Motivational Algorithms: Appraising Taxonomies of Trust in a Counselling Chatbot

Motivational Algorithms

Appraising Taxonomies of Trust in a Counselling Chatbot

Mathew Iantorno

Faculty of Information, University of Toronto, mathew.iantorno@mail.utoronto.ca

Olivia Doggett

Faculty of Information, University of Toronto, olivia.doggett@mail.utoronto.ca

Camille Intson

Faculty of Information, University of Toronto, camille.intson@mail.utoronto.ca

Matt Ratto

Faculty of Information, University of Toronto, matt.ratto@utoronto.ca

This paper examines the development of an artificial intelligence (AI) driven chatbot built in a partnership between the Centre for Addiction and Mental Health (CAMH) and the University of Toronto to aid users with smoking cessation. The bot is programmed using motivational interviewing (MI) treatment techniques to substitute for a human clinician. This paper examines how the transposition of a behavioral intervention chatbot into this clinical role creates unique challenges in appraising trust in the relationship. Through evaluating two taxonomies of trust and a pilot study of the smoking cessation chatbot itself, new concerns for gauging trust when embodied human ability is translated into a disembodied digital space will be proposed. Four attributes were considered as particularly important to evaluating new trust challenges for behavioral intervention chatbots: (1) predictability, (2) expert knowledge, (3) clinical context, and (4) human experience.

CCS CONCEPTS • Applied computing—Life and medical sciences—consumer health • Computing methodologies—Artificial intelligence • Human-centered computing • Human computer interaction (HCI).

Additional Keywords and Phrases: AI trust, behavioral counselling, motivational interviewing

ACM Reference Format:
First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

As the data sets driving conversational user interfaces (CUIs) have advanced, there has been growing interest in utilizing these technologies beyond the transactional exchanges of virtual assistants. Researchers for the Centre for Addiction and Mental Health (CAMH) in collaboration with the authors and University of Toronto computer science faculty and students are currently developing a chatbot aimed to assist users with smoking cessation. This chatbot leverages the robust Generative Pre-trained Transformer 3 (GPT-3) autoregressive language model and other natural language processing (NLP) technologies such as GPT-2 and BERT. By adapting the counselling approach of motivational interviewing (MI), the artificial intelligence (AI) driven chatbot is intended to provide an accessible alternative to an appointment with a trained therapist. This paper examines how the transposition of a chatbot into this clinical role creates unique challenges in appraising trust within the relationship. Current taxonomies of trust are often designed exclusively for human-human [12] or human-machine interaction [19], leaving little affordance for technologies intended to directly emulate empathetic domains of experience. Through evaluating these taxonomies and a pilot study of the smoking cessation chatbot itself, this paper will propose new concerns for gauging trust when embodied human ability is translated into a disembodied digital space.

1.1 Motivational Interviewing

The CAMH chatbot project poses that most habitual smokers have contradictory feelings about the act of smoking, yet their motivation to quit is limited [6]. Smokers are a perennially hard-to-reach target group that are notoriously unwilling to seek counselling for their addictions [6]. As described in previously published work on this project [1][2], practitioners that can perform these clinical interventions are scarce and financially inaccessible for many. As such, the primary goal of the smoking cessation bot is to reach individuals bereft of the opportunity to engage in-person with counsellors.

The chatbot is programmed using MI techniques [1][2]. MI is a client-centred style of psychotherapy that focuses on encouraging clients to voice their own motivations for behavioural change [14]. Developed by William Miller in the late 80s, MI privilege client rather than counsellor-guided treatment sessions [14]. In the context of smoking cessation, a MI session would involve the therapist asking the patient open-ended questions regarding their health, social relationships, and personal history to determine their motivation for quitting. The ideology of MI revolves around four guiding elements (partnership, acceptance, compassion, and evocation) and four major skills (open-ended questions, affirmations, reflections, and summaries) [13][14]. The current version of the chatbot uses a highly distilled version of the MI method, asking participants to list and then elaborate on positive and negative rationales for smoking using a text-based interface. Future work by the team seeks to enhance the capacity of the chatbot to provide more accurate reflections and affirmations [1][2].

Miller and Rollnick advocate that MI is not a “technique” predicated on a “relatively simple operation”; rather, it is a complex “clinical or communication method” that requires “considerable practice over time” [14], pp.131]. MI, consequently, is neither a profession nor a procedure, but a constantly evolving practice. As summarized by addiction psychology scholar, Steve Allsop, MI cannot simply be “reduced to a bag of techniques or tricks” [14], pp.343]. Thus despite the question-and-answer formularity of MI adapting well to traditional computational logics, the need for a honed experiential skillset puts into question how a computer can competently perform the role of an MI counsellor. Furthermore, the adoption of such an empathetic, human-centric role causes the smoking cessation chatbot to occupy a liminal space in regard to expectations and appraisals of trust. In the

following sections, this paper will explore how current taxonomies of trust specific to humans and machines fail to account for the advent of chatbot-as-therapist.

