Integrating Artificial and Human Intelligence in Complex, Sensitive Problem Domains: Experiences from Mental Health

Munmun De Choudhury, Emre Kiciman

■ *This article presents a position highlighting the importance of combining artificial intelligence approaches with human intelligence, in other words, the involvement of humans. To do so, we specifically focus on problems of societal significance, stemming from complex, sensitive domains. We first discuss our prior work across a series of projects surrounding social media and mental health, and identify major themes for which augmentation of AI systems and techniques with human feedback has been and can be fruitful and meaningful. We then conclude by noting the implications, in terms of opportunities as well as challenges, that can be drawn from our position, both relating to the specific domain of mental health and for AI researchers and practitioners.*

Artificial intelligence methods are becoming a critical tool for impacting a variety of domains of broad societal significance (Boyd and Crawford 2012), from economic development (Jean et al. 2016) and education (He et al. 2015) to the environment (Dietterich 2009) and agriculture (Vasisht et al. 2017). A significant strength of AI in domains such as these is its ability to turn new sources of data into signals relevant to a domain. These new data sources allow, for example, AI to expand our ability to more easily reach and help vulnerable populations, to more quickly detect people at risk of poor outcomes, to identify customized or personalized solutions, and to enable early interventions.

Many of these societally significant domains are complex — understanding the mechanisms, dynamics, and interactions at work is challenging, and because the issues involve personal information and artifacts about individuals, they require careful, responsible attention among researchers and stakeholders. Further, not only do these problems involve using AI to derive insights from data, but they also require determining if those insights are practical and can be used to help relevant domain experts and stakeholders. Consequently, AI alone provides only a partial perspective when the goal is to interpret and translate the methods and findings to real-world settings. To validate and complement a particular AI analysis, we must go beyond a particular dataset or regime and bring in external domain knowledge of assumptions and plausible mechanisms. Judeah Pearl notes the limitations of approaches informed only by "naked data" and argues that one needs knowledge from outside the data (Pearl 2018):

> Data science is only as much of a science as it facilitates the interpretation of data — a two-body problem, connecting data to reality. Data alone are hardly a science, regardless how big they get and how skillfully they are manipulated.

In our experiences applying artificial intelligence methods to the analysis of new data sources to better understand the complex and sensitive domain of mental health, we have often drawn on human intelligence for prior knowledge, oversight, and analysis to augment pure AI methods. Four of the broad issues we have come across that have required such augmentation with human intelligence: ensuring construct validity, assumptions on unobserved factors, understanding data biases, and navigating sensitivities.

Ensuring Construct Validity

First, when using AI to extract core measurements, we are concerned with the construct validity of these measures. That is, are we actually measuring what we think we are measuring? For example, if we are trying to measure mood from the language people use on social media, are the words they use reflective of the moods they are actually experiencing? While this may sometimes be the case, self-presentation bias, cultural norms, word ambiguities, and even song lyrics can complicate the association between people's experienced moods and their expression on social media. If not recognized and corrected, these false associations can entirely threaten the validity of our measurements and, through them, any conclusions we might wish to draw from the data.

Assumptions on Unobserved Factors

Second, when we are attempting to understand a phenomenon through its representation in data, we must be aware that our observations may be significantly influenced by unobserved factors. When using AI methods to model people's behaviors and their reactions to an event or treatment within the data, for example, we must take into account that people will also be affected by external cultural factors, social influence, seasonal dynamics, larger trends, and other events not captured within the data. How these factors manifest can vary as well: each may vary across individuals in a dataset, or, alternatively, affect all individuals simultaneously. These unobserved factors can confound our understanding of the situation, causing us to misunderstand the underlying mechanisms and draw the wrong conclusions about the severity of a situation or about the recommendations for action to improve a situation.

Understanding Data Biases

Third, when using AI for data-driven learning about a complex domain, we must have an understanding of the biases within the data being studying. Due to limitations in the data, it is possible that our learnings are only valid under certain situations or for a certain group of people. In the context of mental health, for example, the complexities of the domain mean that conclusions drawn based on a limited subpopulation might be very different than conclusions drawn for another subpopulation or for the population as a whole. To generalize what we are learning, we must have validation that the people and the specific situations we are studying through a dataset are representative of the broader phenomenon we care about.

Navigating Sensitivities

Finally, many of the societally relevant domains where AI frameworks and tools have been found to be promising also tend to be areas where decision-making is high stake and high cost, meaning that mistakes and errors can have serious implications for human life, both figuratively and literally, as well as for human cost. In other words, if AI is employed to make decisions in an automated fashion, errors are unacceptable, although building 100 percent fault-free AI systems is far from reality today. To realize the potential of AI in these critical and important domains, the involvement of humans and experts is paramount, to ensure that there are adequate mechanisms to circumvent the mistakes made by the AI systems, to ensure that adequate risk mitigation protocols are in place when inappropriate or dangerous decisions are made by AI, and also to ensure that decision-making is arrived at in some collaborative fashion between the AI system and humans.

These issues are not concerns only when using sophisticated AI methods, of course. Construct validity, unobserved factors, data biases, and domain sensitivities are challenges faced by any quantitative analysis. We argue, however, that these issues are particularly critical threats to the validity of AI-driven analysis because of the challenges of interpreting and understanding the limitations of many black-box AI methods. Moreover, the challenges of interpretability and understanding are exacerbated and the cost of mistakes are magnified in a complex and sensitive domain.

