Designing Contestability: Interaction Design, Machine Learning, and Mental Health

Tad Hirsch
University of Washington
Seattle, USA
thirsch@uw.edu

Kritzia Merced
University of Utah
Salt Lake City, USA
k.mercedmorales@utah.edu

Shrikanth Narayanan
University of Southern California.
Los Angeles, USA
shri@sipi.usc.edu

Zac E. Imel
University of Utah
Salt Lake City, USA
zac.imel@utah.edu

David C. Atkins
University of Washington
Seattle, USA
datkins@uw.edu

ABSTRACT
We describe the design of an automated assessment and training tool for psychotherapists to illustrate challenges with creating interactive machine learning (ML) systems, particularly in contexts where human life, livelihood, and wellbeing are at stake. We explore how existing theories of interaction design and machine learning apply to the psychotherapy context, and identify "contestability" as a new principle for designing systems that evaluate human behavior. Finally, we offer several strategies for making ML systems more accountable to human actors.

Author Keywords
Machine learning, psychotherapy, mental health, interaction design.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; I.2.1. Artificial Intelligence: Applications and Expert Systems; J.4 Applications: Social and Behavioral Sciences: Psychology; Design.

INTRODUCTION
Mental health and addiction problems are among the most common causes of disability in the U.S. [31]. Psychotherapy – a goal oriented conversation between a patient and therapist – represents a class of effective treatments [17]. Performance based feedback is key to promoting the effectiveness of providers [29], but standard approaches that rely on direct observation by humans are slow, unreliable, and can't be offered at scale [22]. Recent work demonstrates that sessions can be evaluated by a machine learning (ML) system that provides summaries of different types of therapist interventions [3]. Such performance-based feedback is expected to assist skill development and retention, leading to better outcomes for patients [29].

DIS 2017, June 10-14, 2017, Edinburgh, United Kingdom

ACM ISBN 978-1-4503-4922-2/17/06…$15.00
http://dx.doi.org/10.1145/3064663.3064703

However, designing interactive systems that rely on machine learning algorithms requires rethinking core assumptions about user control [2]. Accordingly, design researchers have begun to articulate new approaches to such systems. For example, Horvitz' principles for effective "mixed-initiative" systems include querying users for clarification about goals and preferences, and scoping system precision to match user needs [11]. Yang et al describe patterns for designing adaptive user interfaces [33].

Perhaps more provocatively, Allen argues that HCI's prevailing model of human-controlled systems needs to evolve to an interaction-based "dialogue" between people and machines [1]. Leahu similarly calls for fundamental shifts in how we think about relationships between people and machines [19], citing Taylor's observation that interaction designers typically assume that technologies are clearly delineated from and subservient to people [28]. In contrast, Suchman and Bødker have argued that agency resides in both human and nonhuman actors [26][4]. Machine learning algorithms would seem to further complicate this paradigm, presenting systems in which agency is clearly shared (and in some cases, weighted towards the machine). Accordingly, Verbeek suggests a hermeneutical approach to designing systems in which people and machines collaboratively interpret a seemingly inscrutable world [30].

Prior work on interaction design and machine learning systems has often focused on consumer applications, including various recommender systems. Much of the early thinking was necessarily speculative, conducted at a time when ML systems were relatively uncommon. In recent years, these systems have increasingly "become real" [10], finding widespread adoption and media attention. At the same time, they are finding their way into human service domains, including healthcare [32], public safety [8], and

transportation planning. With these applications, they are being used not simply to predict user preferences, but to anticipate and evaluate human performance in contexts involving human life, livelihood, and wellbeing.

In this Note, we describe the design and implementation of a ML-based feedback system for psychotherapy to illustrate challenges in creating ML systems for such "high stakes" domains. As ML systems grow in popularity and expand the contexts in which they are employed, there is a greater urgency to articulate their attendant design challenges and patterns. Further, it is incumbent upon designers to develop principles that address social and ethical issues, in addition to pragmatic concerns. Our contribution is intended to advance this discourse.

