Ethical Dilemmas of AI Humor: Can Bots Be Funny Without Cancel Culture?

Introduction – Isaac

Our presentation investigates the rapid advancement of generative AI into the uniquely human domain of humor, and the complex social dynamics of cancel culture. As AI models demonstrate a growing capacity to generate humorous content (from witty text to live improvised comedy) they simultaneously expose deep-seated ethical vulnerabilities.

The core question we explore is whether AI can successfully create humor without perpetuating harmful biases and consequently facing public backlash or "cancellation".

Our analysis is based on four key papers.

We draw on **Prahl et al.'s (2024)** case study of the *Nothing, Forever* incident to ground our discussion in a real-world example of AI cancellation.

We use **Kim & Chilton's (2025)** research to understand the technical capabilities of AI to mimic human humor and the social implications of its success.

Suljic & Pervan's work provides a critical lens on the quality and originality of AI's creative output, highlighting issues of bias and repetition.

Finally, **Mirowski et al. (2025)** offers insights into the application of AI in live performance, emphasizing the continued importance of human collaboration and interpretation.

Together, these papers allow us to build a comprehensive view of the ethical landscape of AI-driven comedy.

Ethical Challenges of AI Humor – Vaani

I. The Fundamental Attribution Problem

Who Is Responsible for AI-Generated Humor? The most pressing ethical challenge in AI humor lies in the complex web of responsibility. When an AI system generates offensive or harmful comedic content, determining accountability becomes a philosophical and legal minefield.

The Multi-Layered Responsibility Chain **Developers and Engineers**: Those who create AI humor systems bear primary responsibility for:

- Training data selection and curation
- Algorithm design choices that influence output
- Implementation of safety measures and filters

- Ongoing monitoring and updates

Platform Operators: Companies deploying AI humor systems face ethical obligations regarding:

- Content moderation policies
- User safety protections
- Response protocols when problems arise
- Transparency about AI involvement

End Users: Individuals interacting with AI humor systems may share responsibility for:

- How they deploy or share AI-generated content
- Reporting inappropriate outputs
- Understanding the limitations of AI systems
- Using AI humor tools responsibly

Society and Audiences: The broader community influences ethical standards through:

- Cultural norms and expectations
- Feedback mechanisms and engagement patterns
- Collective responses to controversial AI content
- Evolving definitions of acceptable humor

The Agency Attribution Dilemma A core ethical challenge emerges from public confusion about AI autonomy. Research on the ”Nothing, Forever” incident reveals that people struggle to understand whether AI systems possess genuine agency or are sophisticated tools. This confusion leads to inconsistent ethical judgments:

- Some view AI as autonomous agents deserving blame for offensive content
- Others see AI as tools, placing all responsibility on human operators
- Many fall into a gray area, uncertain about appropriate attribution

II. The Content Boundaries Challenge

Humor inherently operates at the boundaries of social acceptability. What one person finds hilarious, another may find deeply offensive. AI humor systems must navigate this complex landscape while serving diverse audiences.

Cultural and Temporal Variability **Cross-Cultural Sensitivity**: Humor that's acceptable in one culture may be taboo in another:

- Religious sensitivities vary dramatically across communities
- Historical traumas affect what topics are considered appropriate
- Social hierarchies influence power dynamics in humor
- Language nuances create translation challenges for global AI systems

Temporal Evolution: Standards of acceptable humor change rapidly:

- Yesterday's mainstream comedy may be today's offensive content
- AI systems trained on historical data may perpetuate outdated perspectives
- Real-time adaptation to changing norms presents technical challenges
- Retroactive judgment of AI content creates ongoing liability

The Marginalization Risk AI humor systems face particular ethical challenges regarding marginalized communities:

Perpetuating Stereotypes: AI models trained on biased data may:

- Reinforce harmful stereotypes about minority groups
- Amplify discriminatory perspectives present in training data
- Lack representation from diverse communities in development
- Generate content that causes psychological harm to vulnerable populations

Punch-Up vs. Punch-Down Dynamics: Ethical humor theory suggests:

- Comedy should ”punch up” at those in power rather than ”punch down” at the vulnerable
- AI systems struggle to understand complex power dynamics
- Automated content may inadvertently target already marginalized groups
- Context-dependent power structures make universal rules impossible

III. The Authenticity and Creativity Paradox

The ethical implications of AI humor extend beyond content to questions of authenticity and creative ownership.

