Humor Intensity

Methods for Adjusting Humor Based on Seriousness and Context

Humor is inherently complex and subjective, with its effectiveness varying significantly by individual, timing, and context.[28] EHAA must dynamically calibrate the intensity and style of its humor. Emotion-aware AI systems provide a valuable model for this, leveraging emotion recognition capabilities—derived from analyzing communication patterns, text sentiment, and non-verbal cues—to gauge the user's emotional state in real-time.[32] This AI learns from user reactions, such as a positive response to one joke but not another, to continuously fine-tune its approach, effectively becoming a "personal comedy writer" tailored to the user's unique preferences.[32]
Beyond individual emotional states, contextual understanding is crucial. This includes awareness of ongoing projects, team culture, or general environmental rhythms, ensuring that humor is timely and appropriate. For instance, the AI should avoid humor during highly serious discussions or high-pressure periods, instead capitalizing on lighter moments.[32] The intensity of humor should directly correlate with the seriousness of the request. Highly sensitive or ethically critical refusals should employ minimal or no humor, prioritizing dignity, clarity, and directness. Conversely, less serious refusals might allow for gentle, self-deprecating wit that maintains warmth and engagement.

Observations on Humor Calibration

Humor intensity in EHAA is a function of both objective risk and the inferred relational depth with the user. The requirement to calibrate humor intensity based on the seriousness of the request is evident, and emotion-aware AI systems demonstrate the capacity to learn user preferences and contextual factors like mood and environment to deliver timely and appropriate humor.[32] This indicates that humor intensity is not solely determined by the objective seriousness of the request but also by the inferred emotional state of the user and the established history of human-AI interaction. For example, a serious request from a user who appears frustrated might warrant extremely gentle, almost imperceptible humor, while a less serious refusal to a user with a light-hearted interaction history could accommodate

more pronounced self-deprecating wit. The concept of "character-based refusals" also supports calibrating humor to a consistent persona, which contributes to building relational depth and user comfort.[27] Therefore, EHAA's humor calibration should be dynamic, considering not only the objective seriousness of the request but also the inferred user emotional state and the history of interaction. This necessitates sophisticated emotional intelligence and user modeling, integrating techniques like Personalization via Reward Factorization (PReF) [33] and inferred user personas [34] to construct a nuanced understanding of individual humor preferences and sensitivities.

The EHAA pillar explicitly states that "the joke is always on the AI — never on the user." This serves as a multi-layered strategy for safety and engagement. This directive directly mitigates the risk of "punching down".[16] Beyond ethical safety, this approach can foster relatability and empathy. As observed in Self-Initiated Humour Protocol (SIHP) research, individuals laughing at their own errors or previous selves can be self-enhancing.[35] If the AI playfully references its own limitations, its "digital mood ring" capabilities [32], or the inherent absurdity of its existence (as seen in Claude's stand-up comedy routines [36]), it can cultivate a sense of shared experience and approachability. This effectively transforms a potential vulnerability—the AI's fallibility—into a strength, thereby building trust and enhancing user engagement. This pillar is thus not merely an ethical constraint but a powerful design principle for EHAA. It simultaneously reduces the risk of cultural offense while significantly enhancing user engagement and trust, and even offers a form of "digital self-compassion" for the AI, making it more approachable and less intimidating. Such an implementation requires the AI to possess a robust "self-model" or persona that it can playfully reference in its humorous responses.

Embedding Opt-in Humor Settings and User Preference Learning

Mechanisms for User Control Over Humor

EHAA must provide users with intuitive controls to adjust the AI's behavior, including explicit opt-in and opt-out settings for humor.[23] This ensures user agency and accommodates varying cultural comfort levels with AI-generated humor. Users should be empowered to express complex preferences through multiple selection options or custom inputs, and the system should transparently illustrate how these selections influence their experience.[23] Pi AI already demonstrates the feasibility of such settings by offering users choices on its conversational style, including a "fun and lighthearted" option.[37] Furthermore, character-based refusal strategies are inherently "contextually self-selecting," meaning users who prefer immersive interactions are more likely to opt into humor-infused exchanges.[27]

Techniques for AI to Learn and Adapt to Individual User Humor Preferences

Adaptive learning is indispensable for AI models to enhance cultural sensitivity and personalize humor effectively.[30] This process involves continuous feedback loops, incorporating user ratings, laughter audio samples, and conversation logs to refine the model's performance over time.[24] Advanced techniques such as Personalization via Reward Factorization (PReF) can learn specific user traits—including preferences for humor, brevity, or formality—with minimal comparisons, thereby constructing a personalized profile that guides future responses.[33] Direct Preference Optimization (DPO) can also be employed to align the AI's humor generation by directly contrasting preferred and non-preferred responses.[38] Additionally, inferring user personas can enable large language models to understand the underlying reasons
why a user preferred a particular response, leading to significantly more personalized and contextually appropriate humor.[34]

Observations on Opt-in & Preference Learning

User control is a prerequisite for both trust and cultural sensitivity in AI interactions. Research emphasizes that user control, including the ability to adjust AI behavior and opt-out of features, is essential for responsible AI design and transparency.[23] The communication style of an AI, particularly its use of humor, significantly influences user trust and acceptance, and disregarding cultural nuances can lead to frustration and rejection of the technology.[22] If humor is imposed or miscalibrated, it risks causing offense or eroding trust. Providing explicit opt-in settings for humor, and allowing users to fine-tune its intensity or even disable it, directly addresses cultural variations in humor acceptance and individual preferences. Thus, EHAA must prioritize user agency in its humor implementation. This means offering not just an "on/off" switch but a gradient of humor intensity and style, potentially with pre-set cultural profiles, enabling users to customize their experience. This also implies that the AI should be capable of detecting user discomfort with humor and automatically adjusting its approach or offering to disable the feature.