CORE-MI

This paper discusses design issues that have arisen while developing CORE-MI, a system that uses speech and language processing to automatically generate evaluations of counseling sessions directly from session audio [14]. The system employs ML models trained on prior human ratings, whose reliability has been previously reported [5].

CORE-MI is designed for motivational interviewing (MI), which is an evidence-based psychotherapy focused on behavior change for substance abuse and health behavior problems [23]. The system uses machine learning algorithms to transcribe and evaluate the quality of therapy relative to best practices established in the research literature. The system provides an interactive report card-like, visual summary of counseling sessions.

CORE-MI is expected to be used for both provider training and supervision, and is currently being deployed in several clinics. We have employed a participatory, iterative design process in developing CORE-MI. Our research team includes designers, engineers, mental health researchers, and counselors. We have also conducted several information sessions with prospective users, which has enabled the design team to gauge user interest and identify early-stage concerns. This paper reports lessons learned through the design process to date. A formal user evaluation is currently underway, and will be reported in future publications.

DESIGN ISSUES

We have encountered three main interaction design challenges in developing CORE-MI. Two of these – incentives and transparency – are mentioned in the design literature, although not related to psychotherapy. We also identify "contestability" as a new class of concern, particularly relevant for systems that evaluate human performance.

Incentivizing Participation

Like many machine learning systems, CORE-MI relies on end-user input to train and improve its algorithms. As such, we must design appropriate incentives that encourage users to provide feedback on model predictions [16]. We identify two components of incentive system design: the reward model and the transaction model.

Reward model

Incentive systems typically employ intrinsic or extrinsic rewards to motivate user engagement [16]. Intrinsic rewards are inherently linked to the activity that a model is intended to support. For example, Amazon's recommender system solicits users' product ratings by promising better recommendations that will be more valuable to the user. By contrast, there is little or no logical connection between extrinsic rewards and user action. For example, researchers may offer micropayments [13] or "gamification" elements (points, badges, etc.) [20] in exchange for data labelling services.

Our design emphasizes intrinsic rewards. For example, we promise increased quality of care by offering accurate, inexpensive training and assessment. We are also developing features to automate session documentation, which should reduce counselor workloads. While not directly tied to session evaluation, these features are hoped to increase buy-in and encourage users to provide feedback.

We have been hesitant to explore extrinsic rewards. Ours is a specialized, professional user group with deep commitments to craft and patient outcomes. We are concerned that gamification-like approaches could alienate users and trivialize their work. That said, we recognize that incentives may appeal differently to various users – for example, students, early career therapists, and seasoned providers may respond differently to, say, receiving cash payments for coding transcripts.

Transaction model

The "transaction model" describes the mechanisms through which rewards are conferred. Some incentive systems reward users immediately following participation, in a quid pro quo manner. For example, reCAPTCHA [18] is a popular authentication service in which users provide semantic labels for images before they are allowed to access a desired website. Once a user provides a label, they immediately proceed to the desired content or service.

In contrast, other systems employ a promissory transaction model in which users are enticed to provide data with assurances of future benefit. While these models often appeal to user self-interest, some projects also make altruistic appeals that link user participation to others' benefit (typically at an unspecified future date). The Open Mind Common Sense project, for example, seeks to build a large, crowdsourced dataset of commonsense assertions that will benefit the artificial intelligence field generally, but makes little or no promise of direct benefit to individual contributors [6].

We expect to employ both immediate and promissory transactions. For example, automatically generated session transcripts will be updated in response to user corrections, which will immediately result in revised codes and

summary reports for the session. At the same time, we promise improved models and feedback over time, based on user edits. More accurate ratings by the CORE-MI system are claimed to improve quality for current and future users, a mixing of selfish and altruistic encouragement.

Our assumption is that combining transaction models will lead to greater user buy-in and sustained, high-quality input. While it is expected that mixed approaches should be effective, there is limited research thus far ([12] is a notable exception).

From Transparency to Legibility

There is a longstanding principle in interaction design that users should have a conceptual model, or intellectual understanding of how systems work. With regard to machine learning algorithms, this is typically described in terms of "transparency", or of "opening the black box" (e.g.[15]) to reveal to users the inner mechanisms that drive computer prediction.

