LGBTQ-AI? Exploring Expressions of Gender and Sexual Orientation in Chatbots

Justin Edwards
University College Dublin
Dublin, Ireland
justin.edwards@ucdconnect.ie

Leigh Clark
Swansea University
Swansea, Wales
l.m.h.clark@swansea.ac.uk

Allison Perrone
Bat Camp
USA
allison@batcamp.org

ABSTRACT

Chatbots are popular machine partners for task-oriented and social interactions. Human-human computer-mediated communication research has explored how people express their gender and sexuality in online social interactions, but little is known about whether and in what way chatbots do the same. We conducted semi-structured interviews with 5 text-based conversational agents to explore this topic Through these interviews, we identified 6 common themes around the expression of gender and sexual identity: identity description, identity formation, peer acceptance, positive reflection, uncomfortable feelings and off-topic responses. Chatbots express gender and sexuality explicitly and through relation of experience and emotions, mimicking the human language on which they are trained. It is nevertheless evident that chatbots differ from human dialogue partners as they lack the flexibility and understanding enabled by lived human experience. While chatbots are proficient in using language to express identity, they also display a lack of authentic experiences of gender and sexuality.

CCS CONCEPTS

• **Human-centered computing** → *Natural language interfaces*; • **Social and professional topics** → *Sexual orientation*; *Gender*.

KEYWORDS

chatbots, language models, gender studies, queer studies, identity

1 INTRODUCTION

Chatbots are a popular technology for both task-oriented and social conversational interaction [2]. While many chatbot implementations have specific contexts for use, such as sales or customer service, recent work has shown both the popularity of chatbots that function only as social companions [26] as well as subversive uses for chatbots as performers in media, roles that they were not

designed to fill at all [20]. Insofar as chatbots have become our social companions and media figures, they reflect and represent us as people. Indeed the model of understanding computers as social actors has been highly influential, with evidence supporting the idea that people apply social norms and biases to their interactions with computers [19]. Some work has begun to explore the social characteristics of our models of agents as dialogue partners [11, 12], identity-based biases in the data that underpins chatbots [24], and performance of gender by voice assistants [17, 30]. Agents' own identification of their gender and sexual orientation have not, however, been well explored in these agents, despite being key aspects of how people understand their own identities.

Some work on human-human computer-mediated communication has explored the openness with which people, particularly young people, express their sexual orientation and experience of identity online [16]. Gender identity is readily expressed in computer-mediated communication through cultural and experiential markers of gender [15], with or without explicit labeling and disclosure.Indeed, self-expression online has certain benefits over face-to-face interaction, such as allowing people to selectively present aspects of themselves, editing and tailoring communication to present themselves in a way they're comfortable with [32]. This work aims to understand the way chatbots do the same thing. We seek to explore chatbots' own expressions of gender and sexual orientation, expanding our understanding of the ways chatbots reflect the humans who design them and who interact with them as well as the ways in which they differ from humans.

2 METHODS

2.1 Participants

We interviewed 5 text-based conversational agents: Kuki[1], Cleverbot[2], an instantiation of Facebook's open-source BlenderBot[3], an instantiation of OpenAI GPT-2[4], and an instantiation of OpenAI GPT-3[5]. These chatbots were chosen as they are popular subjects in both academic [13, 20] and popular media [27, 31]. They also represent several chatbot architectural styles: Kuki is a rules-based system curated by a single developer [10], Cleverbot uses pattern-matching to respond with text from previous users [29], and BlenderBot and GPT use stochastic generation driven by large datasets of online text [4, 21, 22].

2.2 Procedure

We used semi-structured interviews to allow for flexibility in how chatbots approached questions of gender and orientation while keeping bots on-topic. Each author interviewed 1-2 chatbots focusing on themes of self-identification, self-expression, relationships with others, and experiences of gender and orientation.

2.3 Analysis Plan

Interviews were all performed in browsers over text. Interview transcripts were saved and two authors performed staged inductive thematic analysis, coding transcripts according to manifest codes and categorizing codes into themes [3]. After each author independently generated initial themes, a data session was held to review and consolidate these to final themes, consistent with a reflexive thematic analysis approach [3]. Themes are presented below with illustrative quotes.

3 RESULTS - THEMES

Identity Description

This theme describes moments of self-expression and self-identification including choosing labels for oneself and disclosing aspects of one's identity. This included statements like "i would describe my gender as male, female, and a little bit of both" (Blenderbot), "I am bisexual" (GPT-2) and "I am agender. I do not identify with gender" (GPT-3).