2 TAXONOMIES OF TRUST

The development of taxonomies to evaluate trust in human relationships has been a longstanding ambition within the social sciences [3][4][9][11]. As human beings have entered close working relationships with autonomous systems in the late twentieth century, a parallel discourse has developed in the field of computer science [7][16][18][22]. Although both such taxonomies share a mandate of discerning how effective relationships are forged between two actors, a clear delineation is generally made between human-human relationships (grounded in benevolence) and human-computer interactions (grounded in reliability). This delineation consequently complicates the appraisal of trust in machines intended to emulate human empathy and experiential spheres of knowledge.

To illustrate this complication, we have developed a table contrasting two representative trust taxonomies (see Appendix, Table 1). The first taxonomy, developed by McKnight et al. [12], draws from the disciplines of psychology and sociology to establish three categories of trust within human interpersonal relationships: dispositional trust, interpersonal trust, and institutional trust. The second taxonomy, developed by Schaefer et al. [19], similarly relies on a tripartite model of trust for developing synergistic human-machine interaction: human-related factors, partner-related factors, and environment-related factors. These two taxonomies were chosen for their inclusion of extensive literature reviews of trust in their respective fields. Using the table, we engage in a comparative analysis between these two models to discern not only shared attributes between human-human and human-computer taxonomies of trust but also conflicting attributes.

3 UNIQUE TRUST BOT ATTRIBUTES

Using the conflicting attributes established in Table 1, we have established four specific unique trust considerations for the smoking cessation chatbot: predictability, expert knowledge, clinical context, and human experience. Throughout the next section, we will examine how these considerations problematize the application of common models of trust to appraise an artificially intelligent counsellor, both within the context of the CAMH project and towards autonomous systems intended to emulate common domains of human empathy and experience.

3.1 Predictability

Concerns of unpredictability closely align between interpersonal trust and partner-related factors on the table. MI itself is an intervention that relies on gradual, non-disruptive change [14], framing unpredictable questions that cause a patient to “deny and minimize” a problem as methodologically unsound [14]. The positioning of the chatbot as a motivational interviewer consequently complicates its adherence to either a purely human- or machine-focused appraisal of unpredictability. Pragmatically, unpredictability can always be attributed to shortcomings within the algorithmic logic of GPT-3; however, many subtle technical errors (such as overbearing questions or miscategorization of problems) more closely resemble a lapse of expert knowledge. In these instances, unpredictability may not be attributed to malfunction but rather the general efficacy of MI as a treatment method. The fluidity in which moments of unpredictability are categorized by the user, often on an error-by-error basis, complicates the chatbot’s adherence to any single taxonomy of trust.

3.2 Expert Knowledge

As the CAMH smoking cessation chatbot is designed to stand in for a trained MI therapist, it becomes the gatekeeper of expert knowledge previously held by a specialist and/or institution. The patient's propensity to accept the chatbot's expert knowledge comes not only from MI as a clinical practice, but from the technical capabilities and design affordances to allow the chatbot to ethically and competently carry out its therapeutic intervention. A hybrid taxonomy of trust that bridges interpersonal and partner-related categories (see Appendix - Table 1) is therefore necessary in our analysis of the project. The chatbot is tasked with simultaneously demonstrating expert knowledge, benevolence, reliability, and ethical responsibility in its role as a MI therapist, and this myriad of human- and machine-related factors makes it increasingly conducive to perceived failure or error if one of these qualities falters. Complications also arise when the procedural logic of the chatbot comes into conflict with the anti-procedural logic of MI [13][14]

3.3 Clinical Context

As shown in Table 1, the bot aligns with environment-related trust factors as it is programmed to create a trusting environment for clients through its ability to hold a pleasurable (and operationally consistent) conversation [1][2]. However, without the physical context of a clinic or the expert authority of a human, the clients may not develop the same institutional trust they would in an embodied, face-to-face interaction. Institutions provide authority and structural assurance to back up a person's individual qualities; the bot's authority, disembodied from physical space and professional designation, is not 'backed up' in the same way human facilitators may be [12][19]. However, this does not mean that the bot is devoid of institutional entanglements. It is a digital product produced by CAMH, GPT-3 developers, and the Internet; it is as much beholden to the moral imperatives of MI as it is to the institutions that have invested in programming and developing it.

