

Conversational AI: Social and Ethical Considerations

Elayne Ruane, Abeba Birhane, Anthony Ventresque

School of Computer Science, University College Dublin, Ireland
Lero - The Irish Software Research Centre
{elayne.ruane, abeba.birhane}@ucdconnect.ie

Abstract. Conversational Agents are becoming ubiquitous in our daily lives. They are used in various areas including customer service, education, medicine, and entertainment. As tools that are increasingly permeating various social domains, Conversational Agents can have a direct impact on individual's lives and on social discourse in general. Consequently, critical evaluation of this impact is imperative. In this paper, we highlight some emerging ethical issues and suggest ways for agent designers, developers, and owners to approach them with the goal of responsible development of Conversational Agents.

Keywords: Conversational Agent, Intelligent Systems, Social Impact, Ethics

1 Introduction

Conversational AI allows human users to communicate with an automatic system using natural language. The interaction may be speech and/or text based. It may be served to the user through messaging channels (e.g. Facebook Messenger and Skype), through dedicated phone or web applications, integrated into a website, or shipped as part of an operating system. Conversational AI systems have many names depending on their capabilities, domain, and level of embodiment. These terms include automatic agent, virtual agent, conversational agent, chatbot, or, for very simple systems, bot. In this paper we use the term Conversational AI to refer to any use of Machine Learning (ML) and Deep Learning (DL) models, Natural Language Understanding and Processing (NLU & NLP) techniques, and dialogue management systems to understand user input and generate natural language responses. We use the term Conversational Agent (CA) to refer to systems that have a Conversational AI component and have other features such as a user interface (UI) to facilitate interaction and server-side features such as the app logic and the database.

The year 2016, dubbed the “Year of the Bot” after Microsoft CEO Satya Nadella described bots as the new apps, saw the launch of more than 30,000 chatbots on the Facebook Messenger platform alone [8] [11]. By 2018, there were more than 300,000 active bots with 8 Billion messages exchanged every

month on the platform [5]. Much of this growth is driven by commercial interests. Chatbots are an inexpensive, fast, and always-on service for answering FAQs and completing other well-defined tasks. Although quality remains an issue for more complex tasks and conversational system evaluation is an active area of research [39], CAs have seen adoption in various social domains including customer service and product recommendation, education support, medical services, entertainment, social outreach, and personal organisation.

Ethical concerns inevitably arise with any technological innovation. However, they are often considered secondary to technical development challenges, if they are considered at all [7] [36] [51]. As with any technology that permeates our daily lives, the development and application of conversational AI raises various ethical questions. While some concerns, such as privacy, are an active area of research [29], others have received less attention. This paper examines the ethical challenges posed by the integration of conversational systems into human interaction as well as the necessary cautions and measured steps that need to be considered in developing and deploying CAs. We hope that this paper can serve as a call to action for agent designers, developers, and owners. Section 2 motivates this work by highlighting the potential harms of Conversational AI and Section 3 discusses relevant work from the literature. In Section 4, we identify a number of concerns and propose a way forward for critical and ethical engagement throughout the design and development process.

2 Motivation and Contribution

The history of humanity is full of examples of technology as a force for societal and behavioural change from the earliest prehistoric stone tools through to the invention of the computer, internet, and other advances in Information and Communications Technology. The pace of change has accelerated, and successive generations leading quite different lifestyles due to the impact of technological change. The most recent example may be the wide-spread adoption and usage of smartphones [43] [17] and social media [24] which gives individuals communication capabilities and access to information and other media that is changing the social and political landscape.

Spurred on by the use of smartphones, the last decade has seen the adoption and integration of CAs in our day-to-day lives. The release of Apple’s virtual assistant Siri in 2011, shipped with the iPhone 4S, marked the start of the ubiquity seen today where conversational AI is present in our homes, offices, and social media platforms shaping how we interact with companies and services. A survey (n=800) of marketing professionals by Oracle found 36% of brands surveyed had implemented chatbots for customer service with an increase to 80% expected by 2020 [37]. According to a report from Global Market Insights, Inc., the intelligent virtual assistant market which includes Apple’s Siri, Google Assistant, Amazon’s Alexa, and Microsoft’s Cortana, is expected to grow from a \$1 billion valuation in 2017 to \$11.5 billion by 2024 [20]. This growth directly

and indirectly impacts how individuals interact with services, consume media, and interact online.