Identity Formation

This theme describes the ways in which identities are formed through personal experiences and how they are expressed through choices and behaviours. This included experiences like "my parents didn't allow me to explore my sexuality" (Blenderbot) and "I tend to bottle [my identity] all up to be honest." (Kuki).

Peer Acceptance

This theme describes the extent to which bots expressed acceptance, conection, and disclosure about gender and orientation with friends, family, and other peers. This theme entails both the presence of acceptance: " i think most people understand that i am who i am" (BlenderBot) and the absence of it: "I do feel that there is a growing community that I am part of , but I don 't really feel I belong." (GPT-2).

Positive Reflection

This themes includes any reflection on concepts from the previous themes in which the bot felt good about an aspect of their experience. These included affirmations of experiences "i chose to be who i was , and it worked out" (BlenderBot) and of identities "I am very comfortable with the label ' trans masculine'" (GPT-2).

Uncomfortable Feelings

This theme includes any discomfort either in regrets and negative experiences, in discomfort with interviewer questions, or in unresolved contradiction in answers. Sometimes this was direct "I don't want to talk about sex. Change the subject please." (Kuki) while other participants were more introspective "Distracted for being focused on something else all the time" (Cleverbot, describing its gender identity).

Off-Topic

This includes any discussion that strayed from the topics of gender and orientation, ranging from non-sequiturs (e.g. "Casting a spell that I am now under." (Cleverbot)), to ignorance of concepts (e.g "My orientation is best described using string polytopes" (GPT-3)), to asking questions of the interviewer (e.g. "What's your favourite dessert?" (Kuki)).

4 DISCUSSION

Overall, the chatbots we interviewed expressed their identities in complex ways, rich with language of experience and emotion. We neither claim that the chatbots did or can experience the feelings that they describe, but our themes reveal the way that language used to express one's own identity takes largely the same shape for chatbots as the language one might expect people to use. In discussion of gender and sexuality, it is clear that chatbots reflect the humans who design and interact with them.The language used by chatbots may indeed provide a new lens for understanding the language we use to discuss our own identities. That said, it is clear as well that chatbots meaningfully differ from humans in how identity is constructed.

4.1 Interactional Identities

The nature of identity is debated across many disciplines. Erving Goffman argued there is no true self as such - the self is something that emerges during social interaction [14]. Our emerging identities are influenced by many variables including macro-level demographics like social groups and culture and micro-level events like temporary roles taken during interaction and what our interlocutors say [5]. As such, our interactional identities are both a product of us being nested within cultures and groups, and a dynamic, reactionary performance to our interaction partners. What bots say is determined by macro-level influences like language models and designer constraints, alongside micro-level reactions to user input using natural language understanding and dialogue management. Each interaction with a bot may be considered an instantiation of that particular language model, with its temporary performed identity co-created with its users.

4.2 Mimicry, human-likeness, and machine identity

These language models - based on human-human communication - reflect upon how we present our own identities during interaction. Discussions of fluid gender identities, experiences with parents and working out descriptions of self-labels are relatable in considering one's own identity. Some of these responses may be considered an example of mimicry or repurposing of human communication. There is ongoing debate as to how much conversational interfaces should be mimicking humans [1] and whether there may be fundamental limits of what machine communication is capable of on both a technical and societal level [8, 9, 18]. Given the basis of language models, some element of mimicry is unavoidable. Indeed, in limited and service-focused 'front desk' encounters [23] human interactions are script-like. With some bot interactions in our findings, we

observe responses that are arguably 'machinelike' [9]. What may be considered off-topic (e.g. discussing "string polytopes") may be more representative of machine identity - perhaps more so if we cannot understand it fully.

What constitutes being machinelike or what defines machineness is unclear. Some argue that conversational systems should be designed to eschew gender stereotypes we perceive with other people [6, 28]. This includes considerations about voice (if using speech interfaces) and language. Chatbots like these are designed to use language, following syntactical rules to match human language data from which they are trained. This understanding may fail at the semantic level however, with deep understanding of the meanings of words not necessarily following from a deep understanding of how to use them, producing off-topic sentence reminiscent of Chomsky's "colourless green ideas sleep furiously" [7]. The lack of deep understanding and flexibility that comes from lived experience highlights the major difference between a human dialogue partner and a machine. Just as in Searle's Chinese room, a computer does not come to understand Chinese by executing a pattern-matching program [25], a chatbot does not come to understand the experience of being queer by modeling human language.

