AI & SOCIETY
https://doi.org/10.1007/s00146-023-01674-5

CURMUDGEON CORNER

Losing the information war to adversarial AI

Peter Mantello[1] · **Manh-Tung Ho**[2]

Received: 7 December 2022 / Accepted: 11 April 2023

1 Introduction

Conversational AI is augmenting and replacing human endeavors in finance, mental health counseling, advertising, dating, journalism, and wellness. More commonly referred to as "chatbots" or "social bots," these computational agents can automatically promulgate ideas, generate messages, and act as followers of users. With over 100 trillion parameters trained on a huge amount of texts, large language models such as OpenAI's ChatGPT, Microsoft's Bing or Google's Bard exhibit sophisticated linguistic and conversational capacities that can mimic human behaviors. The popularity of conversational AI has grown exponentially since the release of ChatGPT which has reached more than 100 million users within just three months after its launch in November 2022. More advanced versions, called empathy bots, are growing their ability to read, evaluate, and respond to a user's emotional state, heightening not only their utility as artificial headhunters but importantly, anthropomorphic appeal.

While these artificial agents are intended as tools for societal good, they are also becoming the weapon of choice for agent provocateurs in information warfare and cybercrime. Extremist groups, rogue governments, and criminal organizations are using chatbots to heighten the speed and scale of online mis/disinformation, create fake social media accounts, harvest personal data from unsuspecting users, impersonate friends/associates, and manipulate/disrupt political communication. Already, prominent scientists and Silicon Valley pundits are calling for a 6-month pause to consider the growing risks while Italy, China, Russia, Iran, and North Korea have banned the use of ChatGPT within their sovereign borders.

Problematically, the war against adversarial AI chatbots is failing miserably. It is not simply that chatbot programmers are better at making artificially intelligent agents behave and respond like humans to avoid detection. The battle is also being lost by content moderation stakeholders and their competing interests and agendas. Social media companies and security agencies are locked in an unofficial zero-sum game, where one stakeholder's gain is another stakeholder's loss. At the heart of this contest are the moderation systems themselves, which often go against the economic interests of big-tech companies. This, in turn, leads to a legitimacy gap between censorship claims and actual filtering practices.

Here are some reasons why we believe the fight against the malevolent use of chatbots may be unwinnable. The first reason surrounds the highly competitive nature of the industry that dictates social media providers to continually expand their roster of automated features and services to attract, sustain, and grow their customer base. This includes making artificial intelligent agents more human-like. First, such anthropomorphizing strategies include creating chatbots that post at irregular time intervals and with less consistency, or purposely make spelling errors or use trendy words and phrases. Second, another industry tactic is lowering the entry bar to bot-making and given the lowering computational and technical barrier of entry, this trend will only continue into the future. A good example is Telegram, a popular site that provides users with free and easy-to-create "Matchmaker" bots on its privately encrypted channels. According to Telegram's instructional literature, a user can seek new contacts simply by entering a username in the Telegram settings and uploading a photo. Alternatively, a Telegram user can simply type into a Telegram desktop app, "BotFather", a multi-purpose application program interface (API), that helps users to create new bot accounts and manage existing bots.

Yet these same services and features are what make social media platforms attractive to extremist organizations, criminals and rogue states. Although Telegram is heavily policed by security agents and platform administrators,

✉ Manh-Tung Ho
tung.homanh@phenikaa-uni.edu.vn

[1] Ritsumeikan Asia Pacific University, Beppu, Oita 874-8577, Japan

[2] Centre for Interdisciplinary Social Research, Ha Dong, Hanoi 100803, Vietnam

Published online: 28 April 2023

the "Matchmaker" bot remains the go-to tool for extremist organizations and other bad actors. Net-savvy agent provocateurs use it to provide tactical information to aspiring lone-wolf attackers, deliver updates on operational missions to supporters, share videos with sympathizers, and/or forewarn followers of sites being surveilled by security agencies. With advances in affective computing, conversational AI can evaluate and respond to human emotions and feelings. Originally, purposed for mental health counseling, these bots can facilitate an emotional bond in the indoctrination process, making applicants feel more comfortable and less mistrustful than when dealing with a real person (Ho, et al. 2021; Mantello et al. 2021). In a study by iN2 (2018), researchers found that Islamic extremist bots were more successful at convincing prospective members to join up than human headhunters. The researchers attributed this motivational efficacy to the bot's upbeat tone, quick response, and non-judgmental attitude to applicant queries.