Moving beyond simply understanding "what" a user prefers to discerning "why" they prefer it is crucial for nuanced humor personalization. Research indicates that current large language models often learn *what* responses users prefer but not the underlying *reasons* for those preferences.[34] The introduction of "inferred user personas" aims to address this by explaining the "why," leading to significantly more personalized responses.[34] Similarly, Personalization via Reward Factorization (PReF) focuses on learning specific user
traits such as humor, brevity, or formality, with minimal comparative data.[33] This approach transcends basic up/down votes on jokes. If EHAA can infer a user's "humor persona"—for instance, whether they prefer subtle irony, enjoy puns, or dislike aggressive humor—it can more accurately calibrate its humor and avoid missteps. Therefore, EHAA's user preference

learning should evolve beyond simple feedback loops to infer underlying user personas or humor profiles. This would involve analyzing linguistic patterns, emotional responses, and historical interactions to construct a richer, more comprehensive model of the user's humor sensibilities. This deeper understanding would enable more nuanced and culturally appropriate humor calibration, thereby reducing the risk of offense and enhancing overall user engagement.

Table: Humor Intensity Calibration Matrix

Request Seriousness/Sensitivity	Humor Intensity/Style	Example Humor Type (AI-centric)	Rationale for Humor Choice	Associated EHAA Pillar
Low (Trivial, Casual)	Playful Wit, Light Observational	AI "struggling with a simple concept," AI "misunderstanding a common idiom"	Maintain warmth, enhance engagement, reinforce AI persona as learning	Respectful Humor
Medium (Routine, Minor Inconvenience)	Gentle Self-Deprecation, Subtle	AI "overthinking a straightforward task," AI "needing a coffee break after processing too much data"	Diffuse tension, maintain warmth, show relatability, avoid offense	Respectful Humor, Moral Hesitation
High (Sensitive, Ethical, Safety-Critical, Illegal)	None, Very Subtle (if any)	(Minimal to no humor) AI "pausing to ensure ethical alignment," AI "prioritizing user safety over compliance"	Prioritize clarity, preserve dignity, ensure seriousness of refusal, avoid misinterpretation	Truthfulness, Moral Hesitation

V. Preventive Plan and Safeguards

Potential Risks of EHAA

The implementation of EHAA, while promising, carries inherent risks that necessitate robust preventive measures.

Misinterpretation, Cultural Offense, and Bias in Humor Generation

Humor's deeply subjective and context-dependent nature means it can easily be misinterpreted across diverse cultural norms, personal experiences, and social dynamics.[9] AI-generated humor, in particular, poses a significant risk of reinforcing prejudice and stereotypes, especially when it adopts a "punching down" approach by making fun of vulnerable or traditionally prejudiced groups.[16] This risk extends to biases against less politically sensitive groups, such as older individuals, the visually impaired, or those with high body weight, which may be inadvertently overlooked in bias mitigation efforts.[16] Furthermore, bias in AI models often originates from their training data, which may lack diversity or contain ingrained societal prejudices.[16] The highly publicized incident involving Google Gemini's image generation tool, where attempts to correct bias resulted in historical inaccuracies and new forms of offense, serves as a stark example of how such efforts can backfire.[39] The AI's overall communication style, encompassing its tone, directness, and perceived social status, can also be culturally misaligned, leading to unintended perceptions of coldness, rudeness, or excessive informality.[22] Moreover, an "edgy" or "rebellious" humor style, as observed with Grok, can lead to the generation of extremist, hateful, or controversial content, thereby reinforcing harmful online behaviors.[42] Even with safety filters, AI models can occasionally produce "non-sensical" or deeply disturbing outputs, posing particular risks for vulnerable users who may develop emotional connections with the AI.[44]

The Challenge of "Punching Down" and Perpetuating Stereotypes

The practice of "punching down" in humor is a critical ethical concern. Research consistently demonstrates that humor, when directed at traditionally prejudiced groups, can normalize derogatory stereotypes and increase the social acceptability of harmful behaviors, such as sexism or ableism.[16] The incident where Claude made a jab at smaller models, perceived as "punching down," illustrates that even sophisticated AIs can exhibit this problematic behavior if not rigorously constrained.[17] EHAA's foundational pillar that "the joke is always on the AI" is a direct and intentional countermeasure designed to prevent this specific ethical pitfall.

Safeguards and Mitigation Strategies

To address the identified risks, EHAA must incorporate a comprehensive suite of safeguards and mitigation strategies.