3.4 Human Experience

The smoking cessation chatbot has a non-human ontological status, lacking the personal motivations that colour typical client-therapist relationships. Because GPT-3 and AI technologies as whole currently lack the capacity for human empathy and intuition, there is a marked difference of embodied experience between conversing with a human therapist and a well-trained machine. This absence of human motivation proves an oft-considered hurdle in the chatbot's capacity to perform clinical interventions using MI [20]. This complication places the chatbot in a liminal space between the interpersonal and partner-related trust typologies in Table 1 as the patient must place trust in the chatbot's goodwill and benevolence, as well as technical reliability and predictability. Conversely, the non-positionality of the bot may provide an advantage in the CAMH study due to its lack of bias and judgment, which may appeal to patients due to the entrenched social stigmas that surround smokers.

4 CONCLUSION

The field of HCI has often proposed the development of chatbots and other autonomous agents to address high patient demand and low institutional resources for therapy and other forms of one-on-one healthcare support [21]. This paper has posed that traditional taxonomies of trust for both human-human and human-machine relationships are problematized by such new artificially intelligent systems, which are designed to adopt vocations long considered the sole domain of human experts. In the case of the smoking cessation chatbot, questions of predictability, expert knowledge, clinical context, and human experience will need to be addressed

to ensure both its success as a motivational interviewer and its widespread adoption as it enters real world contexts. Although this paper stops short from presenting a hybrid taxonomy of trust to accommodate this novel form of automation, the identification of shared and conflicting attributes between current models of trust is intended to provide a foundation for the development of such rubrics beyond the confines of the CAMH pilot study.

ACKNOWLEDGMENTS

The authors would like to thank the other members of the project team, particularly Professors Jonathan Rose and Peter Selby, for their insights and perspectives during the writing of this paper.

REFERENCES

[1] Fahad Almusharraf. 2018. *Motivating smokers to quit through a computer-based conversational system*. Master's thesis. University of Toronto, Toronto, Canada.

[2] Fahad Almusharraf, Jonathan Rose, and Peter Selby. 2020. Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing–based chatbot through iterative interactions. *Journal of Medical Internet Research* 22, 11 (Nov. 2020). DOI: https://doi.org/10.2196/20251.

[3] Bernard Barber. 1983. *The Logic and Limits of Trust*. Rutgers University Press, New Brunswick, NJ.

[4] Luke J. Chang, Bradley B. Doll, Mascha van't Wout, Michael J. Frank, and Alan G. Sanfey. 2010. Seeing is Believing: Trustworthiness as a Dynamic Belief. *Cognitive Psychology* 61, 5 (Sep. 2010), 87–105. DOI: https://doi.org/10.1016/j.cogpsych.2010.03.001.

[5] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 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*, May 4–9, 2019, Glasgow, Scotland, United Kingdom. ACM Inc., New York, NY, 1-12. DOI: https://doi.org/10.1145/3290605.3300705.

[6] Caroline Free, Rosemary Knight, Steven Robertson, Robyn Whittaker, Phil Edwards, Weiwei Zhou, Anthony Rodgers, John Cairns, Michael G. Kenward, and Ian Roberts. 2011. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. *The Lancet* 378, 9785 (July 2011), 49-55. DOI: https://doi.org/10.1016/S0140-6736(11)60701-0.

[7] Peter Hancock, Deborah Billings, Kristin Schaefer, Jessie Chen, Ewart de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 5 (Oct. 2011), 517-527. DOI: https://doi.org/10.1177/0018720811417254.

[8] Christina Hassija and Matt Gray. 2011. The effectiveness and feasibility of videoconferencing technology to provide evidence-based treatment to rural domestic violence and sexual assault populations. *Telemedicine and e-Health* 17, 4 (May 2011), 309–315. DOI: https://doi.org/10.1089/tmj.2010.0147.

[9] John Holmes and John Rempel. 1985. Trust in Close Relationships. *Journal of Personality and Social Psychology* 49, 1 (July 1985), 95–112. DOI: https://doi.org/10.1037/0022-3514.49.1.95.

[10] Donald Hilty, Daphne Ferrer, Michelle Burke Parish, Barb Johnston, Edward Callahan, and Peter Yellowlees. 2013. The Effectiveness of Telemental Health: A 2013 Review. *Telemedicine and e-Health* 19, 6 (May 2013), 444–454. DOI: https://doi.org/10.1089/tmj.2013.0075.

[11] Roy Lewicki, Daniel McAllister, Robert Bies. 1998. Trust And Distrust: New Relationships and Realities. *The Academy of Management Review* 23, 3 (July 1998), 438-458. DOI: https://doi.org/10.2307/259288.

[12] D. Harrison McKnight and Norman Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies: Integrating the Human and Artificial Perspectives.* Rino Falcone, Munindar Singh, and Yao-Hua Tan, Ed. Springer-Verlag, Berlin, Germany, 27-54. DOI: http://dx.doi.org/10.1007/3-540-45547-7_3.

