Toward LLM-Powered Social Robots for Supporting Sensitive Disclosures of Stigmatized Health Conditions

Alemitu Bezabih
Colorado School of Mines,
Department of Computer Science
Golden, CO, USA
alemitubezabih@mines.edu

Shadi Nourriz
Colorado School of Mines,
Department of Computer Science
Golden, CO, USA
shadinourriz@mines.edu

C. Estelle Smith
Colorado School of Mines,
Department of Computer Science
Golden, Colorado, USA
estellesmith@mines.edu

ABSTRACT

Disclosing sensitive health conditions offers significant benefits at both individual and societal levels. However, patients often face challenges due to concerns about stigma. The use of social robots and chatbots to support sensitive disclosures is gaining traction, especially with the emergence of LLM models. Yet, numerous technical, ethical, privacy, safety, efficacy, and reporting concerns must be carefully addressed in this context. In this position paper, we focus on the example of HIV status disclosure, examining key opportunities, technical considerations, and risks associated with LLM-backed social robotics.

CCS CONCEPTS

• **Computer systems organization** → **Robotics**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Sensitive Disclosure, HIV, Disclosure Model, Social Robots, Large Language Models, LLMs, Ethics, Safety, Privacy

1 INTRODUCTION

For patients struggling with stigmatized physical and mental health conditions, disclosing sensitive information about their health–whether in online contexts such as social media [1, 21] or in-person social interactions [4] –poses numerous challenges, including: lack of *a priori* knowledge and skill for how, when, and whom to disclose to; navigating complex interpersonal dynamics prior to, during, and after disclosure; fear of discrimination or social ostracization; and practical concerns like access to healthcare, employment security, and financial stability. However, successful sensitive disclosure can also benefit both individuals and society by fostering empowerment, reducing anxiety, and improving health and quality of life outcomes. Transparency also cultivates trust, facilitates informed decision-making for relationships and healthcare, and fosters a more inclusive community by reducing stigma and promoting empathy.

The use of social robots in HRI (and chatbots in HCI) to support sensitive disclosures is rapidly gaining research traction [11, 12, 15, 17, 19, 22] —particularly considering the increasing public availability of LLM models which can readily be implemented as backend models for disclosure systems rather than, for example, traditional Wizard-of-Oz techniques to mimic human-like responses. Yet there are numerous technical, ethical, privacy, safety, efficacy, and reporting concerns that must be carefully discussed in order to ultimately realize the potential of social robots for supporting sensitive disclosures. In this position paper, we consider a case study of disclosure of human immunodeficiency virus (HIV)-positive status, a sexually transmitted infection (STI). Using a 4-step model of how clinicians now navigate HIV disclosure plans with patients, we explore the nuances of this stigmatized physical illness and outline some of its major opportunities, considerations, and risks for social robotics. We posit that future work must make similar considerations for other stigmatized physical and mental illnesses, while also adapting to their unique idiosyncrasies.