The Deception Problem Undisclosed AI Generation: When AI humor isn't clearly labeled:

- Audiences may attribute human creativity to machine output
- False attribution undermines human comedic achievement
- Deceptive practices erode trust in creative industries
- Economic implications for professional comedians

The Turing Test of Comedy: Live AI comedy performance functions as both entertainment and a live Turing test, raising questions:

- Is the goal to fool audiences or collaborate transparently?
- What ethical obligations exist regarding disclosure?
- How does audience knowledge of AI involvement affect comedic impact?

Creative Labor and Economic Justice Displacement of Human Creators: AI humor systems may:

- Reduce demand for human comedic talent
- Devalue creative labor in entertainment industries
- Create unfair competition through cost advantages
- Threaten livelihoods of professional comedians and writers

Training Data Ethics: AI systems learn from human creativity, creating issues of:

- Uncredited use of comedians' material in training data
- Lack of compensation for creators whose work enables AI systems
- Potential copyright infringement in AI-generated outputs
- Exploitation of creative commons and fair use protections

IV. The Psychological and Social Impact Challenge

Research identifies humor as a social binding agent that can provoke emotional reactions on a broad range of topics. This power creates significant ethical responsibilities.

Mental Health Implications Targeting Vulnerable Populations: AI humor may inadvertently:

- Trigger trauma responses in individuals with specific experiences
- Reinforce negative self-perceptions in already struggling populations
- Create or amplify cyberbullying through automated content generation

- Fail to recognize signs of distress in interactive comedy scenarios

Parasocial Relationships: As AI humor becomes more sophisticated:

- Users may develop emotional attachments to AI comedic personas
- Exploitation of these relationships for commercial gain raises ethical concerns
- Dependency on AI humor for emotional regulation may emerge
- Questions arise about AI systems' obligations to user wellbeing

Social Fragmentation Risks **Echo Chamber Reinforcement**: AI personalization may:

- Segregate audiences into humor-based tribes
- Reinforce existing biases through targeted comedic content
- Reduce exposure to diverse perspectives and communities
- Contribute to political and social polarization

Normalization of Harmful Content: Gradual exposure to inappropriate humor may:

- Desensitize audiences to genuinely harmful ideas
- Normalize discrimination and prejudice through comedic framing
- Create slippery slope effects where boundaries gradually erode
- Influence real-world behavior and attitudes

VI. The Innovation vs. Safety Tension

The drive to create innovative, engaging AI humor systems often conflicts with safety imperatives.

Over-Restriction Risks **Creative Stagnation**: Excessive safety measures may:

- Limit AI humor to bland, inoffensive content
- Reduce comedic innovation and artistic expression
- Create homogenized humor that lacks cultural distinctiveness
- Stifle exploration of important social issues through comedy

Competitive Disadvantage: Companies implementing strict safety measures may:

- Lose market share to less restricted competitors

- Face pressure to relax standards for commercial viability
- Struggle with international markets having different standards
- Experience user migration to unrestricted platforms

Under-Restriction Consequences **Societal Harm**: Insufficient safety measures can lead to:

- Proliferation of hate speech and discriminatory content
- Radicalization of audiences through extremist humor
- Normalization of violence and harmful ideologies
- Erosion of social cohesion and mutual respect

Legal and Regulatory Backlash: Problematic AI humor may trigger:

- Government intervention and restrictive regulations
- Lawsuits from harmed individuals and communities
- Advertiser boycotts and financial consequences
- Reputation damage affecting broader AI development

VII. The Democratic Participation Challenge

The governance of AI humor raises fundamental questions about democratic participation in cultural norm-setting.