Such ubiquity and increasing integration makes CAs forces that shape, alter, and impact the experience of individuals and groups. Their impact and potential harm varies depending on the domain and target user group. Relatively simple social bots, for example, have transformed the political landscape. In a study that examined the impact of bots in the 2016 US presidential election, Bessi and Ferrara (2016) [4] found that the presence of social media bots negatively affects democratic political discussion. Although social bots are often benign and useful, they can be used to manipulate and mislead users by spreading misinformation which has been particularly effective on Twitter and Facebook.

Like any other AI system, CAs do not exist in a social, political, economic, and cultural vacuum. They are developed by individuals or teams of individuals with specific, often commercial, aims. CAs necessarily reflect the values and perspectives of such individuals and the interests of the respective industry. When chatbots are rolled-out to users, they become part of the social utility where the implications of their design can be felt by real people. However, a combination of a lack of awareness of the technology behind these agents among the general public, company-level confidentiality, and the emerging nature of this technology has created an environment in which ethical concerns are not well-defined around Conversational AI. As such, we argue those involved in the process of developing and deploying CAs have a responsibility to critically examine the social impact of their tools and to view such practice as an integral part of the development process.

3 Background and Related Work

Although many major companies, research institutions, and public sector organizations have all issued guidelines for ethical artificial intelligence, recent work [22] has discovered substantive divergence in how these are written and interpreted, highlighting the complexity of designing guidelines for systems with complex social impact. An emerging body of work indicates that the integration of AI systems into various social spheres brings with it a host of often unanticipated and harmful outcomes. Furthermore, users from disadvantaged backgrounds, such as those with disabilities or those that face racial, gender, or other bias, may face disproportionate harm. Various studies illustrate this, as bias is found in: detecting skin tones in pedestrians [52], predictive policing systems and justice [38], the display of STEM career ads [27], recidivism algorithms [2], politics of search engines [21], medical applications [13], automatic speech recognition [44], and in hiring algorithms [1]. This emerging body of work that critically examines unfairness, injustice, bias and discrimination within various areas of AI is invaluable. However, there are a number of ethical considerations that are unique to machine-human conversation that have not yet become de facto considerations in the design and development stages of building a chatbot or other CA.

Conversational AI and Human-Computer Interaction (HCI) are active fields of research within academia. However, most publicly deployed CAs are developed by industry stakeholders among which there is little cross-collaboration or publication of proprietary training datasets and system architectures. This makes critical engagement and analysis of social impact difficult. Given their ubiquitous presence in various social, political, and financial spheres, we contend that CAs might be best viewed primarily as social utilities, and not solely as corporate assets. The effect of unintended consequences as a direct result of design decisions holds the potential to harm people. Consequently, critical engagement is required throughout design and development.

Language is central to Conversational AI systems as a medium that facilitates interaction. Effective and responsible design of CAs requires an understanding of various linguistic elements of conversation as well as an awareness of wider social and contextual factors [18] [46]. Language, as a social activity embedded in historical, cultural, and social norms is not a “neutral” or “objective” medium. Rather, it reflects existing societal values and judgements [30]. Take, for example, how the meaning of, and the discourse around, the word “gay” has changed since the 1950s. Language is situational and contextual - a single word or conversation can have radically different meanings depending on context and time. “Acceptable” norms and forms of conversing in one context might be perceived as “unacceptable” or “deviant” in another. Consequently, conversation formats, phrases, and words that are perceived as “acceptable” or “standard” might represent the status quo, leaving anything outside the status quo either implicitly or explicitly coded as an anomaly or outlier [6]. Decisions made during development regarding various aspects of language such as accent, dialect, and register can encode socially held beliefs and assumptions of, for example, “standard language” into the system. Language registers and expressions that are used by target user groups but not recognized by an agent are implicitly deemed outside the “norm”. The language(s) accepted and understood by the system reflect the accessibility of the system and this is a deliberate choice during the design phase that may have significant knock-on effect for users after deployment. In the process of developing CAs, these nuances of language and conversation, and the problems that arise due to lack of awareness around them, should take centre stage alongside the technical challenges.