The second issue concerns the way social media companies leverage the arbitrary nature of government overwatch. For decades now, social media companies and security agencies have waged counter-offensive after counter-offensive against adversarial bots using both human and non-human tactics. Yet the velocity, scalability, and resilience of bad bots make disruption efforts extremely difficult. Moderators use the phrase, 'playing digital whack-a-mole' to describe their frustration of fighting non-human armies that continually repopulate themselves. As malicious interactive bots gain the higher ground in twenty-first-century information warfare, security agencies exert greater pressure on social media companies to gain access to and oversight of their platforms and content moderation systems.

Problematically, law enforcement agents tend to concentrate their regulatory scrutiny on social media platforms operating in Western countries. As a result, social media providers neglect their moderation and disruption efforts in non-Western regions of the world. Yet it is often in these same places where not only political turmoil and violence are endemic but malicious content and speech are also prevalent. The lack of intense regulatory scrutiny in non-Western countries also allows social media companies to lower their operational costs by avoiding the development of costly AI moderation systems and hiring on skeleton crews of underpaid and poorly trained human moderators often not fluent in local dialects. Aggravating this situation is the untold psychological demands placed on human moderators, having to suffer daily images of graphic violence. Inevitably, these unfavorable working conditions create high turnover which, in turn, means social media companies must constantly replenish their ranks. Lax practices and concerns to identify and disrupt adversarial chatbots have led many host governments to shut down or suspend a social media company's operation. Conversely, social media providers are known to jettison ethical principles to stake a claim or retain a foothold in authoritarian countries. For example, Facebook and Twitter have a long track record of appeasement in Myanmar, Egypt, Jordan, and Saudi Arabia (York, 2021). This placation usually entails violating user privacy, blocking content of rival political parties, or deleting posts that challenge a ruling regime's official narrative. Increasingly, these actions are done under the rhetorical banner of combatting the spread of mis/disinformation.

On the other hand, social media companies have proven reluctant to permit outside researchers a deeper inspection of their moderation systems and practices. For many years social media companies have garnered an infamous reputation for stone-walling external researchers' efforts, hiding data or handing out corrupt or limited datasets. Indeed, researchers are not the only ones prevented from inspection. Governments are increasingly experiencing push-back from legacy social media companies as whistleblowers such as Francis Haugen reveal more of how the algorithms of these platforms intentionally incite and feed off primal emotions such as fear, hate, and anger.

2 Conclusion

The growth of artificially intelligent agents that can sense, read, and respond to human emotions illustrates the speed at which AI-driven social media platforms are accelerating and nuancing developments in affective capitalism. Concomitantly, it also personifies how less-resourced non-state belligerents and criminal organizations are exploiting these same AI tools and strategies for malevolent and egregious purposes. As chatbots become increasingly human-like, their appeal and efficacy as automated propagandists and agent provocateurs will only increase. Thus, this short article highlights the need for further research on the contradictions and tensions that lay not only at the heart of content moderation systems but also on a more phenomenological level, where the increasingly anthropomorphic quality of conversational AI may accelerate the dangers of online radicalization.

Acknowledgements This study is part of the project "*Emotional AI in Cities: Cross Cultural Lessons from UK and Japan on Designing for an Ethical Life*" funded by JST-UKRI Joint Call on Artificial Intelligence and Society (2019), Grant No. JPMJRX19H6. Author Manh-Tung Ho would like to express his gratitude toward the SGH Foundation for their support of his doctoral study.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Data availability statement There are no data associated with this paper.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Ho M-T, Mantello P, Nguyen H-KT, Vuong Q-H (2021) Affective computing scholarship and the rise of China: a view from 25 years of bibliometric data. Humanities and Social Sciences Communications 8(1):282. https://doi.org/10.1057/s41599-021-00959-8

iN2 (2018) The Envoy and the Bot: Tangibility in Daesh's Online and Offline Recruitment. https://thescli.org/the-envoy-and-the-bot-tangibility-in-daeshs-online-and-offline-recruitment/. Accessed January 18, 2022

Mantello P, Ho M-T, Nguyen M-H, Vuong Q-H (2021) Bosses without a heart: socio-demographic and cross-cultural determinants of attitude toward Emotional AI in the workplace. AI & Soc. https://doi.org/10.1007/s00146-021-01290-1

York JC (2022) Silicon values: The future of free speech under surveillance capitalism. Verso Book.