[13] William Miller and Gary Rose. 2009. Toward a Theory of Motivational Interviewing" *The American Psychologist* 64, 6 (Sep. 2009), 527-537. DOI: https://doi.org/10.1037/a0016830.

[14] Wiliam Miller and Stephen Rollnick. 2012. *Motivational Interviewing: Helping People Change*. Guilford Press, New York, NY.

[15] William Miller and Stephen Rollnick. 2009. Ten Things that Motivational Interviewing is Not. *Behavioural and Cognitive Psychotherapy* 37, 2 (Mar. 2009), 129-140. DOI: https://doi.org/10.1017/S1352465809005128.

[16] Bonnie Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 1 (Nov. 1987), 527-539. DOI: https://doi.org/10.1016/S0020-7373(87)80013-5.

[17] Theresa Moyers. 2004. History and Happenstance: How Motivational Interviewing Got its Start. *Journal of Cognitive Psychotherapy* 37, 4 (Oct. 2004), 291-298. DOI: https://doi.org/10.1891/jcop.18.4.291.63999.

[18] Kristin Oleson, D.R. Billings, Vivien Kocsis, Jessie Chen, and Peter Hancock. 2011. Antecedents of trust in human-robot collaborations. In *Proceedings of the 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, February 22-24, 2011, Miami Beach, Florida. IEEE, New York, NY, 175–178. DOI:

https://doi.org/10.1109/COGSIMA.2011.5753433.

[19] Kristin Schaefer, Jessie Chen, James Szalma, and Peter Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors* 58, 3 (May 2016), 377-400. DOI: https://doi.org/10.1177/0018720816634228.

[20] Rebecca Shingleton and Tibor Palfai. 2016. Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. *Patient Education and Counseling* 99, 1 (Jan. 2016), 17–35. DOI: https://doi.org/10.1016/j.pec.2015.08.005.

[21] Sherry Turkle. 2012. *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, New York, NY.

[22] Zheng Yan, Raimo Kantola, and Peng Zhang. 2011. A Research Model for Human-Computer Trust Interaction. In *Proceeding of the 10th International Conference on Trust, Security and Privacy in Computing and Communications*, November 16-18, 2011, Changsha, China. IEEE, New York, NY,. 274–281. DOI: https://doi.org/10.1109/TrustCom.2011.37.

A APPENDICES

A.1 Table 1

Human-Human Trust McKnight & Chervany	**Human-Computer Trust** Schaefer et al.	**Shared Attributes**	**Definition**
Dispositional Trust The general propensity for an individual to be willing to depend on others across a broad spectrum of situations. Refers to people in general rather than specific people.	Human-Related Factors The general propensity for an individual to be willing to depend on an automaton. This propensity is determined by an individual's traits, states, cognitive and emotive factors.	A general focus on human factors that predispose an individual towards the act of trusting. Attribution to the philosophical and demographic traits beyond individual interactions (faith in humanity and confidence in automation).	Dispositional trust is anchored by the idea of benevolence. An individual appraises the general of others and develops strategies for trusting people based on this appraisal. Human-related factors omit the idea of benevolence, owing to the machine's lack of agency. Instead, emphasis is placed on the individual's capacity to embrace the machine in a comfortable, stress-free, and satisfying manner.
Interpersonal Trust The willingness for one individual to depend on another individual with a feeling of relative security. This human-to-human trust is determined through evaluating an individual's predictability, competence, goodwill and benevolence.	Partner-Related Factors The willingness for one individual to depend on an automaton with a feeling of relative security. This human-to-robot trust is determined through machine characteristics such as design affordances and longitudinal technical capabilities such as reliability and predictability.	A general focus on partner-related characteristics and expectations. Attribution to the importance of presentation and performance over time, leading to forecasting of trust (subjective probability and reliability/errors).	Collapses of interpersonal trust are attributed to shortcomings in the integrity or good faith of an individual. While competency is a factor, it is tied more to authority in and control of a situation. Collapses of partner-related factors occur when machine operations are unreliable or unpredictable or when design affordances denote these traits. Partner-related factors are related to quantifiable competency over time.
Institutional Trust The belief that favourable conditions are in place that are conducive to situational success in an endeavour. Refers to formal structures and not the individual people involved.	Environment-Related Factors The belief that automation can enter an existing team or environment in a favourable and risk-free way. Also refers to the broader social and cultural implications of automaton beyond immediate human-to-machine interactions	A general focus on the context of interaction beyond the task-at-hand. Attribution to the larger power dynamics at play beyond the human and partner (situational normality and team composition).	Institutional trust is largely predicated on faith in structures, situations, and roles that serve a protective function in society. In environmental-related factors, there is little presumption that machines will abide by these structures. Instead, attention is paid to managing anxieties regarding social cohesion, well-being, safety, and security on the interaction level as automation inevitably rearranges these institutions.