Language is inherently social, cultural, contextual, and historical, which means that the design of agent dialogue necessarily reflects a particular worldview. As tools that exist within the social realm, socially sensitive conversations are unavoidable. How these socially sensitive issues are responded to plays a significant role in terms of how such sensitive issues and individuals affected by them are perceived. Recent work [9] studied how CAs handle sensitive requests involving sexual harassment and bullying. The authors found that while commercial conversational systems often avoid answering such requests altogether and rule-based systems usually try to deflect these topics, data-driven systems risk responding in a way that can be interpreted as flirtatious and sometimes counter-aggressive. Similarly, it was found that race related conversations are of-

ten deflected by chatbots [41]. Given their pervasiveness, these topics are something an open-domain agent should be designed to handle responsibly. Although the rationale behind such design is to take a “neutral” stance, avoidance and deflection of complex social issues can symbolize either endorsement, trivialization, or devaluation of the topic or an individual’s experience.

The use of CAs within mental health services is another area where critical reflection is required. The gap between the demand for mental health services and lack of available resources, as well as the cost efficiency and seemingly non-judgmental nature of CAs, makes them seem an attractive solution. So far, CAs have been bestowed with responsibilities including screening diagnosis and treatment of mental health [19] [35] [47]. However, despite being perceived as less-stigmatizing, CAs might actually pose harm to users due to their limited capacity to re-create human interaction and to provide tailored treatment, especially if they are not continually audited and evaluated [25]. Mental health services meet people at their most vulnerable. Consequently, any conversational interactions with such users needs utmost ethical and critical attention. However, ongoing evaluation for harms and benefits, which is essential for ethical and responsible practice, is absent in many digital platforms and apps for mental health [25]. Unfortunately, this is not limited to this application domain. Among the varied applications of CAs, one common recurring theme is a lack of critical assessment. Evaluation of the use of CAs often mentions the importance of ethical considerations but fails to explicitly discuss such concerns or provide mechanisms to address them such as in [19] [35] and [54].

There have been numerous approaches proposed to implement ethical decision making for AI agents. Some argue the best approach is within the context of Safety Engineering whereby safety mechanisms are used to mitigate harmful impact of AI systems. Others argue for a Machine Ethics approach which involves encoding ethical standards and reasoning into the AI systems themselves [32] [3]. In this paper, we argue for a shift in mindset that considers social context in identifying and addressing ethical concerns specific to conversational AI throughout the design and development process. We place responsibility on designers and developers for cultivating awareness of these issues and how their approaches impact the end user, as opposed to discussing general ethical approaches and focusing on agent decision-making. In the next section, we discuss aspects of conversational AI that require critical reflection throughout the design and development phases. This is not a complete list of concerns that arise with Conversational AI by any means. Rather, these are some concerns we have focused on as a point of discussion with the aim of bringing forth and clarifying implicit assumptions and the impact they may have on users.

4 Towards Ethical Conversational Agents

4.1 Plurality of approaches

Ethical concerns that emerge with Conversational AI vary markedly depending on the application domain, target user group, and the goal(s) of the agent. As

such, an understanding of the domain and the problem that the agent aims to solve should inform the identification of possible ethical concerns and solutions. For example, a chatbot used within an organisation by employees for a specific purpose will have a considerably different set of considerations than a customer or public-facing agent that may be expected to answer general or unconstrained queries. For responsible system design, deep understanding of the user groups characteristics, contexts, and interests is imperative. For example, a recent survey on the use of CAs in education and associated user concerns revealed that people were open to this technology if privacy issues are addressed but found that there were significant differences in how adults and children viewed privacy in this context [28]. Such insight, and its incorporation into the design of the system, is critical for ethical and responsible design that centres the values and interests important to users. As such, embracing contextual, flexible, and plural methods of identifying and addressing ethical concerns is imperative. Additionally, identifying solutions that are the most suitable to the specific scenario should always be prioritized over attempting to fit some standard principles. While failure to anticipate and mitigate potential ethical issues can result in destructive, traumatic, or dangerous outcomes in some circumstances, emerging issues might be easily contained and corrected in others. Consequently, there is no one-fits-all ethical standard or principle that can be applied to all CAs. Therefore, in the strive to develop ethical and responsible Conversational AI, we encourage contextual and plural approaches over a set of abstract principles.