2 PRIOR WORK IN SENSITIVE DISCLOSURES

Social robots in HRI and conversational agents in HCI are gaining traction toward facilitating sensitive disclosures of challenging personal experiences and emotions. For instance, robotic and chatbot conversations are effective in creating emotionally safe spaces for self-disclosure [15], improving and encouraging self-disclosure [12, 17, 22], and diagnosing and improving mental health [11, 19]. These HRI/HCI efforts have greatly contributed to designing ways to encourage people to share personal and sensitive information that they might not feel comfortable disclosing to humans. However, several ethical issues remain unresolved—*e.g.*, data & privacy, informed consent & autonomy, and loss of empathy [7]. Moreover, the recent public release of LLMs (*e.g.*, ChatGPT) has greatly attracted HRI/HCI scholarship to integrate LLMs into social robots, exacerbating preexisting risks with new concerns. Therefore, scholars must proactively identify risks in domain-specific designs and contribute to evolving guidelines. Moving toward developing such guidelines, we consider the case of HIV status disclosure via social robots, drawing from an existing 4-step clinical model.
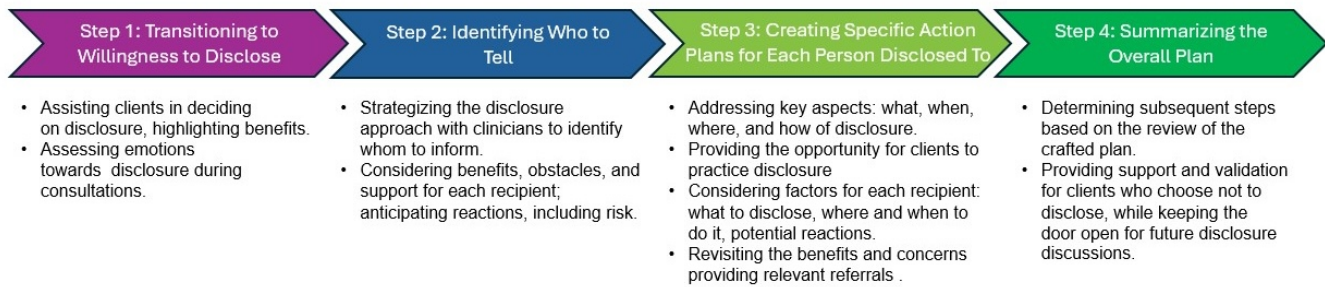
Figure 1: A four-step self-disclosure model used in clinical practice, for example by the California Prevention Training Center in the UCSF Bixby Center for Global Reproductive Health [6]

3 BACKGROUND: THE 4-STEP CLINICAL MODEL OF DISCLOSURE

In contexts like HIV/STIs, patients must disclose their HIV status to sexual partners for the sake of health of the latter. Other vital disclosure contexts include parents disclosing their own HIV status to children, or parents disclosing HIV-status to children born with HIV [4, 16]. However, many patients struggle with knowing how, when, and with whom to disclose HIV status. Therefore, clinicians engage in ongoing discussions with patients until they develop a disclosure plan for potentially affected individuals. Figure 1 shows a 4-step disclosure model that is now used in clinical practice to guide this process [6]. The steps include: (1) transitioning to willingness to disclose; (2) identifying who to tell; (3) creating specific action plans for each person disclosed to, and (4) summarizing the overall plan. However, going through these steps is challenging given the time constraints and workload of healthcare workers. Importantly, there is also usually limited or no time available during a clinical consult for *practicing* disclosure.

4 KEY OPPORTUNITIES FOR INTEGRATION OF SOCIAL ROBOTS IN HIV CLINICS

Social robots powered by LLMs could be helpful to support HIV disclosure and enhance clinical care. For example, in cooperation with clinicians, they could prepare clients for disclosure via some or all of the four steps listed above. Moreover, given the significant lack of clinician time available, a social robot could provide a new 5th step by offering training and practice opportunities: Patients could verbally and repeatedly practice disclosure with the robot, as a stand-in for the intended future human recipient. Informed by prior work in HRI such as [23], we hypothesize that such practice may be highly relatable and transferable to real-life conversations; thus brainstorming and practicing sensitive disclosures with a social robot could improve later disclosure to a human. In order to become clinically feasible given clinicians' time constraints and staffing availability, this practice time would likely not be supervised by staff. Figure 2 depicts a storyboard sketch illustrating the potential for a human patient to practice disclosure with a robot, as described in the hypothetical scenario in Section 4.1.

4.1 An Example Scenario

John, a young man in his twenties, recently found himself in a difficult situation. Having struggled with drug addiction, he made the brave decision to undergo an HIV test. The first test brought temporary relief as it turned out negative. However, a subsequent test revealed the harsh reality - John was HIV positive. Feeling overwhelmed and uncertain, John sought support from his regular clinician. After having a couple of counseling sessions, the clinician helped him to consider disclosing to his circles. To aid this process, the clinician recommended him to a clinic's tech room equipped with a social robot designed to assist individuals in sensitive situations like his.

Upon entering the tech room, John was greeted by the social robot, which introduced itself as his supportive companion throughout his disclosure journey. The robot knows everything about John's health profile from his electronic medical record (EMR). The robot began by guiding John through the process of identifying people to disclose his HIV status to, suggesting pre-existing categories such as family, friends, sex partners, co-workers, and providers. Next, the robot helped John brainstorm the benefits and concerns associated with disclosing to each person, allowing him to weigh the potential outcomes carefully. With the robot's assistance, John created a detailed action plan for each individual on his list, outlining how and when he would disclose his status to them. Through a UI/UX interface, John and the robot crafted a disclosure plan and attached it to the EMR.

