Coller School of Management
Tel Aviv University

Coller Lab for Artificial Intelligence and Business Analytics

Deliberative AI – Generating an HCI Content Moderating Codebook

Hofit Wasserman-Rozen, Inbal Yahav

Abstract

we propose a method to formulate a crowd-AI-driven codebook to define acceptance/rejection rules for moderating potentially misleading information. This initiative involves assembling a panel of human decision-makers, supported by AI, to establish a comprehensive codebook. It recognizes the inherent ambiguity of the informational "ground truth", often inaccessible as well as perpetually shifting over time and context. Therefore, leveraging the linguistic capabilities of Language Learning Models (LLMs), an AI-mediated online interface will explore nuanced scenarios, aiming to de-bias content and foster a deeper understanding of the crowds accept/reject boundaries. The interaction between the crowd and the AI aims to ensure: (1) a comprehensive codebook, (2) wide agreement among humans, (3) high consistency across decisions, (4) optimized results by leveraging reciprocal learning principles, and (5) dynamic adaptation to changes in the data. Its enforcement will also be managed by AI to secure consistent and objective applications.

Background and rationale

Disinformation is intentionally false information. On the other hand, misinformation involves incorrect or misleading information but is not necessarily created with the intent to deceive. This distinction presents a unique challenge for governance. Unlike disinformation, misinformation is often ambiguous, containing elements of truth mixed with mischaracterizations that can unintentionally deceive and distort reality.

Detecting disinformation has been the focus of many research efforts. Common approaches involve using natural language processing (NLP) techniques, which are known to struggle in this context, and strategies that aim to identify the spreaders—whether bots or humans—based on their social characteristics. Misinformation, however, presents a more formidable challenge because the spreaders are often unaware that they are disseminating false information. Consequently, they are likely to share similar characteristics with the general population, making it difficult to distinguish them through social traits alone. Thus, the detection of misinformation must rely more heavily on analyzing the subtle textual signals within the information being spread.

Coller School of Management
Tel Aviv University

Coller Lab for
Artificial Intelligence
and Business Analytics

Objective

This Deliberative AI Initiative is dedicated to addressing the murky territory of misinformation. Clearly, effectively governing misinformation requires enforcing comprehensive and consistent rules, devoid of political, social, or moral biases. Two important questions therefore lie at the heart of properly moderating misinformation: 1. Who creates the rules?, and 2. Who enforces the rules? Fleshing out the rules is a complex and nuanced challenge by itself, amplified by the requirement for a neutral and consistent implementation of those rules once established.

The ambiguity of the misinformation "gray zone" underscores the need for effective governance. Without clear rules, similar content may be inconsistently treated—flagged as misinformation in one instance and deemed permissible in another. The objective of this initiative is to rectify this by establishing a consensus-driven codebook through collaboration among diverse human participants.

Methodology

Deliberative AI harnesses LLMs to create an interface that fulfills these criteria. Humans excel at judgment, seeing the big picture, and contextual understanding, leveraging their ability to think abstractly. In contrast, machines excel in pattern recognition and linguistic analysis, approaching problems with a more concrete perspective. LLMs can educate participants about misinformation, generate diverse scenarios, mitigate biases, and elucidate participants' belief systems, complementing human capabilities by providing detailed and specific insights.

The Deliberative AI Initiative synthesizes human judgment with AI capabilities through the following process:

1. A diverse and representative panel of human decision-makers will be tasked with creating a comprehensive misinformation-moderating codebook.
2. An AI group-moderating interface will assist panel members in deliberating amongst themselves. The unique linguistic capabilities of LLMs will support a dialectic process, generating an endless trove of potential examples, follow-up questions, scenarios, and analogies. It will also expose and highlight misalignments and inconsistencies in views. This dialogue process is designed to help the human decision-makers panel clearly define their decision boundaries, red lines, and objectively acceptable rules. The interaction will follow the guidelines of RHML - Reciprocal Human-Machine Learning (Te'eni, Yahav, and colleagues, 2023), supported by the new version of RHML software, called Fusion (the beta version is available here: https://fusion.nisaba-technologies.com, access granted with proper authorization. A previous open version can be found here: https://github.com/TAUCollerLab/Fusion).

3. The agreed-upon codebook will be enforced by an AI, thus eliminating subjective human influences due to personal incentives, fatigue and burnout.

Summary

The Deliberative AI Initiative aims to combat misinformation by assembling a diverse panel of human decision-makers, supported by AI, to establish a comprehensive codebook. This codebook sets boundaries for acceptable and unacceptable posts, with enforcement handled by AI to ensure consistency and objectivity. Leveraging the linguistic capabilities of LLMs, an AI-mediated interface facilitates dialogue among panel members to define decision boundaries and enforce agreed-upon rules. By leveraging RHML principles, this initiative aims to mitigate human biases and inconsistencies, foster wide agreement among humans, ensure high consistency across decisions, and dynamically adapt to changes in the data.