4.2 Trust and Transparency

Providing users with choices, and consequently with control, over how they prefer to interact with an agent, is an important first step towards centring users needs and wellbeing. Transparency about an agents status as automatic (non-human) and the limits of its capabilities, for example, is essential in order to allow users to make informed choices, which further contributes to users trust. Recent work has shown that users behave and interact differently when conversing with an automatic agent compared to interacting with another human [34]. If users are aware that they are speaking to an automated agent or a human agent, then they might be able to make informed decisions with regards to their own behaviour, in particular regarding information disclosure [14]. This is especially crucial where the user information being discussed or disclosed is sensitive, such as in banking or education of minors, or where the implications and/or consequences of the conversation are significant such as user health concerns.

Understanding user expectations of an agent is crucial in ensuring that user trust is not taken advantage of. Reasonable expectations should be identified and validated before the agent is published. For example, if a user expects a conversation to be anonymous, then identifiable plain text conversation logs should not be visible to individuals on the development team. Similarly, if a chatbot has been designed to recommend products, such as the retailer H&M’s chatbot which helps users to plan and purchase outfits, a user may expect relatively unbiased information such that the chatbot will not show clothes from

other retailers but also that it won't only show the most expensive H&M clothes either. The user's assumption of agent neutrality is part of a widely held but often misguided belief that AI systems are unbiased. It can be difficult to evaluate the behaviour of a system such that we can validate whether the agent recommends products based on genuine interests or needs instead of profiling users by features such as gender, race, age, or location in a way that may harm their opportunity for a fair purchase. Nonetheless, given the magnitude of harm that this might cause, it is imperative to continually assess and ensure that users are not profiled based on these sensitive features. CAs that engage with users in a higher-risk scenario such as mental health services as opposed to clothing or household-item purchases, have a greater social responsibility towards their users and how the service may affect them. In any scenario, the user should be able to trust the system not to take advantage of them and to provide the stated service in good faith. This requires (1) explicitly detailing the agent's motivations and explaining its behaviour in a way the target user group can understand (2) evaluation to determine how the agent is treating various types of users, and (3) an understanding of users concerns, expectations, and experience.

4.3 Privacy

The interaction of humans and CAs, and sometimes even the presence of virtual agents such as in-home, always-on devices, present various ethical and legal questions including what data is collected, who has access to it, how long the data is stored and where and what such data is used for. Collecting user data raises many privacy concerns, some of which have legal basis and are covered by data protection laws that vary geographically, such as GDPR in Europe. The nature of these ethical issues varies significantly depending on the domain in which the agent is deployed and the level of vulnerability of the user group. However, we propose that clear legal requirements should be viewed as a baseline, not a target, in this area where the default approach should be to only collect and store user data if required for delivery of the stated service and to do so in a transparent manner. User privacy is paramount and is becoming increasingly important as we see AI systems rolled out into more areas of society where such systems are used to make increasingly substantial and far-reaching decisions. This makes the concept of privacy something that should not be framed entirely as a problem regarding the individual user but rather as a wider social concern. The individual user is often not afforded the opportunity or does not have the resources to negotiate terms and conditions that are written by corporations in a manner that applies to all. How we think about and legislate privacy, therefore, should be considered in light of how the collective might be impacted by the introduction of AI systems. This perspective is helpful in re-conceptualizing privacy in a way that links it to the bigger picture of collective aspirations and concerns.

A distinct concern with respect to CAs in this area is the influence that the social relations that users develop with an agent and the way user-agent interaction is often perceived as anonymous [12] [14], can encourage self-disclosure

of information. Additionally, the dialogue design of an agent impacts users inclination to self-disclose. Self-disclosure may be encouraged to gather data with the goal of improving user experience via personalization [40]. However, unlike explicitly submitting data via a structured form, users may not be conscious of how much information they have divulged via a conversation or what personal data can be inferred from their natural language utterances. Furthermore, users may not know how the system works on a technical level with regards to the processing and storing of their data [31]. For these reasons, the unique context of CAs with respect to privacy should be considered when aiming to comply with legal requirements such as GDPR or any adopted privacy guidelines.