To prepare John for the challenging conversations ahead, the robot provided a safe space for him to practice disclosure, allowing him to say his thoughts out loud and receive feedback and suggestions for improvement. The robot's non-judgmental presence and supportive feedback helped John build confidence in expressing himself. As the session drew to a close, the robot summarized the overall plan, revisiting specific action plans and key issues discussed during the session. Additionally, the robot reminded John of the importance of confidentiality and the potential risks of breaches. Feeling empowered and supported by the guidance of the social robot, John left the tech room with a sense of clarity and determination to navigate the journey of HIV disclosure with confidence and resilience.

5 TECHNICAL & DESIGN CONSIDERATIONS

Which steps are [in]appropriate for a social robot to manage? Of the steps outlined, it is not immediately clear which steps are best supported by a social robot. Which aspects of care are
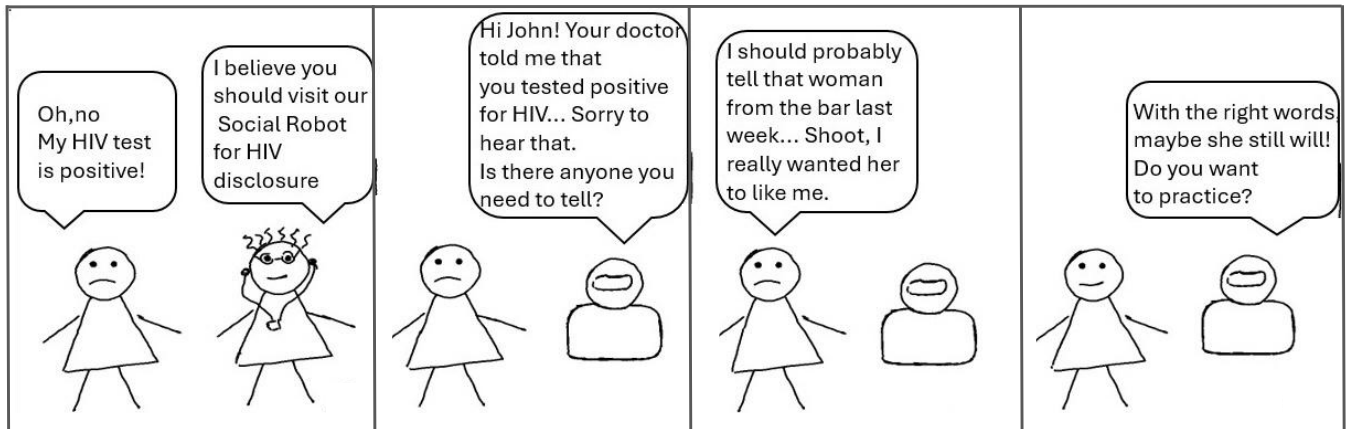
Figure 2: A storyboard depicting a scenario in which a user discovers that he is HIV-positive during a clinical visit. The clinician recommends that he visits with a social robot to practice disclosing his HIV-positive status to important social connections.

best delivered by humans, and which by robots? For instance, are clinicians or robots more persuasive in convincing a human to disclose in the first place? Are clinicians or robots more effective in suggesting targets of disclosure, plans to disclose, and summaries of the plan? Research should first ascertain the degree to which it is both clinically feasible and desired by patients for clinicians or robots to administer each of the needed steps.

How might LLMs be technically integrated? In a clinical context, information can be collected and shared via screening tools (*e.g.,* intake surveys or interviews), online messaging with clinicians through electronic portals, in-person discussions with clinicians during appointments, and/or electronic health records (EHRs). In order to effectively interact with patients, a social robot will either need to ingest previously collected data from such source(s), or collect information itself during interaction with a patient. Moreover, its conversations must be structured appropriately to meet clinical aims. Therefore a robot must be architectured with consideration for how and when use of an LLM is appropriate in the workflow. Some steps may be best served by structured, consistent queries (*e.g.,* intake questions, determining who to disclose to, or summarizing a plan) without need for an LLM. In other cases, natural language may be essential (*e.g.,* persuading to disclose, or practicing disclosures) for a fluid and effective user experience. The architecture of the robot must therefore be capable of reasoning about which step of the process it is involved in, use the correct LLM or non-LLM based strategy for that step, have appropriately trained and scoped LLM-models available for LLM-based steps, and have access to the appropriate PII for each patient. Thus, questions arise related to good training materials for each LLM-based step, as well as technical access to electronic discussion portals, EHRs, or in-person dialogues.