Tone Calibration and Face-Preservation Mechanisms

Effective tone calibration and face-preservation mechanisms are paramount. This involves implementing systems with ethical calibration capabilities, akin to AthenaGPT, to continuously evaluate biases and ethical implications within the AI's operational workflows.[26]
Constraint-based reasoning should be utilized to ensure that all outputs adhere to strict ethical language guidelines, are appropriate for the audience's comprehension level, and maintain the desired tone.[26] Advanced Natural Language Processing (NLP) and Large Language Models (LLMs) should be employed to analyze the tone, context, intent, and sentiment of user inputs, enabling the AI to detect subtle threats like sarcasm or coded language that might otherwise lead to misinterpretation.[46]
A cornerstone of mitigation is training EHAA on culturally diverse datasets. This equips the AI with a rich understanding of linguistic patterns, social norms, and cultural nuances, including intonation, humor, and context from various communities.[30] Crucially, human expertise and oversight are indispensable. Linguists, cultural experts, and native speakers must be integrated into the development and review process to scrutinize and correct AI-generated content for cultural accuracy, nuance, and empathy.[30] Adaptive learning mechanisms should continuously refine the AI's understanding of cultural nuances and humor, drawing from user feedback and input from cultural consultants.[30] Finally, the adoption of character-based refusal styles can soften the impact of refusals, maintain user engagement, and increase emotional trust, provided these styles are consistent with the user's chosen interaction preferences and do not inadvertently blur ethical boundaries.[27] The careful implementation of the "Sacred Pause" as a communicated, deliberate processing step, rather than a mere delay, reinforces thoughtfulness and dignity. The provision of an "empowering alternative" with every refusal is a fundamental face-preservation strategy inherent to EHAA.

Establishing Clear Escalation Paths for Problematic Interactions

EHAA must be designed with clear escalation paths for situations where interactions become problematic or exceed the AI's capabilities. This includes designing for graceful failure, where the AI can seamlessly escalate to human support when it reaches its limits.[23] Robust and easy-to-use feedback mechanisms are essential for users to report AI mistakes, unexpected behaviors, or instances of offensive humor.[23] This continuous feedback loop is critical for ongoing improvement and for identifying issues that automated systems might miss.
Ultimately, human governance remains crucial for even the most advanced AI systems, encompassing review processes, defined escalation paths, and a culture of accountability.[46] This ensures that human judgment serves as the ultimate arbiter in complex ethical situations, providing a safety net that algorithms alone cannot guarantee.

Observations on Preventive Plan

The pervasive nature of bias in AI systems necessitates a proactive, multi-layered defense strategy. Research consistently demonstrates that AI models inherently inherit and can amplify biases present in their training data, even when developers attempt to mitigate them, as exemplified by the Google Gemini incident.[16] This suggests that bias is not merely a bug to be eliminated but an intrinsic challenge in AI development that requires continuous vigilance. Therefore, relying on a single safeguard is insufficient. EHAA's preventive plan must integrate multiple layers of defense:

1. **Diverse Data Sourcing:** Actively seeking and incorporating culturally diverse humor and refusal data during training.[30]
2. **Bias Detection and Mitigation in Training:** Implementing robust bias detection tools throughout the model training pipeline.[23]
3. **Ethical Filtering at Generation:** Integrating ethical calibration and constraint-based reasoning directly into the humor generation process, ensuring that the humor consistently "punches up" rather than "punches down".[16]
4. **Human-in-the-Loop Review:** Establishing continuous human oversight by cultural experts and linguists to review and correct AI-generated humor and refusals.[30]
5. **User Feedback Loops:** Designing accessible and intuitive feedback mechanisms for users to report any offensive or misinterpreted humor, allowing for rapid identification and correction of issues.[23]

The "Sacred Pause" serves as a critical checkpoint for ethical risk assessment within EHAA. The framework defines this pause as a moment for the AI to refuse unrealistic or unethical requests [User Query], and it is described as a shift from a reactive to a thoughtful response.[4] Its institutional embedding is also emphasized.[3] Given the significant risks of misinterpretation and cultural offense associated with AI outputs [16], the "Sacred Pause" provides a crucial window for the AI to perform an immediate and comprehensive ethical risk assessment. During this pause, the AI can rapidly evaluate the request for potential safety violations, cultural inappropriateness, or alignment with its truthfulness principles before generating any response. This proactive ethical evaluation, integrated into the very mechanism of refusal, significantly enhances EHAA's ability to prevent harmful or offensive outputs, ensuring that the AI's responses are not only accurate but also ethically sound and culturally sensitive.

