Jongbeen Song
2025020399
Prof. Shin
01 October 2025

Research Proposal

Background: The “Dead Internet Theory” suggests that much of the content on today’s internet is not produced by humans but by algorithms and bots. This speculation provides a critical lens to examine how AI-driven, automated discourse production alters public perception and social debate. In particular, the rapid spread of controversial and provocative content on news and social media blurs the line between human-generated and machine-generated texts. This project aims to experimentally implement this tension by building two systems: **Internet Killer**, which generates provocative discourse, and **Internet Guard**, which attempts to detect and counter it.

Methodology:

1. **Internet Killer**
    - Building an API that retrieves the top five trending news topics
    - Allowing users to either select one of these topics or input a topic in natural language
    - Generating multiple pieces of text on the chosen topics using different rhetorical strategies:
        - Provocating framing
        - Conspiracy undertones
        - Populist rhetoric
        - Irony, or satire
        - Emotional amplification
    - Aiming to simulate the kinds of online posts or comments that attract strong reactions
2. **Internet Guard**
    - Training a discriminator model to distinguish between machine-generated and human-generated texts.
    - Using adversarial training with data consisting of Internet Killer outputs and real online comments/posts.
    - Introducing intentional artifacts if the Guard consistently fails into the Internet Killer’s outputs to mandate disclosure and traceability of AI-generated texts.

Projected Timeline:

- October 20: Developing Internet Killer
- November 3: Developing Internet Guard
- November 17: Code Refactoring & Preparing for Deployment
- December 1: Completing Lightning Talks Video & Final Product

Goals:

The project has two primary goals. First, to demonstrate how automated systems can generate controversial discourse and amplify conflict, providing an experimental framework to interrogate the Dead Internet Theory. Second, to test the boundaries of detectability between human and machine texts, and to critically explore what it means when such boundaries collapse. By introducing the possibility of mandatory artifacts, the project highlights the ethical and practical challenges of AI content disclosure. Ultimately, this creative project seeks to interrogate how the dynamics of **generation and detection** shape contemporary debates in AI and the media environment.

Peer Review Feedback Questions:

I plan to provide the final product in the form of an API. However, I am not yet certain how many simultaneous requests it can handle. Scaling the system up to a true product-level service would not be suitable for the scope of this course and would consume excessive resources. To address this, I am considering offering API access to a limited number of volunteer students as a demo. If you have better ideas for demonstrating the system, I would greatly appreciate your suggestions.

Regarding the generation strategies of Internet Killer, are there any existing studies in the field of media and communication that could serve as useful references? For example, research on trolling, deliberate misinformation, or other related discursive practices. I would also welcome any personal ideas for generation strategies you might have.

On the issue of artifact insertion, I believe the most effective approach would be one that is invisible to users but still detectable within the system. For instance, Naver Blog has a policy where copying and pasting automatically includes the source link in the clipboard, but this is easily bypassed by simply deleting the link. Similarly, if AI-generated images include visible watermarks along the edges, they can be bypassed by simple cropping. For textual content, what would be the most effective way to insert such artifacts? I would appreciate your ideas on potential strategies.