Chatmods, Counterspeech & Disinformation Disruption

John Wihbey, Northeastern University

Workshop on Evolving Technological, Legal and Social Solutions to Counter Disinformation in Social Media

National Academies of Sciences, Engineering, and Medicine

April 2024

Northeastern University

Platform-directed Moderation Using Gen AI Chat Tools

- **Idea:** Platforms deploy gen AI interactive agents: "Chatmods"/ "Modbots" that interact with users about problematic behavior
- **Ecosystem context:** New intervention tool (with high degree of risk) adding to classic triad of platform tools: Reduce-remove-inform … chat
- **Possible advantage I:** Disinformation campaigns rely on ambiguity and often seek to amplify true or partially true information for strategic purpose. Reduce-remove-inform is limited toolkit in face of ambiguity
- **Possible advantage II:** Chatmods could 1) create friction, 2) activate awareness, and 3) impose costs, while also being speech-preserving
- **Higher level:** Reinforcing fundamental democratic norms: remedy to bad speech is counterspeech and dialogue; differentiating from censorship

Northeastern University

Chatmods: Context and literature

- **Literature:** Potential uses of gen AI in social tech: Counterspeech, mediation, information assistance, discourse moderation (Kapoor & Narayanan, 2023)
- **Industry:** Deep learning already used in classifiers, removal, application of static labels; public is used to "unauthorized" bots; OpenAI is developing content moderation capacity; Snap has MyAI bot; Meta has experimented with Jane Austen bot, etc., bots in metaverse; Reddit and Discord
- **Public opinion:** Northeastern survey across US, UK, Canada found:

  -Publics OK with chatbots starting a private or public conversation with users to address abusive behavior publicly (~50%)

  -Acceptance of chatbots in social space correlated with prior experience with company-sponsored bots (customer service), suggesting norming/conditioning is key

  -Major concerns (85%+) AI chatbots won't understand context/words; and chatbots may create divisiveness: 60% (US); 56% (UK); 62% (CA)

Northeastern University

Open Questions for ChatMods Approach

- **Form:** What posture/voice should we normatively prefer for such agents – police, referee, informational, gentle inquisitor? What is the UI/UX?
- **Space:** What are criteria for chatmods taking action publicly vs. privately?
- **Tradeoffs:** Ethical AI frameworks must be applied to authorized agents exerting power in human social space. Considerations of nonmaleficence, justice, autonomy.
- **Monitoring:** What could accountability and transparency look like? Considerations of data forms, third-party auditing, Digital Service Act rules.
- **Partnerships:** Could platforms partner on LLM reinforcement learning with civil society groups in tricky domains, e.g., elections, public health, terrorism, etc.?
- **Public policy:** Given that agents would be clearly "publishing," how does this affect Section 230 safe harbor? Is a further exemption required?

Northeastern University