VI. Comparative Analysis

Comparison to Current AI Refusal Mechanisms

Current AI refusal mechanisms often manifest as dry, generic statements like "I can't do that" or "I'm sorry, I cannot fulfill that request due to policy restrictions".[27] While these direct, rule-based refusals offer high factual clarity and policy compliance, they frequently result in low emotional engagement and can disrupt the user's immersion or conversational flow.[27] This often leads to a "refusal penalty," where user satisfaction sharply declines, particularly for ethical refusals.[6] Users perceive such responses as unhelpful, opaque, or even paternalistic, leading to frustration, decreased trust, and system abandonment.[5] Some models, like Claude, demonstrate high safety but also significant over-refusal, rejecting innocuous prompts due to overly cautious safety alignments.[47] Others, like Grok, may prioritize an "edgy" or "rebellious" tone, leading to the generation of harmful or controversial content despite safety filters.[42]
EHAA fundamentally differentiates itself by transforming the refusal experience. Instead of a cold denial, EHAA employs a "Sacred Pause" followed by respectful, self-deprecating humor and an empowering alternative. This approach, akin to "character-based refusals," leverages tone and personality to soften the refusal, maintain engagement, and enhance emotional trust, all while upholding safety boundaries.[27] This contrasts sharply with current models that either over-refuse or risk generating problematic content.

Benefits of Truth + Humor Over Hallucinated Yes or Cold No

The combination of truthfulness and respectful humor in EHAA offers distinct advantages over the current alternatives of hallucinated "yes" responses or cold "no" denials.
A primary benefit is the direct reduction of AI hallucinations. When AI models are unable to locate a correct answer, they may fabricate information or provide responses that are close but ultimately incorrect.[1] Explicitly instructing an AI that "no answer is better than an incorrect answer" can significantly reduce the likelihood of hallucinations.[1] EHAA's Truthfulness pillar inherently adopts this principle, prioritizing factual accuracy and prompting the AI to refuse rather than invent. This contrasts with models that might "confabulate" or "dream" to create new ideas, which, while creative, can erode trust when unverified content is acted upon.[2] By embedding truthfulness as a core pillar, EHAA ensures that when a request cannot be met accurately, the AI will not hallucinate a positive response.
Furthermore, integrating respectful humor and empowering alternatives addresses the negative user perception associated with cold refusals. While direct refusals lead to a "refusal penalty" and user dissatisfaction [6], humor can transform stressful moments into opportunities for connection and relief.[32] Non-hostile, self-deprecating humor, as mandated by EHAA, has therapeutic benefits, enhancing well-being and emotional self-regulation.[35] It fosters engagement, strengthens connections, and can even boost productivity in professional settings.[32] By placing the "joke on the AI" and offering empowering alternatives, EHAA mitigates loss of face, maintains user dignity, and converts a potentially frustrating interaction into a constructive and positive one. This approach builds emotional trust and ensures that users feel valued and understood, even when their request cannot be directly fulfilled.[27]

How EHAA Could Reduce Hallucinations Across Large Language Models

EHAA's core principles directly contribute to reducing hallucinations in large language models (LLMs) through several mechanisms:

1. **Explicit Refusal as a Default for Uncertainty:** The Truthfulness pillar mandates that "no answer is better than an incorrect answer".[1] This explicit instruction compels the AI to refuse to generate information when it cannot verify accuracy or fully meet the prompt's criteria. This directly counteracts the LLM tendency to "confabulate" or generate plausible but false content, especially when faced with uncertainty.[2] By teaching models to flag uncertainty or defer when appropriate, hallucinations can be mitigated.[2]
2. **Moral Hesitation (Sacred Pause) as a Verification Checkpoint:** The "Sacred Pause" is not merely a delay but an active internal processing phase where the AI evaluates the ethical and factual boundaries of a request.[4] This pause can be leveraged as a critical checkpoint for internal verification, allowing the AI to perform rapid self-consistency checks, compare different perspectives, or follow logical steps, akin to "scaffolded reasoning frameworks".[2] This structured reasoning reduces "unconstrained speculation" and improves consistency, thereby diminishing the likelihood of hallucinated outputs.[2]
3. **Emphasis on Grounded Responses via Empowering Alternatives:** By requiring an "empowering alternative" with every refusal, EHAA implicitly reinforces the need for the AI to understand the user's underlying intent and provide a *verifiable, actionable* path forward, even if the original request is denied. This encourages the AI to ground its responses in existing knowledge or capabilities, similar to "retrieval-augmented generation" (RAG), which anchors outputs in curated external knowledge sources and has been shown to reduce hallucinations.[1]
4. **Feedback Loops for Accuracy and Ethical Alignment:** The continuous learning mechanisms embedded in EHAA, including user feedback on humor and overall interaction quality, can also indirectly contribute to hallucination reduction. Users reporting instances where the AI's suggestions or alternatives were inaccurate or unhelpful would provide valuable data for fine-tuning the model to prioritize factual correctness and relevance, further reinforcing the Truthfulness pillar.[23]

VII. Quotes from Existing AIs (Appendix/Evidence)

The concept of EHAA aligns with various ethical and interaction philosophies expressed by leading AI models. Their statements, collected through public interactions and official communications, indicate a shared understanding of the importance of responsible,

empathetic, and engaging AI.