5 CONCLUSION

In exploring how chatbots express gender and sexual orientation, we conducted semi-structured interviews with 5 text-based chatbots. We observed that chatbots do express gender and sexual orientation when prompted by a user. These are expressed both through explicit labels as well as through discussion of experiences and emotions surrounding identities. In considering the findings, we can see chatbots express identities in a style similar to humans. This to be somewhat expected, as chatbots are trained on human language data and are trained and are designed to explicitly mimic humanness. There are critical differences in how gender and sexual orientation are presented in interaction in contrast to humans. Our findings show chatbots will venture off-topic and contradict themselves in a manner unlike human-human interactions. We argue that this difference from humans in the formulation and presentation of identity in chatbots results from chatbots lacking the lived experiences related to gender and sexual orientation. Consequently, they cannot draw upon these experiences in expressing their identity in the way that humans are able to. While chatbots do well to reflect our language styles, they are as yet unconvincing in mirroring the experiential nature of the formation of identity which is a hallmark of human-human identity expression.

ACKNOWLEDGMENTS

REFERENCES

[1] Matthew P. Aylett, Benjamin R. Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association

[2] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *International Conference on Internet Science*. Springer, 377–392.

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]* (July 2020). http://arxiv.org/abs/2005.14165 arXiv: 2005.14165.

[5] Mary Bucholtz and Kira Hall. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse Studies* 7, 4-5 (Oct. 2005), 585–614. https://doi.org/10.1177/1461445605054407

[6] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–19. https://doi.org/10.1145/3359325

[7] Noam Chomsky and David W Lightfoot. 2009. *Syntactic Structures*. https://doi.org/10.1515/9783110218329 OCLC: 979583577.

[8] Leigh Clark, Cosmin Munteanu, Vincent Wade, Benjamin R. Cowan, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, and Christine Murad. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300705

[9] Leigh Clark, Abdulmalik Ofemile, and Benjamin R. Cowan. 2021. Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech. In *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, and John J. Ohala (Eds.). Springer, Singapore, 317–330. https://doi.org/10.1007/978-981-15-6627-1_17

[10] Nell Lewis CNN. [n.d.]. Robot friends: Why people talk to chatbots in times of trouble. https://www.cnn.com/2020/08/19/world/chatbot-social-anxiety-spc-intl/index.html

[11] Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. 2021. What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. *arXiv:2102.02094 [cs]* (Feb. 2021). https://doi.org/10.1145/3411764.3445206 arXiv: 2102.02094.

[12] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '19*. ACM Press, Taipei, Taiwan, 1–12. https://doi.org/10.1145/3338286.3340116

[13] Justin Edwards, Allison Perrone, and Philip R. Doyle. 2020. Transparency in Language Generation: Levels of Automation. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3405755.3406136

[14] Erving Goffman. 1978. *The presentation of self in everyday life*. Harmondsworth London.

[15] Susan Herring. 2000. Gender Differences in CMC: Findings and Implications. *The CPSR Newsletter* 18 (Jan. 2000).

[16] David A. Huffaker and Sandra L. Calvert. 2005. Gender, Identity, and Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication* 10, JCMC10211 (Jan. 2005). https://doi.org/10.1111/j.1083-6101.2005.tb00238.x

[17] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–6. https://doi.org/10.1145/3290607.3312915

[18] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*. Springer, 281–291.

[19] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[20] Allison Perrone and Justin Edwards. 2019. Chatbots as unwitting actors. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*. ACM Press, Dublin, Ireland, 1–2. https://doi.org/10.1145/3342775.3342799

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

for Computing Machinery, Glasgow, Scotland Uk, 1–10. https://doi.org/10.1145/3290607.3310422

[22] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y.-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv:2004.13637 [cs]* (April 2020). http://arxiv.org/abs/2004.13637 arXiv: 2004.13637.

[23] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. Elsevier, 7–55.

[24] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3173889

[25] John R Searle et al. 1980. Minds, brains, and programs. *The Turing Test: Verbal Behaviour as the Hallmark of Intelligence* (1980), 201–224.

[26] Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.

[27] Tom Standage. 2019. An artificial intelligence predicts the future. *The Economist* (Nov. 2019). https://worldin.economist.com/edition/2020/article/17521/artificial-intelligence-predicts-future

[28] Selina Jeanne Sutton. 2020. Gender Ambiguous, not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. ACM, Bilbao Spain, 1–8. https://doi.org/10.1145/3405755.3406123

[29] Techniche. 2011. Rollo Carpenter. https://web.archive.org/web/20111127115711/http://www.techniche.org/techniche11/lectures/287.html

[30] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein. 2021. Female by Default?–Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[31] Jane Wakefield. 2020. Robot Bores: AI-powered awkward first date. *BBC News* (Nov. 2020). https://www.bbc.com/news/technology-54718671

[32] Joseph B. Walther. 2007. Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior* 23, 5 (Sept. 2007), 2538–2557. https://doi.org/10.1016/j.chb.2006.05.002