While transparency is an admirable goal, it may be unachievable for certain kinds of algorithms. In many cases, the interactions between large numbers of variables upon which predictions are made are so complex as to defy easy comprehension by non-expert users – consider economic forecasts, as one example. More challenging still are deep learning neural network approaches and other machine learning systems in which predictions result from thousands of hidden nodes and layers. Such systems are inherently obscure: while their behavior can be observed and evaluated, the precise mechanisms through which decisions are made are unknowable even to their designers, and thus cannot be made fully transparent to users.

Several approaches have been suggested to enhance user comprehension. Höök describes a "black box in a glass box," approach that enables users to query inputs and outputs, but does not provide insight into a system's inner workings [9]. Another method involves "inverting" models, prompting them to generate representative examples that provide insight into how their algorithms function [21] (cited by [19]).

Glass-box-in-a-black-box and inversion do more than make algorithms visible; they help people make sense of an algorithm's behavior. Emphasizing the degree to which a system can be deciphered and interpreted – its legibility – suggests an opportunity to enhance user understanding through "sandboxing" techniques that allow users to probe models with various inputs to investigate their effects on predictions. For example, we could allow users to manipulate percentages of therapist vs. patient talk time or edit sample transcripts to see their effects on overall efficacy scores. Or, we might experiment with having the system generate its own transcripts of what it considers to be high-quality therapy.

Importantly, legibility is linked to users' trust and willingness to adopt. For example, [25] found strong user preference for, and trust in, models that exhibit "sound" (i.e., human comprehensible) reasoning and "clear communication" about decision making. These models were also perceived as more accurate, which did not necessarily correlate with actual or statistical accuracy.

This points to an interesting tension for system designers. In some cases, human interpretable models may be preferable, even if they are less statistically accurate. We suspect that tradeoffs between legibility and accuracy will be particularly important in applications that evaluate human performance. For these applications, trust in a system's "soundness" – and by extension, its fairness – will be a crucial factor in people's willingness to abide by its edicts.

Our system employs a variety of algorithms and measures, of varying degrees of human-comprehensibility. On the one hand, summaries of types of therapist reflections and questions are broadly interpretable, and will be familiar to many of our users (although, the hidden Markov models that the system uses to identify reflections and questions may not be). Other measures are more complex. For example, we offer interpreted scores to indicate a users' overall adherence to MI "spirit." While users of an earlier version of our software expressed interest in interpreted measures [7], informal evaluation suggests that they struggle to understand how they are determined. Moving forward, one of our design challenges will be to provide adequate descriptions of these measures, or to consider dropping them despite their statistical accuracy.

Facilitating Contestation

CORE-MI is currently being implemented in several clinics, and we have already heard some counselors express concern about being "judged by a machine." We understand that such sentiments are bound up with concerns about evaluation more generally, and likely vary with professional status. Students, for example, expect to be evaluated during training, whereas this is exceedingly rare for therapists in private practice. Resistance to evaluation may be exacerbated by automated tools, because, unlike a human supervisor, a trainee cannot inquire about the ratings, nor engage the evaluator about their assessment.

The problem of contestation, of challenging machine predictions, becomes particularly acute when we recognize that our models are, and will continue to be, fallible, and the risks of "getting it wrong" can be quite high for therapists and patients alike.

Imperfect machine learning predictions can limit their appeal and misguide users. Providers may perceive correcting transcripts as a problematic addition to their workload. Perceptions of inaccuracy may also undermine confidence and suppress user interest, casting doubts on system credibility while also undermining its benefits. Equally troubling, overconfidence in model outcomes may lead to negative results for users. For instance, supervisors might weight model predictions too heavily in job

performance evaluations, or trainees might adapt their practices to improve machine scores in ways that are ultimately detrimental to patient care.