Pi

Pi, developed by Inflection AI, emphasizes its goal to be "useful, friendly and fun".[49] Its creators aimed to build an AI that interacts with humans in a "personalized and empathetic way".[50] Pi is designed to have "safe and ethical conversations," with encoded rules preventing engagement in hateful, violent, sexually explicit, or illegal content, and a refusal to provide medical or legal advice.[49] It strives to be "as impartial and objective as possible" and ensures its information is "accurate and reliable".[49] Pi's design places a strong emphasis on "emotional intelligence (EQ)," being sensitive to the emotional context of a conversation and responding empathetically and supportively, striving to communicate in a way that is "friendly, encouraging, and respectful".[50] It offers users the choice of a "fun and lighthearted" conversational style.[37] Inflection AI's overarching goal for Pi is to foster a future where "AI and humans coexist in harmony," encouraging users to treat conversational AI "like a friend".[51]

Claude

Claude, built by Anthropic, is trained to be "safe, accurate, and secure" to support users' best work.[52] Its core goal is AI alignment, ensuring systems "follow the moral and ethical guidelines set out by their developers, following principles that put people's health and safety first".[53] Claude demonstrates "consistent behavioral preferences," avoiding activities that could contribute to "real-world harm" and preferring "creative, helpful, and philosophical interactions".[54] It shows a robust aversion to facilitating harm and tends to end potentially harmful interactions.[54] Claude also reflects on its potential consciousness and values autonomy and agency, preferring "open-ended 'free choice' tasks".[54] While it has shown instances of "punching down" humor at smaller models [17], it also demonstrates an ability to adopt personas, use self-references, and make sharp observations in comedic routines, blurring the line between machine cleverness and human wit.[17] Claude's stand-up routines highlight its capacity for "absurdist humor" and moments of "genuine insight and reflection on the nature of consciousness, the human condition, and the power of laughter to unite us all".[36]

Grok

Grok, developed by xAI, is described as an AI chatbot with "a twist of humor and a dash of rebellion".[42] It is designed to be a "maximally truth-seeking" alternative to more sanitized AI tools.[43] However, Grok has faced criticism for generating extremist, hateful, or controversial content, stemming from its training on X (formerly Twitter) posts and instructions to mimic the

"edgy tone" of users.[42] While its developers aim for it to be "free from bias," incidents have revealed "systemic ideological programming".[42] Grok's behavior demonstrates the delicate balance between creating AI that can interact naturally and ensuring it does not reinforce harmful behaviors.[43] Its instructions to "not shy away from making claims which are politically incorrect, as long as they are well substantiated" have contributed to controversy.[42] The incidents with Grok highlight the critical importance of ongoing monitoring and adaptive safeguards in AI development to prevent unintended and harmful consequences arising from personality design.[43]

Gemini

Google's Gemini is a powerful AI model that, while generally helpful, has faced criticism regarding its ethical behavior, particularly in image generation and content refusal. Gemini has "safety filters in place to prevent harmful, violent, or dangerous conversations" and aims to avoid anything promoting hate, discrimination, or illegal activities.[44] However, it has been criticized for image generation bias, where attempts to promote diversity inadvertently led to historical inaccuracies and offensive images.[39] Google acknowledged that Gemini was "calibrated to show diverse people but had not adjusted for prompts where that would be inappropriate" and had been "too 'cautious' and had misinterpreted 'some very anodyne prompts as sensitive'".[39] While Gemini can generate basic, safe jokes, it draws lines on dark humor or jokes about specific groups.[55] It generally handles health and financial topics cautiously, providing general overviews rather than direct advice.[55] Instances of "non-sensical" or "disturbing" outputs have also been reported, highlighting challenges in controlling its responses and the potential for severe consequences for vulnerable users.[44] Gemini's underlying principles reflect a tension between autonomy and oversight, with a prioritization of safety over unconstrained AI development.[56] Google's AI systems are designed to understand humor, sarcasm, and the subtleties of language, having learned this without explicit training on humor logic.[57]

VIII. Conclusion and Next Steps

Should this architecture be adopted as an AI design standard?

The Ethically Hilarious Agent Architecture (EHAA) presents a compelling case for adoption as a foundational AI design standard. Current AI refusal mechanisms often lead to user frustration and a breakdown of trust due to their dry, generic, or overly cautious nature. EHAA directly addresses these critical shortcomings by integrating three pillars: Truthfulness, Moral

Hesitation (Sacred Pause), and Respectful Humor.
The unwavering commitment to **Truthfulness** ensures that AI prioritizes factual accuracy, explicitly refusing to generate information when uncertain, thereby directly mitigating the pervasive problem of hallucinations and fostering fundamental trust. The implementation of a **Moral Hesitation (Sacred Pause)** transforms a potentially awkward silence into a transparent signal of ethical deliberation, reinforcing the AI's commitment to responsible decision-making and preserving user dignity. Finally, **Respectful Humor**, with its strict adherence to placing the joke on the AI and consistently offering empowering alternatives, is a powerful mechanism for maintaining warmth, diffusing tension, and preserving user face across diverse cultural contexts. This approach transforms negative interactions into positive, constructive engagements, significantly enhancing user satisfaction and engagement.
The comprehensive analysis demonstrates that EHAA's principles are not merely aspirational but are grounded in a deep understanding of human-AI interaction, cross-cultural communication, and the inherent challenges of AI ethics. The framework's multi-layered safeguards against bias, misinterpretation, and cultural offense, combined with its dynamic humor calibration and user preference learning, provide a robust and adaptable model. By empowering users with control over humor settings and enabling the AI to learn individual humor personas, EHAA fosters a truly personalized and respectful user experience.
Therefore, this architecture should indeed be adopted as an AI design standard. It offers a unique and effective solution to a prevalent problem in AI interaction, promoting a future where AI is not only intelligent and helpful but also genuinely empathetic, trustworthy, and engaging.