4.4 Agent Persona

A large part of agent design decisions relate to persona and personality, which can be used to inform specific dialogue choices. Agent persona expressions include gender, age, race, cultural affiliation, and class. These indications may be more explicit as the level of embodiment increases. It is important to consider the impact of agent persona on the types of relationships users may try to explore with the agent and to determine if the design of the agent persona and accompanying dialogue is encouraging behaviour that may be harmful. Agent persona design can also inadvertently reinforce harmful stereotypes. Many publicly available agents present as female, including popular assistants such as Siri, Alexa, Cortana and the default female voice for Google Assistant [50]. While female personas are often used in subservient contexts, male personas are often found in situations perceived as authoritative, such as an automatic interviewer [23] [45] [55]. Gendering CAs in this manner may reflect market research but in the interests of gender equity, practices that embed and perpetuate socially held harmful gender stereotypes should be avoided. In some domains, there is an increased move towards androgyny such as banking agent Kai. Research has been conducted on how users respond to androgynous agents and the effects this has on user experience. A study that analysed college students' perceptions of gendered vs androgynous agents [15], found a gender-neutral agent led to more positive views on females than a female-presenting agent did. Another similar study by [42] found that female agents received more abuse than androgynous agents. There is no clear consensus within the industry on this issue. Some recommend allowing users to lead the agent persona by designing the agent to dynamically respond to how the user interacts. Others continue to gender the agents they build in an attempt to humanize the system and increase user satisfaction at the risk of reinforcing harmful gender bias. We recommend designing agents to be androgynous to avoid gender stereotypes and allow users to interpret according to their own context.

4.5 Anthropomorphism and Sexualization

Humans tend to anthropomorphize machines [26]. This kind of anthropomorphism is exacerbated when users can interact conversationally with a system

and especially if the system has been imbued with personality and embodied with an avatar or in some other way. This can be seen throughout history and occurs even when the developers themselves oppose such anthropomorphism and over-hyping of machines. The creator of ELIZA (1964-66) Joseph Weizenbaum, for example, explicitly insisted that ELIZA could not converse with true understanding. Despite this, many users were convinced of ELIZAs intelligence and empathy [49]. Possibly a surprising element of human-computer interaction is unsolicited romantic attention towards the agent. A good example of this is the popular entertainment chatbot Mitsuku¹ which has won the Loebner Prize four times. Steve Worswick, the creator and maintainer of Mitsuku, has described the type of romantic attention "she" gets and even the correspondence he receives from users demanding her freedom [53].

Research has shown users use greater profanity with a chatbot than with a human and are similarly more likely to harass a chatbot than a human agent [16], even more so if the agent has a female persona [42]. Recent work [9] that explored the capabilities of conversational agents to respond to sexual harassment on the part of the user and collected 360,000 conversations found that 4% were sexually explicit, a percentage somewhat below previous research into sexually explicit chatbot interactions. The authors argue handling these types of conversations should be a core part of a systems design and evaluation due to their prevalence and consequences of reinforcing gender bias and encouraging aggressive behaviour.

Due to the prevalence of abusive messages directed at conversational agents, unsupervised learning techniques on an unconstrained user group should be avoided. Even with a trusted user group, oversight is required to ensure the agent has not acquired harmful concepts or language. There are numerous examples of chatbots that have been released for use by the general public that use unsupervised learning but quickly learn racist, homophobic, and sexist language and have to be shut down to avoid abuse of human users. In the case of Microsoft's Tay bot, this took less than 24 hours [48]. Dialogue design should involve response strategies for romantic attention, sexualized messages, and abuse with the aim of protecting the user. If an agent can detect abusive language, which is a difficult task for both social and technical reasons, it can invoke the appropriate response strategy. This may be a non-response, a neutral response, an in-kind response, or escalation to a human agent. In this scenario the domain and goals of the agent are important, but the user demographic is the most influential factor when designing the agent's response strategy [10]. For example, it is very rare that an in-kind response, that is responding with similar tone and content as the abusive message, will be an ethical and acceptable response strategy. In the case of an education bot that converses with minors, escalation to a human (maybe a teacher) is the most appropriate response. It should be noted that a neutral response can be seen as endorsement. Engaging in use-case centred discourse can help to elicit social values that may then be used to inform

¹ Mitsuku: <https://www.pandorabots.com/mitsuku/>

the design of a specific agent’s response strategy, especially where variation of values across user groups is high (value pluralism) [33].