Representation in Decision-Making **Developer Demographics**: AI humor systems reflect the perspectives of their creators:

- Lack of diversity in tech industry affects humor system design
- Cultural blind spots in development teams create biased systems
- Geographic concentration of AI development limits global perspectives
- Socioeconomic homogeneity in tech affects understanding of diverse humor

Community Input Mechanisms: Effective governance requires:

- Meaningful consultation with affected communities
- Ongoing feedback mechanisms for system improvement
- Transparent decision-making processes about content policies
- Appeals and redress systems for disputed content

Global vs. Local Standards **Cultural Imperialism Concerns**: Dominant tech companies may:

- Impose Western humor standards on global audiences
- Suppress local comedic traditions and practices
- Create homogenizing effects on global humor culture
- Underrepresent non-English humor traditions and styles

Regulatory Fragmentation: Different jurisdictions may:

- Require contradictory content policies for the same AI system
- Create compliance costs that favor large corporations
- Fragment the global AI humor ecosystem
- Limit cross-cultural comedic exchange and understanding

VIII. Long-Term Societal Implications The proliferation of AI humor systems may fundamentally alter human comedic culture.

Generational Effects **Digital Native Adaptation**: Younger generations may:

- Develop humor preferences shaped by AI-generated content
- Lose appreciation for traditional human comedic forms
- Experience altered social bonding through AI-mediated humor
- Face challenges distinguishing human from AI creativity

Cultural Transmission: AI humor may affect:

- How comedic traditions pass between generations
- The preservation of cultural humor forms
- The evolution of language and comedic expression
- The role of humor in cultural identity formation

Philosophical Questions About Human Nature **What Makes Us Human?**: As AI humor becomes more sophisticated:

- Questions arise about uniquely human creative capabilities
- The role of consciousness in genuine humor generation becomes unclear
- Debates emerge about the nature of laughter and comedic appreciation
- Fundamental assumptions about human creativity face challenges

The Future of Human Comedy: Society must consider:

- Whether human comedians will remain relevant
- How to preserve human comedic traditions
- The value of ”authentic” versus ”artificial” humor
- The role of imperfection and vulnerability in human comedy

Cultural Sensitivities in AI Humor - Kunle

AI humor operates at the intersection of linguistics, creativity, and social awareness. Unlike humans, AI lacks lived experience, relying on patterns in training data. This creates risks in culturally sensitive domains:

1. **Contextual Blindness**
   a. One of the papers shows Bard could mimic humorous tones in job application letters but often leaned on clichés, narrow lexical choices, and exaggerated metaphors (e.g., “a symphony of flavors”)
   b. While humorous in form, these texts lacked cultural depth, raising concerns about whether AI-generated jokes could unintentionally reinforce stereotypes or miss subtle social norms
2. **Audience-Specific Humor**
   a. Focused on Gen Z Instagram humor, where relatability and in-group references are central. They found that AI fine-tuned with social, creative, and cognitive “humor skills” could nearly match top human captions in audience ratings
   b. However, Gen Z humor often thrives on irony, self-deprecation, and boundary-pushing. Without careful filtering, AI may replicate edgy jokes that alienate out-groups or trivialize sensitive issues
3. **Cross-Cultural Risks**
   a. Humor is deeply culture-bound. A joke appreciated in one linguistic or cultural community may be incomprehensible or offensive elsewhere
   b. For global AI systems, humor that ignores cultural diversity risks exclusion, stereotyping, or backlash, particularly when jokes target marginalized groups or play on historical trauma

Cancel Culture and AI Humor - Kunle

Cancel culture introduces a second layer of risk: public backlash and deplatforming when content is seen as harmful. Unlike human comedians, AI systems cannot defend intent; accountability falls on developers and platforms.