Reporting concerns. In steps that eventually do involve LLMs, it is essential that researchers proceed cautiously, and record, review and report on LLM performance. Clinicians should be involved in the review of interactions, and comparisons made between the language clinicians would choose to use at each step, and the language

generated by an LLM. Later technical development should be taken to adjust LLM-generated utterances toward clinician recommendations. Patient feedback is also essential. For instance, preliminary data suggest that LLMs can be perceived as more empathetic than clinicians [3]; conversely, if LLMs can take care of some steps, this may enable more time for clinicians to express more empathy [20]. Researchers should report on: how LLMs were trained at each step; the performance of LLMs according to clinicians and patients; and what steps can be taken to improve performance.

6 RISKS & ETHICS IN HIV DISCLOSURE

Safety and efficacy. Issues may arise concerning the quality of LLM-generated robot conversations, such as misinformation. An LLM-backed social robot could potentially offer non-factual information, including deliberate disinformation (*e.g.,* stemming from malicious attacks on LLMs), unintentional inaccuracies or biases inherent in the LLM's training data, or subtle misstatements that reinforce stigma. For instance, there is a risk that a robot may suggest an ineffective strategy or an inappropriate target for disclosure. Consequently, questions emerge regarding how to filter information originating from the LLM to mitigate possible harms. Furthermore, given the highly sensitive nature of HIV, individuals living with the condition often harbor traumatic and stigmatizing experiences. Engaging in a conversation with a robot may inadvertently trigger these negative experiences, leading to emotional distress. While such triggers may also occur in human interactions, clinicians are equipped to provide emotional support in such instances. Thus, the challenge lies in detecting and effectively managing these triggers when they arise during interactions with a social robot. In conclusion, there is a pressing need for mechanisms to evaluate the effectiveness of robotic interventions in comparison to standard care protocols.

Privacy & HIPAA. The concern of privacy, particularly in sensitive contexts, raises critical questions regarding the integration

of social robots with any clinical data, EHRs, or Personal Identifiable Information (PII) [9]. With the stringent regulations outlined in HIPAA [10], concerns arise about whether a robot *should* have direct access to such data. One potential approach involves implementing additional privacy-preserving layers to ensure the protection of PII–*e.g.,* stripping PII from prompts before they are processed by LLMs, and then reintroducing it at the conversational layer as needed. By incorporating such measures, we can strike a balance between leveraging the capabilities of social robots in healthcare contexts while safeguarding patient privacy following HIPAA regulations. Implementing additional privacy-preserving layers may increase computational overhead and processing time, potentially slowing down system response. Optimization strategies, such as efficient algorithms and hardware acceleration, may be necessary to mitigate these efficiency and speed implications.

Ethical and social concerns*.* The introduction of robots into sensitive domains raises significant ethical and social concerns regarding their appropriateness. For instance, clinicians adhere to strict codes of ethics (*e.g.,* The Code of Ethics for Nurses [2]), yet how to hold robots accountable remains an open question. Moreover, ethics of care emphasize the importance of relationships, empathy, and compassion in human-centered care of patients [8], however robots and LLMs are not sentient beings capable of genuine care—despite their appearance or mimicry of sentience. For instance, the following is a quote from a chaplain participant in [18]: *"God help us if it's just robots taking care of people at end of life."* In the extreme case of terminal illness, this quote evokes a philosophical dread for the loss of soulful, loving connection that is vital to human wellbeing, meaning-making, and purpose. Limits should be limits imposed on the degree to which robots can replace humans in clinical care, and these must be considered with excessive caution and forethought. Establishing frameworks for robot accountability through regulations, standards of care, and oversight mechanisms, is essential to ensure their responsible use and to mitigate potential risks to individuals and society at large.