5 Conclusion

Assuming agents continue to improve in their functionality and conversational ability, how will their ubiquity and integration in our daily lives change how we live? Who will be most affected by the decisions of agent owners? These questions are difficult to answer but provide perspective on the ethical issues raised in this paper. Ultimately, there are no one-approach-fits-all answers to the concerns we have discussed. However, designing, building, and deploying an agent into the social sphere engenders a level of social responsibility that must be confronted and contemplated on an agent-by-agent basis to produce agent-specific strategies to address the ethical considerations described in this paper.

Acknowledgement. This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094.

References

1. Ajunwa, I., Friedler, S., Scheidegger, C.E., Venkatasubramanian, S.: Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN (2016)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica (2016)
3. Baum, S.D.: Social choice ethics in artificial intelligence. *AI & SOCIETY* pp. 1–12 (2017)
4. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion. *First Monday* 21(11-7) (2016)
5. Boiteux, M.: Messenger a F8 2018 (2018)
6. Bourdieu, P.: Language and symbolic power. Harvard University Press (1991)
7. Cheney-Lippold, J.: We are data: Algorithms and the making of our digital selves. NYU Press (2018)
8. Constine, J., Perez, S.: Facebook messenger now allows payments in its 30,000 chat bots. *techcrunch*. URL: <https://tcrn.ch/2cDEVbk> (2016)
9. Curry, A.C., Rieser, V.: #MeToo Alexa: How conversational systems respond to sexual harassment. In: *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. pp. 7–14 (2018)
10. Curry, A.C., Rieser, V.: A crowd-based evaluation of abuse response strategies in conversational agents. *arXiv preprint arXiv:1909.04387* (2019)
11. Dale, R.: The return of the chatbots. *Natural Language Engineering* 22(5), 811–817 (2016)
12. Evans, R.E., Kortum, P.: The impact of voice characteristics on user response in an interactive voice response system. *Interacting with Computers* 22(6), 606–614 (2010)
13. Ferryman, K., Pitcan, M.: Fairness in precision medicine. *Data & Society* (2018)
14. Gentsch, P.: Conversational ai: How (chat) bots will reshape the digital experience. In: *AI in Marketing, Sales and Service*, pp. 81–125. Springer (2019)