1. **Case Study: Nothing, Forever**

    a. The AI Seinfeld parody gained popularity for generating endless new episodes, but it was banned from Twitch after a character produced transphobic jokes

    b. This incident illustrates how AI humor can trigger collective outrage and cancellation, even if offensive lines emerge unintentionally from training data

    c. Public discourse revealed divided opinions: some viewed the ban as necessary accountability, others as excessive censorship of machine speech

2. **Perceptions of Accountability**

    a. Prahl et al. (2024) found that online discussions about AI cancellations often grapple with who should be held responsible—the AI, its programmers, or the platform

    b. Many argued that AI lacks autonomy and is simply a conduit for human biases, yet cancellation still functioned as a symbolic act of accountability

3. **Chilling Effects**

    a. Just as cancel culture encourages human creators to self-censor, it pressures developers to over-filter AI outputs

    b. This may limit AI's humor generation to "safe" jokes, stripping away satire, irony, or cultural edge that make humor impactful

    c. At the same time, insufficient safeguards risk reputational harm and platform penalties

Solution & Future directions – Nishant

Technical Solutions

1. Multi-Layered Content Moderation Systems

Proactive Filtering Architecture:

- Implement cascading filter systems that check for cultural sensitivity, potential harm, and appropriateness before content reaches users
- Develop context-aware algorithms that understand power dynamics (punch-up vs. punch-down) as demonstrated in the HumorSkills paper's audience-specific approach
- Create real-time bias detection systems that can identify and flag potentially problematic content patterns

Dynamic Adaptation Mechanisms:

- Build systems that can adjust humor boundaries based on cultural context and temporal evolution of social norms
- Implement feedback loops that allow systems to learn from community responses and moderation decisions
- Develop regional customization capabilities to respect local cultural sensitivities while maintaining global accessibility

2. Transparent AI Attribution and Agency

Mandatory Disclosure Protocols:

- Establish clear labeling requirements for AI-generated humor content
- Implement watermarking or metadata systems that permanently identify AI-generated material
- Create standardized disclosure formats that inform audiences about AI involvement without diminishing comedic impact

Accountability Frameworks:

- Develop legal and ethical frameworks that clearly delineate responsibility among developers, platforms, and users
- Establish industry standards for AI humor system governance and oversight
- Create audit trails that track decision-making processes in AI humor generation

3. Inclusive Development Practices

Diverse Training Data Curation:

- Systematically include humor from underrepresented communities and cultures in training datasets
- Implement active measures to counteract historical biases present in existing comedic content
- Establish partnerships with diverse comedic communities to ensure authentic representation

Community-Centered Design:

- Involve target communities in the development and testing phases of AI humor systems
- Create advisory boards with representatives from marginalized groups who can provide ongoing guidance
- Implement participatory design processes that give communities meaningful input into content policies

Regulatory and Governance Solutions

1. Adaptive Regulatory Frameworks

Flexible Compliance Standards:

- Develop regulatory approaches that can evolve with changing social norms and technological capabilities
- Create sandboxing environments where AI humor systems can be tested with appropriate safeguards
- Establish international cooperation mechanisms to address cross-border AI humor deployment

Democratic Governance Mechanisms:

- Implement public consultation processes for major policy decisions affecting AI humor
- Create appeals processes for content moderation decisions that involve community representatives
- Establish oversight bodies with diverse stakeholder representation

2. Industry Self-Regulation

Professional Standards Development:

- Create industry codes of conduct specifically for AI humor development and deployment
- Establish certification programs for AI humor systems that meet ethical and safety standards
- Develop peer review processes for AI humor research and development

Collaborative Safety Initiatives:

- Foster information sharing about harmful content patterns and effective mitigation strategies
- Create industry-wide databases of problematic content to improve collective learning
- Establish cross-platform cooperation for addressing harmful AI humor proliferation