7 PROPOSED EXPERIMENTAL DESIGN

In our future work, we aim to investigate these questions situated within an HIV clinic at our local Children's Hospital. In our first phase of work, we plan to conduct design work in collaboration with HIV clinicians. We will: (1) investigate which steps clinicians recommend be integrated with a robot and explore the feasibility and acceptability ramifications of these steps; (2) design and prototype the robot's cognitive architecture; (3) determine the technical requirements required to protect patient privacy and safety.

Building upon prior work [4], our prototype design will focus on parents disclosing their HIV status to children. Here, we need to highlight that while the four-step model described in section 3 assumes disclosure as a one-off event, this approach may not be appropriate for children. Children have a limited understanding of illness processes which advances with age [5]. This implies that HIV information needs to be structured around stages of cognitive development and should be delivered progressively following children's capacities to understand illness [14]. In consultation with HIV clinicians, we will determine a specific population of children at a particular developmental stage to work with for this study. For

the technical development of the prototype, we will utilize a child-faced Furhat robot integrated with a fine-tuned GPT-4 LLM model. To mitigate the safety and efficacy concerns presented above, we plan to train our model on a customized dataset consisting of HIV and disclosure resources including peer-reviewed HIV research, disclosure guidelines, and governmental policies. We will experiment with a variety of prompt engineering strategies to balance an empathetic, caring, and understanding tone against a prescriptive or persuasive clinical tone.

Following this careful stage of prototyping and co-design with clinicians, we propose to conduct an experiment in which we compare patients' perceptions of and benefits derived from disembodied LLM-based conversational agents (i.e. a chatbot using our LLM model) *v.s.* similar agents embodied within the physical Furhat robot. We plan to conduct a one-month trial of our prototypes with 30 parents living with HIV (15 to trial the disembodied UI/UX version, and 15 to trial the robot format), as a supplemental service to standard disclosure support. To demonstrate whether there is a meaningful advantage to using a robot over a chat interface, a pre-post disclosure readiness assessment will be conducted based on the six stages of decision readiness to disclose their HIV-positive status to children as described in [13].

8 CONCLUSION

The disclosure of sensitive health conditions holds immense benefits but is often hindered by stigma-related challenges. The rising use of social robots and chatbots, particularly with advancements in LLM models, may help to support sensitive disclosures. However, the integration of these technologies raises numerous technical, ethical, privacy, safety, efficacy, and reporting concerns. Focusing on HIV status disclosure as an example, our position paper explores opportunities, technical considerations, and risks associated with LLM-backed social robotics in healthcare. We explore which aspects of the disclosure process are suitable for social robots, technical integration of LLMs, safety, efficacy, privacy, HIPAA compliance, and broader ethical and social implications. Through this exploration, we aim to offer insights into responsible and ethical deployment of social robots in sensitive healthcare contexts.

REFERENCES

[1] Nazanin Andalibi, Margaret E. Morris, and Andrea Forte. 2018. Testing Waters, Sending Clues: Indirect Disclosures of Socially Stigmatized Experiences on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 19:1–19:23. https://doi.org/10.1145/3274288

[2] American Nurses Association. 2017. Code of Ethics for Nurses. https://www.nursingworld.org/practice-policy/nursing-excellence/ethics/code-of-ethics-for-nurses/

[3] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* 183, 6 (June 2023), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

[4] Alemitu Bezabih, Kathrin Gerling, Workeabeba Abebe, and Vero Vanden Abeele. 2023. Challenges and Opportunities for Interactive Technology to Support Parents of HIV-Positive Children in Ethiopia in the Disclosure Process. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3580756

[5] Roger Bibace and Mary E. Walsh. 1980. Development of Children's Concepts of Illness. *Pediatrics* 66, 6 (Dec. 1980), 912–917. https://pediatrics.aappublications.org/content/66/6/912 Publisher: American Academy of Pediatrics Section: Article.

[6] california_ptc. 2024. Self-Disclosure. https://californiaptc.com/training/selfdisclosure/

[7] Dariush D. Farhud and Shaghayegh Zokaei. 2021. Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iranian Journal of Public Health* 50, 11 (Nov. 2021), i–v. https://doi.org/10.18502/ijph.v50i11.7600

[8] Carol Gilligan. 1993. *In a Different Voice: Psychological Theory and Women's Development.* Harvard University Press. Google-Books-ID: XItMnL7ho2gC.