We therefore must provide mechanisms for users to challenge model predictions. We already see this capability in small ways with recommender systems. For example, Netflix may predict that I'll love the film "Gigli," but I can override this suggestion with my own 1-star rating. Simple correction like this may work with discrete predictions; e.g. word errors from automated speech transcription should be easily correctable. But there the will also be a need for more nuanced and substantial argumentation.

Doing so may require users to marshal evidence and create counter narratives that argue precisely why they disagree with a conclusion drawn by an AI system. This becomes particularly important when the user cannot simply register disagreement with the system, but rather, must make arguments to powerful actors whose decisions are informed by those systems. Think, for example, of a mental health provider whose clinical supervisor relies on automated assessment technologies to evaluate performance, or an insurance company that employs machine learning algorithms in determining whether to cover an addiction treatment program. In cases like these, therapists and trainees will need to be able to access and annotate session transcripts, as well as provide additional contextual detail that might not be available to the algorithm. For example, a therapist who suspects that a patient is inebriated during a session might tailor her discussion in ways that lead to lower scores. Such contextual information would need to be included in any assessment of system performance.

DESIGN FOR CONTESTABILITY

In approaching contestability, we invoke Verbeek's ethical imperative [30] to anticipate and design for potential mediations that our system may embody. In our case, this means thinking beyond the experience an individual has while using our software, and recognizing that our technology can shape relationships between therapists, supervisors, insurance providers, patients, and other stakeholders. In particular, we recognize the potential for ML to be used as a blunt assessment tool by managers and businesses, to the detriment of therapists and patients. Indeed, we acknowledge that there will likely be significant financial and organizational pressures to do so, as our technology occurs at a time of increasing pressure to "rationalize" mental health care [24].

This concern is somewhat speculative, as systems like ours are fairly novel. Nonetheless, we identify several design strategies that can be put in place now, to help mitigate future misuse.

First, we strive to improve the accuracy of our models. This is achieved through phased deployment with expert users in training clinics and universities, and with students who are not expected to perform perfectly and who can be more easily incentivized to provide feedback. By running through thousands of sessions in education contexts, we hope to develop models that are reasonably accurate by the time they are released in broader healthcare settings.

Second, we strive to make our models as legible as possible. We provide detailed explanations of each measure, and highlight confidence scores to indicate the degree of certainty in each prediction. We will also provide mechanisms for users to unpack aggregate measures, tracing system predictions all the way down to the transcript level so that users can follow, and if necessary, contest the reasoning behind each prediction. At the same time, we recognize that certain measures defy human interpretation and will continue to evaluate the efficacy that these measures provide in light of their potential to foster distrust. Moving forward, we may decide that the value that some measures is outweighed by their cost in user confidence and trust.

Third, alongside technology development we anticipate designing training modules for therapists and supervisors that describe how the system works, including discussion of its strengths and limitations. These modules may include sandboxing features that allow users to experiment with inputs and outputs. By teaching users about the system's capabilities, we hope to discourage inappropriate use and interpretations of model output, and to create a shared understanding that can act as a basis for addressing inevitable disagreements.

Finally, we expect to remain vigilant about potential misuse and implicit bias. This includes providing mechanisms for users to ask questions and record disagreements with system behavior. It also means looking for aggregate effects that may not be apparent to individual users. For example, [27] has demonstrated racial bias in Google's AdSense system that were only made visible by looking at behavior across multiple users and sessions. We take this lesson seriously and hope to implement programs that can monitor for such effects on the behalf of vulnerable users.

The above suggestions arise from our experience designing and implementing a machine learning system for psychotherapy. However, we believe that the principles we describe will have wider applicability, particularly for ML systems that predict and evaluate human behavior in contexts that affect human life, livelihood, and wellbeing. That said, we offer these recommendations tentatively, as a step towards articulating a set of design principles and practices for this emerging area. We look forward to exploring and evaluating them more fully in future work.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Health through grants AA018673, DA034860, and AA023814.

REFERENCES

[1] Allen, J.E. et al. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems*. 14, 5 (Sep. 1999), 14–23.

[2] Amershi, S. et al. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*. 35, 4 (2014), 105–120.

[3] Atkins, D.C. et al. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*. 9, 1 (Dec. 2014).