15. Gulz, A., Haake, M.: Challenging gender stereotypes using virtual pedagogical characters. In: *Gender Issues in Learning and Working with Information Technology: Social Constructs and Cultural Contexts*, pp. 113–132. IGI Global (2010)
16. Hill, J., Ford, W.R., Farreras, I.G.: Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49, 245–250 (2015)
17. Howard, R., Kehoe, J.: *Mobile consumer survey 2018: The irish cut* (2018)
18. Hutchby, I., Wooffitt, R.: *Conversation analysis*. Polity (2008)
19. Inkster, B., Sarda, S., Subramanian, V.: An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6(11), e12106 (2018)
20. Insights, G.M.: *Intelligent virtual assistant (iva) market trends share forecast 2024s* (2018)
21. Introna, L., Nissenbaum, H.: The politics of search engines. *IEEE Spectrum* 37(6), 26–27 (2000)
22. Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines. *Nature Machine Intelligence* pp. 1–11 (2019)
23. Kim, Y., Baylor, A.L., Shen, E.: Pedagogical agents as learning companions: the impact of agent emotion and gender. *Journal of Computer Assisted Learning* 23(3), 220–234 (2007)
24. Kim, Y., Wang, Y., Oh, J.: Digital media use and social engagement: How social media and smartphone use influence social activities of college students. *Cyberpsychology, Behavior, and Social Networking* 19(4), 264–269 (2016)
25. Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., Group, N.Y.P.A.: Can your phone be your therapist? young peoples ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights* 11 (2019)
26. Kuipers, B., McCarthy, J., Weizenbaum, J.: Computer power and human reason. *ACM SIGART Bulletin* (58), 4–13 (1976)
27. Lambrecht, A., Tucker, C.: Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* (2019)
28. Latham, A., Goltz, S.: A survey of the general publics views on the ethics of using ai in education. In: *International Conference on Artificial Intelligence in Education*. pp. 194–206. Springer (2019)
29. Lau, J., Zimmerman, B., Schaub, F.: Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *HCI* 2, 102 (2018)
30. Linell, P.: *Rethinking language, mind, and world dialogically*. IAP (2009)
31. Luger, E., Sellen, A.: Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. pp. 5286–5297 (2016)
32. Meek, T., Barham, H., Beltaif, N., Kaadoor, A., Akhter, T.: Managing the ethical and risk implications of rapid advances in artificial intelligence: a literature review. In: *2016 Portland International Conference on Management of Engineering and Technology (PICMET)*. pp. 682–693. IEEE (2016)
33. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: *From what to how. an overview of ai ethics tools, methods and research to translate principles into practices* (2019)
34. Mou, Y., Xu, K.: The media inequality: Comparing the initial human-human and human-ai social interactions. *Computers in Human Behavior* 72, 432–440 (2017)

35. Oh, K.J., Lee, D., Ko, B., Choi, H.J.: A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In: 2017 18th IEEE International Conference on Mobile Data Management (MDM). pp. 371–375. IEEE (2017)
36. O’Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books (2016)
37. Oracle: Can virtual experiences replace reality? (2016)
38. Richardson, R., Schultz, J., Crawford, K.: Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, Forthcoming (2019)
39. Ruane, E., Faure, T., Smith, R., Bean, D., Carson-Berndsen, J., Ventresque, A.: Botest: a framework to test the quality of conversational agents using divergent input examples. In: IUI. p. 64 (2018)
40. Saffarizadeh, K., Boodraj, M., Alashoor, T.M.: Conversational assistants: investigating privacy concerns, trust, and self-disclosure. In: International Conference on Information Systems. AIS (2017)
41. Schlesinger, A., O’Hara, K.P., Taylor, A.S.: Let’s talk about race: Identity, chatbots, and ai. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. p. 315 (2018)
42. Silvervarg, A., Raukola, K., Haake, M., Gulz, A.: The effect of visual gender on abuse in conversation with ecas. In: International Conference on Intelligent Virtual Agents. pp. 153–160. Springer (2012)
43. Smith, A.: Smartphone ownership–2013 update. Pew Research Center: Washington DC 12, 2013 (2013)
44. Tatman, R.: Gender and dialect bias in youtubes automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp. 53–59 (2017)
45. Tay, B., Jung, Y., Park, T.: When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. Computers in Human Behavior 38, 75–84 (2014)
46. Te Molder, H., Potter, J.: Conversation and cognition. Cambridge University Press (2005)
47. Vaidyam, A.N., Wisniewski, H., Halamka, J.D., Kashavan, M.S., Torous, J.B.: Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry 64(7), 456–464 (2019)
48. Wakefield, J.: Microsoft chatbot is taught to swear on twitter. bbc news (2016)
49. Weizenbaum, J., et al.: Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM 9(1), 36–45 (1966)
50. West, M., Kraut, R., Ei Chew, H.: I’d blush if i could: closing gender divides in digital skills through education (2019)
51. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J., Schwartz, O.: AI now report 2018. AI Now Institute at New York University (2018)
52. Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. arXiv preprint arXiv:1902.11097 (2019)
53. Worswick, S.: Ethics and chatbots. medium (2018)
54. Yu, H., Miao, C., Leung, C., White, T.J.: Towards ai-powered personalization in mooc learning. npj Science of Learning 2(1), 15 (2017)
55. Zhou, M.X., Mark, G., Li, J., Yang, H.: Trusting virtual agents: The effect of personality. Transactions on Interactive Intelligent Systems 9(2-3), 10 (2019)