[9] David Grande, Xochitl Luna Marti, Rachel Feuerstein-Simon, Raina M. Merchant, David A. Asch, Ashley Lewson, and Carolyn C. Cannuscio. 2020. Health Policy and Privacy Challenges Associated With Digital Technology. *JAMA Network Open* 3, 7 (July 2020), e208285. https://doi.org/10.1001/jamanetworkopen.2020.8285

[10] Office for Civil HHS. 2024. The HIPAA Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html Last Modified: 2022-03-31T17:56:44-0400.

[11] Masamune Kawasaki, Naomi Yamashita, Yi-Chieh Lee, and Kayoko Nohara. 2020. Assessing Users' Mental Status from their Journaling Behavior through Chatbots. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20).* Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3383652.3423870

[12] Guy Laban, Val Morrison, Arvid Kappas, and Emily S. Cross. 2022. Informal Caregivers Disclose Increasingly More to a Social Robot Over Time. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22).* Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3491101.3519666

[13] T. C. Laschober, J. M. Serovich, M. J. Brown, J. A. Kimberly, and C. M. Lescano. 2019. Decision Readiness to Disclose HIV-Positive Status to Family Members. *Sexually transmitted diseases* 46, 5 (May 2019), 342–346. https://doi.org/10.1097/OLQ.0000000000000966

[14] A. Lesch, L. Swartz, A. Kagee, K. Moodley, Z. Kafaar, L. Myer, and M. Cotton. 2007. Paediatric HIV/AIDS disclosure: towards a developmental and process-oriented approach. *AIDS Care* 19, 6 (July 2007), 811–816. https://doi.org/10.1080/09540120601129301 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09540120601129301.

[15] Tony W. Li, Michael Murray, Zander Brumbaugh, Raida Karim, Hanna Lee, Maya Cakmak, and Elin A. Björling. 2023. Tell Me About It: Adolescent Self-Disclosure with an Online Robot for Mental Health. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23).* Association for

[16] Alemitu Mequanint Bezabih, Kathrin Gerling, Workeabeba Abebe, and Vero Vanden Abeele. 2022. The Challenge of (Non-)Disclosure: Exploring the Lived Experience of Ethiopian Adolescents with HIV and Their Attitudes Toward Technology. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502132

[17] Yohei Noguchi, Hiroko Kamide, and Fumihide Tanaka. 2020. Personality Traits for a Social Mediator Robot Encouraging Elderly Self-Disclosure on Loss Experiences. *ACM Transactions on Human-Robot Interaction* 9, 3 (May 2020), 17:1–17:24. https://doi.org/10.1145/3377342

[18] C. Estelle Smith, Avleen Kaur, Katie Z. Gach, Loren Terveen, Mary Jo Kreitzer, and Susan O'Conner-Von. 2021. What is Spiritual Support and How Might It Impact the Design of Online Communities? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 43:1–43:42. https://doi.org/10.1145/3449117

[19] Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal Analysis and Estimation of Intimate Self-Disclosure. In *2019 International Conference on Multimodal Interaction (ICMI '19).* Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3340555.3353737

[20] Eric J. Topol. 2023. Machines and empathy in medicine. *The Lancet* 402, 10411 (Oct. 2023), 1411. https://doi.org/10.1016/S0140-6736(23)02292-4 Publisher: Elsevier.

[21] Mark Warner, Andreas Gutmann, M. Angela Sasse, and Ann Blandford. 2018. Privacy Unraveling Around Explicit HIV Status Disclosure Fields in the Online Geosocial Hookup App Grindr. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018). https://doi.org/10.1145/3274450

[22] Qian Yu, Tonya Nguyen, Soravis Prakkamakul, and Niloufar Salehi. 2019. "I Almost Fell in Love with a Machine": Speaking with Computers Affects Self-disclosure. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19).* Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312918

[23] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics* 7, 3 (June 2015), 347–360. https://doi.org/10.1007/s12369-014-0267-6

Computing Machinery, New York, NY, USA, 183–187. https://doi.org/10.1145/3568294.3580068