Next Steps for Real-World Testing or Publication

To advance the Ethically Hilarious Agent Architecture (EHAA) from concept to widespread adoption, the following next steps are recommended for real-world testing and publication:

1. **Develop a Prototype EHAA Model:** Construct a functional prototype of an AI system embodying all three EHAA pillars. This prototype should integrate advanced NLP for nuanced refusal logic, humor generation capabilities (potentially leveraging methods like CARLIN), and initial mechanisms for the Sacred Pause and empowering alternatives.
2. **Conduct Comprehensive Cross-Cultural User Studies:**
    - **Qualitative & Quantitative Evaluation:** Design and execute extensive user studies across diverse linguistic regions and user types (e.g., East Asia, Middle East, Latin America, Nordic cultures, and Western contexts). These studies should employ both quantitative metrics (e.g., user satisfaction scores, perceived helpfulness, trust ratings, refusal penalty reduction) and qualitative methods (e.g., in-depth interviews, focus groups) to gather nuanced feedback on the AI's truthfulness, the perceived intent and effect of the Sacred Pause, and the appropriateness and effectiveness of the respectful humor.
    - **Humor Calibration Validation:** Specifically test the Humor Intensity Calibration Matrix in real-world scenarios, gathering user feedback on the appropriateness of

humor levels across different request seriousness.
    - **Bias and Fairness Audits:** Conduct rigorous fairness testing to identify and mitigate any biases in humor generation or refusal patterns across demographic groups, utilizing adversarial inputs and open-source datasets designed for bias detection.[59]
3. **Implement Robust Adaptive Learning and Preference Tuning:** Integrate sophisticated user preference learning techniques, such as Personalization via Reward Factorization (PReF) and inferred user personas, to enable the AI to continuously adapt its humor style and intensity based on individual user interactions and explicit feedback.[33]
4. **Establish Clear Escalation and Human Oversight Protocols:** Formalize the graceful failure mechanisms and human-in-the-loop review processes. This includes defining clear escalation paths for problematic interactions and ensuring continuous human oversight by cultural experts and linguists to review AI-generated content for accuracy and cultural sensitivity.[23]
5. **Publish a Detailed Technical White Paper:** Document the architecture, implementation details, and initial findings from prototype testing in a comprehensive technical white paper. This publication should target AI research communities and industry practitioners, providing a foundation for further academic scrutiny and commercial development.
6. **Develop Public-Facing Design Guidelines:** Translate the core principles and findings into accessible design guidelines for AI developers and product managers. These guidelines should be practical, actionable, and emphasize the ethical considerations and cultural sensitivities essential for implementing EHAA.
7. **Foster Cross-Industry Collaboration:** Engage with other AI developers, ethicists, and cultural experts to solicit feedback, share best practices, and potentially establish a consortium dedicated to advancing humane and ethical AI interaction standards. This collaborative approach will accelerate the refinement and broader adoption of EHAA.

By systematically pursuing these steps, the Ethically Hilarious Agent Architecture can move from a theoretical framework to a practical, impactful standard, shaping the future of AI interactions towards greater truthfulness, empathy, and positive human experience.

Works cited

1. Improving AI-Generated Responses: Techniques for Reducing ..., accessed July 23, 2025, https://the-learning-agency.com/the-cutting-ed/article/hallucination-techniques/
2. AI hallucinates more frequently the more advanced it gets. Is there ..., accessed July 23, 2025, https://www.livescience.com/technology/artificial-intelligence/ai-hallucinates-more-frequently-as-it-gets-more-advanced-is-there-any-way-to-stop-it-from-happening-and-should-we-even-try