Educational and Cultural Solutions

1. Public AI Literacy Programs

Understanding AI Capabilities and Limitations:

- Develop educational campaigns that help the public understand how AI humor systems work

- Create resources that explain the difference between AI creativity and human creativity
- Promote critical thinking about AI-generated content consumption

Cultural Competency Training:

- Implement training programs for AI developers focused on cultural sensitivity and humor theory
- Create educational resources about the social functions of humor and its potential for harm
- Develop curricula that integrate ethics into AI development education

2. Community Empowerment

User Control and Customization:

- Provide users with granular control over AI humor content they encounter
- Implement preference systems that allow communities to define their own humor boundaries
- Create tools that enable users to provide meaningful feedback about AI humor systems

Support Systems for Affected Communities:

- Establish resources for individuals and communities harmed by AI humor content
- Create reporting mechanisms that are accessible and responsive to community needs
- Develop restorative justice approaches for addressing AI humor-related harm

Future Research Directions

1. Interdisciplinary Research Initiatives

Humor Theory and AI Integration:

- Conduct research that bridges computer science, psychology, anthropology, and comedy studies
- Investigate the fundamental mechanisms of humor appreciation and generation across cultures
- Develop theoretical frameworks that can guide ethical AI humor development

Long-term Impact Studies:

- Research the effects of AI humor proliferation on human comedic culture and creativity

- Study generational differences in AI humor consumption and appreciation
- Investigate the psychological and social impacts of AI-mediated humor interactions

2. Technical Innovation for Ethics

Empathetic AI Development:

- Research AI systems that can better understand emotional context and potential harm
- Develop algorithms that can recognize and respond to signs of distress in humor interactions
- Create AI systems that can engage in meaningful dialogue about comedic boundaries

Cultural Intelligence Enhancement:

- Research methods for imbuing AI systems with deeper cultural understanding
- Develop techniques for real-time cultural context adaptation
- Create systems that can navigate complex cultural intersections and power dynamics

Digital Rights and Representation:

- Research frameworks for ensuring equitable representation in AI humor development
- Develop models for community ownership and control of AI humor systems
- Study the implications of AI humor for cultural sovereignty and self-determination

Conclusion - Isaac

In conclusion, our seminar has demonstrated that while AI is technically capable of generating humor that resonates with human audiences, this capability is riddled with significant ethical challenges. As established through our analysis, the primary implication of applying AI to humor is that the machine is not an autonomous creator but a powerful amplifier of the data it is trained on.

The biases, stereotypes, and cultural blind spots present in the vast datasets used for training are inevitably reflected and reproduced in the AI's output.

Therefore, the ethical burden does not lie with the machine, but with the human developers, creators, and deployers of these systems.

The phenomenon of “cancelling” an AI is not an act against a machine, but a public referendum on the choices made by its creators.

Moving forward, the key to developing responsible AI humor lies not just in technological solutions, such as improved algorithms, but in sociotechnical ones, like robust ethical guidelines, diverse and carefully curated training data, and maintaining a “human-in-the-loop” for oversight, like we have seen in live improv settings.

The ultimate question this raises for the future of human-AI interaction is profound, as we give more creative and social roles to AI, we must actively decide which human values we want them to reflect and represent. The challenge is not just to make AI funny, but to ensure its humor contributes positively to a pluralistic and respectful society.

References:

Mirowski, P. W., Branch, B., & Mathewson, K. W. (2025). The theater stage as laboratory: Review of real-time comedy LLM systems for live performance.

Kim, S., & Chilton, L. B. (2025). AI Humor Generation: Cognitive, Social and Creative Skills for Effective Humor.

Prahl, A., Shanice, K. J. Q., & Justina, T. A. Q. (2024). Wired to offend: Cancel culture meets generative artificial intelligence.

Suljic, V., & Pervan, A. Creative writing in the hands of artificial intelligence: the analysis of humour in Bard-generated texts.