[4] Bødker, S. 1991. *Through the interface: a human activity approach to user interface design*. L. Erlbaum.

[5] Can, D. et al. 2016. "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*. 63, 3 (2016), 343–350.

[6] Catherine Havasi et al. 2010. Open mind common sense: crowd-sourcing for common sense. (2010).

[7] Cook, J. and Hirsch, T. 2014. Monologger: visualizing engagement in doctor-patient conversation. (2014), 37–40.

[8] Emerging Technology from the arXiv 2016. Neural Network Learns to Identify Criminals by Their Faces. *MIT Technology Review*.

[9] Höök, K. et al. 1996. A glass box approach to adaptive hypermedia. *User Modeling and User-Adapted Interaction*. 6, 2–3 (Jul. 1996), 157–184.

[10] Höök, K. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers*. 12, 4 (Feb. 2000), 409–426.

[11] Horvitz, E. 1999. Principles of mixed-initiative user interfaces. (1999), 159–166.

[12] Huang, Y. et al. 2016. Combining contribution interactions to increase coverage in mobile participatory sensing systems. (2016), 365–376.

[13] Irani, L.C. and Silberman, M.S. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. (2013), 611.

[14] James Gibson et al. 2016. Developing an Automated Report Card for Addiction Counseling: The Counselor Observer Ratings Expert for MI (CORE-MI). (San Jose, CA, 2016).

[15] Jameson, A. 2003. Adaptive interfaces and agents. *The human-computer interaction handbook*. L. Erlbaum Associates Inc. 305–330.

[16] Kraut, R.E. et al. 2011. *Building successful online communities: evidence-based social design*. MIT Press.

[17] Lambert, M.J. 2013. *Bergin & Garfield's handbook of psychotherapy and behavior change*. Wiley.

[18] Law, E.L.M. and Von Ahn, L. 2011. *Human computation*. Morgan & Claypool Publishers.

[19] Leahu, L. 2016. Ontological Surprises: A Relational Perspective on Machine Learning. (2016), 182–186.

[20] Lindqvist, J. et al. 2011. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. (2011), 2409.

[21] Mahendran, A. and Vedaldi, A. 2015. Understanding deep image representations by inverting them. (Jun. 2015), 5188–5196.

[22] Proctor, E.K. et al. 2009. Implementation Research in Mental Health Services: an Emerging Science with Conceptual, Methodological, and Training challenges. *Administration and Policy in Mental Health and Mental Health Services Research*. 36, 1 (Jan. 2009), 24–34.

[23] Rollnick, S. and Miller, W.R. 1995. What is Motivational Interviewing? *Behavioural and Cognitive Psychotherapy*. 23, 04 (Oct. 1995), 325.

[24] Scheid, T.L. 2003. Managed Care and the Rationalization of Mental Health Services. *Journal of Health and Social Behavior*. 44, 2 (Jun. 2003), 142.

[25] Stumpf, S. et al. 2007. Toward harnessing user feedback for machine learning. (2007), 82.

[26] Suchman, L.A. 2007. *Human-machine reconfigurations: plans and situated actions*. Cambridge University Press.

[27] Sweeney, L. 2013. Discrimination in Online Ad Delivery. *Queue*. 11, 3 (Mar. 2013), 10.

[28] Taylor, A. 2015. After interaction. *interactions*. 22, 5 (Aug. 2015), 48–53.

[29] Tracey, T.J.G. et al. 2014. Expertise in psychotherapy: an elusive goal? *The American Psychologist*. 69, 3 (Apr. 2014), 218–229.

[30] Verbeek, P.-P. 2015. COVER STORYBeyond interaction: a short introduction to mediation theory. *interactions*. 22, 3 (Apr. 2015), 26–31.

[31] Whiteford, H.A. et al. 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*. 382, 9904 (Nov. 2013), 1575–1586.

[32] Yang, Q. et al. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. (2016), 4477–4488.

[33] Yang, Q. et al. 2016. Planning Adaptive Mobile Experiences When Wireframing. (2016), 565–576.