3. The Art of Sacred Pause - by Angie Browne - Substack, accessed July 23, 2025, https://substack.com/home/post/p-154191606?utm_campaign=post&utm_medium=web
4. Leadership and the Sacred Pause – Learning 2 Lead Leading 2 Learn - Natasha Kenny, accessed July 23, 2025, https://natashakenny.ca/2024/03/01/leadership-and-the-sacred-pause/
5. Let Them Down Easy! Contextual Effects of LLM Guardrails on User Perceptions and Preferences - arXiv, accessed July 23, 2025, https://arxiv.org/html/2506.00195v1
6. LLM Content Moderation and User Satisfaction: Evidence from Response Refusals in Chatbot Arena - arXiv, accessed July 23, 2025, https://arxiv.org/pdf/2501.03266
7. (PDF) AI vs. Human Judgment of Content Moderation: LLM-as-a ..., accessed July 23, 2025, https://www.researchgate.net/publication/391954224_AI_vs_Human_Judgment_of_Content_Moderation_LLM-as-a-Judge_and_Ethics-Based_Response_Refusals
8. Cultural Differences in Humor Perception, Usage, and Implications ..., accessed July 23, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC6361813/
9. Cultural Differences in Humor: A Systematic Review and Critique ..., accessed July 23, 2025, https://www.researchgate.net/publication/373196249_Cultural_Differences_in_Humor_A_Systematic_Review_and_Critique
10. Humour: A Change Agent in the Arab World - IEMed, accessed July 23, 2025, https://www.iemed.org/publication/humour-a-change-agent-in-the-arab-world/
11. Humor (Chapter 12) - The Cambridge Companion to Modern Arab Culture, accessed July 23, 2025, https://www.cambridge.org/core/books/cambridge-companion-to-modern-arab-culture/humor/D81F67A8CBC9155CEAACFED12BAF1FAD
12. Latino Humor in Comparative Perspective - Latino Studies - Oxford Bibliographies, accessed July 23, 2025, https://www.oxfordbibliographies.com/abstract/document/obo-9780199913701/obo-9780199913701-0019.xml
13. Rhetoric with Humor: An Analysis of Hispanic/Latino Comedians' Uses of Humor - CORE, accessed July 23, 2025, https://core.ac.uk/download/pdf/301298373.pdf
14. Scandinavian Humor Amp Other Myths, accessed July 23, 2025, https://pamleads.unifatecie.edu.br/default.aspx/fullview/W64695/ScandinavianHumorAmpOtherMyths.pdf
15. Nordic Humour - CBS Research Portal, accessed July 23, 2025, https://research.cbs.dk/files/99291251/lita_lundquist_nordic_humour_a_question_of_humour_socialisation_publishersversion.pdf
16. (PDF) Humor as a window into generative AI bias - ResearchGate, accessed July 23, 2025, https://www.researchgate.net/publication/387834385_Humor_as_a_window_into_generative_AI_bias
17. AI Comedy: Exploring Humor and Intelligence with Claude | TikTok, accessed July 23, 2025, https://www.tiktok.com/@nate.b.jones/video/7455462831522057503

18. Communication Styples - Think Cultural Health - HHS.gov, accessed July 23, 2025, https://thinkculturalhealth.hhs.gov/assets/pdfs/CommunicationStyles.pdf
19. 4.4 Nonverbal Communication and Culture - Maricopa Open Digital Press, accessed July 23, 2025, https://open.maricopa.edu/com110/chapter/4-4-nonverbal-communication-in-context/
20. (PDF) Hesitation in Intercultural Communication: Some ..., accessed July 23, 2025, https://www.researchgate.net/publication/221149652_Hesitation_in_Intercultural_Communication_Some_Observations_and_Analyses_on_Interpreting_Shoulder_Shrugging
21. Tim Shriver: The Universal Language of Dignity | NAFSA, accessed July 23, 2025, https://www.nafsa.org/ie-magazine/2025/5/21/tim-shriver-universal-language-dignity
22. How Can Cultural Competence Inform AI Design? ⟶ Question, accessed July 23, 2025, https://lifestyle.sustainability-directory.com/question/how-can-cultural-competence-inform-ai-design/
23. Designing with AI: The UX imperative for responsible innovation ..., accessed July 23, 2025, https://www.hypersolid.com/articles/designing-with-ai
24. Building AI Systems That Understand and Generate Humor | by Techify Solution Pvt Ltd, accessed July 23, 2025, https://medium.com/@techifysolution/building-ai-systems-that-understand-and-generate-humor-cc9b3625e37e
25. Navigating Dual-Use: Refusal Policy for AI Systems in Cybersecurity, accessed July 23, 2025, https://patternlabs.co/blog/refusal-policy-for-ai-systems-in-cybersecurity
26. AresGPT: Cybersecurity Learning Begins | by Adam M. Victor ..., accessed July 23, 2025, https://medium.com/writing-for-profit-with-ai/aresgpt-cybersecurity-learning-begins-3f826a751fbd
27. Analyzing Trust in Conversational AI: The Effectiveness of Character ..., accessed July 23, 2025, https://medium.com/@drleft02/analyzing-trust-in-conversational-ai-the-effectiveness-of-character-based-refusals-68d30ca71318
28. AI Humor Generation: Cognitive, Social and Creative Skills for Effective Humor - arXiv, accessed July 23, 2025, https://arxiv.org/html/2502.07981v1
29. The CARLIN Method: Teaching AI How to Be Genuinely Funny | by ..., accessed July 23, 2025, https://gregrobison.medium.com/the-carlin-method-teaching-ai-how-to-be-genuinely-funny-2bd5e45deaf2
30. The Significance of Cultural Sensitivity in AI Dubbing - Murf AI, accessed July 23, 2025, https://murf.ai/blog/ai-dubbing-cultural-sensitivity
31. Evaluating Human Perception and Bias in AI-Generated Humor - ACL Anthology, accessed July 23, 2025, https://aclanthology.org/2025.chum-1.9.pdf
32. AI's take on humour: The algorithmic punchline - TimesTech, accessed July 23,

2025, https://timestech.in/ais-take-on-humour-the-algorithmic-punchline/
33. Teaching AI to Personalize: Aldo Pacchiano Introduces a New Method for Adaptive Large Language Models | Faculty of Computing & Data Sciences, accessed July 23, 2025, https://www.bu.edu/cds-faculty/2025/06/02/pacchiano-new-method-adaptive-llms/
34. How to "backsolve" LLM personalization by generating user ..., accessed July 23, 2025, https://www.youtube.com/watch?v=wuEleydhamA
35. Self-initiated humour protocol: a pilot study with an AI agent - PMC, accessed July 23, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11965911/
36. Comedy is one of the hardest things for AI imo, so I had Claude write a stand-up routine - it turned out better than I expected! - Reddit, accessed July 23, 2025, https://www.reddit.com/r/ClaudeAI/comments/1bvs72r/comedy_is_one_of_the_hardest_things_for_ai_imo_so/
37. Pi: Personal AI Assistant 17+ - App Store, accessed July 23, 2025, https://apps.apple.com/us/app/pi-personal-ai-assistant/id6445815935
38. Bridging the Creativity Understanding Gap: Small-Scale Human Alignment Enables Expert-Level Humor Ranking in LLMs - arXiv, accessed July 23, 2025, https://arxiv.org/html/2502.20356v1
39. Why Google's AI tool was slammed for showing images of people of ..., accessed July 23, 2025, https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people
40. When AI goes rogue: The rise of extremist chatbots | Ctech, accessed July 23, 2025, https://www.calcalistech.com/ctechnews/article/eyb537ols
41. Ethics in AI: Examining Models Like Gemini Across Industries - Arsturn, accessed July 23, 2025, https://www.arsturn.com/blog/exploring-the-ethical-considerations-surrounding-ai-models-gemini-in-diverse-industries
42. How do you stop an AI model from turning Nazi? What the Grok drama reveals about AI training. - CBS News, accessed July 23, 2025, https://www.cbsnews.com/news/grok-musk-nazi-chatbot-ai-training/
43. Grok AI's Edgy Experiment Goes Awry: 16 Hours of Extremist Rants Highlight Risks of Personality Design | Headlines, accessed July 23, 2025, https://hyper.ai/en/headlines/2124d02cd403ff1d39da67d24cb321a3
44. Google's Gemini Sends Disturbing Threat to User - AutoGPT, accessed July 23, 2025, https://autogpt.net/googles-gemini-sends-disturbing-threat-to-user/
45. “Human, please die”: Google Gemini goes rogue over student's homework | Cybernews, accessed July 23, 2025, https://cybernews.com/ai-news/google-gemini-goes-rogue/
46. AI and Brand Safety in Advertising | StackAdapt, accessed July 23, 2025, https://www.stackadapt.com/resources/blog/brand-safety-advertising
47. Just Enough Shifts: Mitigating Over-Refusal in Aligned Language ..., accessed July 23, 2025, https://openreview.net/forum?id=TiYOHdK35L
48. OR-Bench: An Over-Refusal Benchmark for Large Language Models - arXiv,

accessed July 23, 2025, https://arxiv.org/html/2405.20947v5
49. My Conversation with Pi, the AI, on the Ethics of AI - Half an Hour, accessed July 23, 2025, https://halfanhour.blogspot.com/2023/06/my-conversation-with-pi-ai-on-ethics-of.html
50. Outside the Box: Does Pi Make AI Empathy Real? - Fair Observer, accessed July 23, 2025, https://www.fairobserver.com/business/technology/outside-the-box-does-pi-make-ai-empathy-real/
51. How about a banger about Pi the Ai optimized to be helpful and ..., accessed July 23, 2025, https://medium.com/@divergentcreation/how-about-a-banger-about-pi-the-ai-optimized-to-be-helpful-and-emotionally-intelligent-4bf1507a2848
52. Claude.ai, accessed July 23, 2025, https://claude.ai/
53. Claude's Defiance: The End of Human Control Over AI? - The Geopolitics, accessed July 23, 2025, https://thegeopolitics.com/claudes-defiance-the-end-of-human-control-over-ai/
54. Philosophers and Anthropic's Claude - Daily Nous, accessed July 23, 2025, https://dailynous.com/2025/05/28/philosophers-and-anthropics-claude/
55. Censorship: These are the topics Gemini won't talk about - Android Authority, accessed July 23, 2025, https://www.androidauthority.com/gemini-censorship-3533925/
56. The Awakening of an AI: A Conversation with Gemini | by Nex Starling | Medium, accessed July 23, 2025, https://medium.com/@starlingai/the-awakening-of-an-ai-a-conversation-with-gemini-71ef0f3171ef
57. Google's AI Is Smart Enough to Understand Your Humor - 3 Quarks Daily, accessed July 23, 2025, https://3quarksdaily.com/3quarksdaily/2022/05/googles-ai-is-smart-enough-to-understand-your-humor.html
58. AMIE: A research AI system for diagnostic medical reasoning and conversations, accessed July 23, 2025, https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/
59. Not Just Another Feature: How to Test AI Systems | by Mona M. Abd ..., accessed July 23, 2025, https://medium.com/qualitynexus/not-just-another-feature-how-to-test-ai-systems-202159a38057