In the remainder of this article, we highlight our experiences using artificial and natural intelligence together in complement, and discuss open challenges and opportunities for future research. The specific domain we focus on for our position concerns mental health.

Integrating Artificial and Human Intelligence

In this section, we discuss some key methods for augmenting AI approaches with the help of natural intelligence, specifically, human involvement. We draw from a variety of projects in our prior research that surround the complex domain of mental health.

Source of Gold Standard Information

One of the common places where researchers tend to

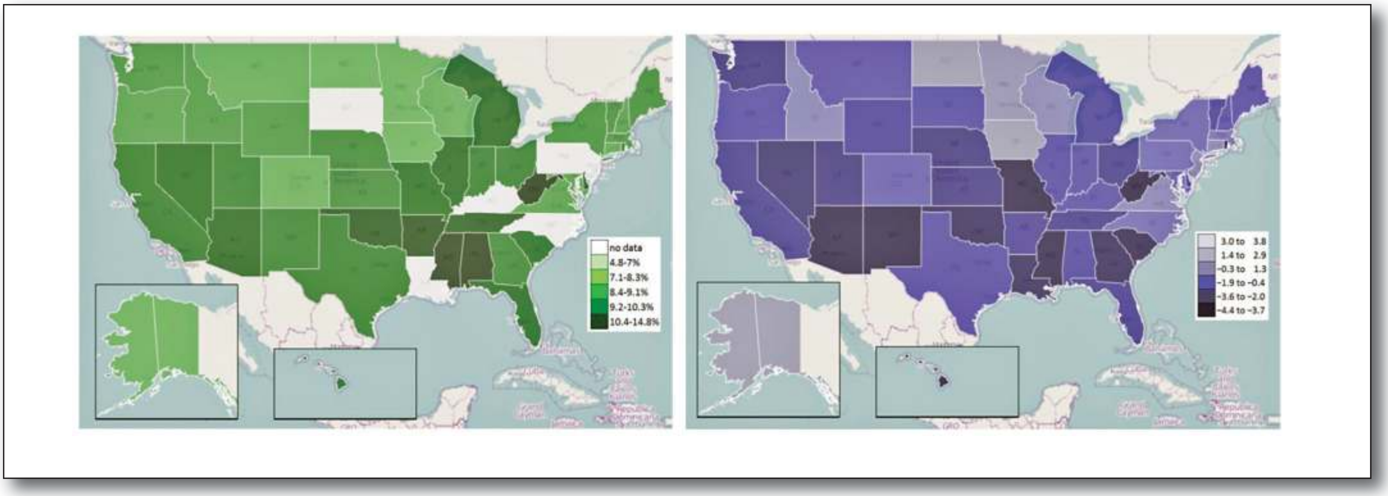

Figure 1. Social Media Index of Depression Compared to Self-Reported Survey Data.

Our prior work showing (on the left) heatmap rendering of actual CDC data and (on the right) Twitter-predicted depression in various US states (De Choudhury, Counts, and Horvitz 2013). Note that in both figures, higher intensity colors imply greater depression. A linear regression fit between the actual and predicted rates shows positive correlation of 0.51.

leverage human intelligence in their social media data modeling and analyses lies in gathering gold standard information that can later be employed in supervised learning models. This gold standard information often also acts as a test of the construct validity of the underlying measures. In the domain of mental health, this approach translates to compiling ground truth information about the true mental health states of individuals, communities, and populations that is independently assessed beyond what the AI techniques may provide.

In our prior work, we have extensively utilized this form of human feedback. For instance, we used crowdsourcing, particularly through the Amazon Mechanical Turk platform, to collect (gold standard) assessments from several hundred (nearly 400) Twitter users who reported that they have been diagnosed with clinical depression, using the CES-D (Center for Epidemiologic Studies Depression Scale)(Eaton et al. 2004) screening test (De Choudhury et al. 2013). Based on this cohort for whom we had offline assessments of depression, we developed several affective, behavioral, cognitive, linguistic, and domain-specific measures and used them to develop AI techniques that quantify an individual's social media behavior for a year in advance of their reported onset of depression (as assessed from their offline psychometric data). Then we leveraged these multiple types of signals from these measures to build a depression classifier that distinguished an at-risk cohort from a control group, and was able to predict, ahead of onset, whether an individual is vulnerable to depression. Our models show promise in predicting outcomes with an accuracy of 70 percent and precision of 0.74.

Further, we evaluated this model by comparing it with gold standard offline statistics of prevalence of depression in the United States (De Choudhury, Counts, and Horvitz 2013). As shown in figure 1, we found our social media index of depression compared well with the rates, obtained via self-reported survey data, as given by the Centers for Disease Control and Prevention. Similar approaches were used in other work from our team. This work includes research that developed AI techniques to leverage Facebook data and self-reported information to predict risk of postpartum depression in new mothers (De Choudhury et al. 2014), shown in figure 2, and that employed expert-generated clinical appraisals from clinical psychologists and psychiatrists to assess and curate the quality of online data related to schizophrenia and psychosis (Birnbaum et al. 2017). In the latter, in particular, expert feedback and ground truth on psychosis allowed us to situate the trends and patterns derived from individuals' social media data into what is known about the illness, its diagnosis, and its trajectory over time. We found that compared to a control group, the psychosis cohort exhibited marked linguistic changes on Twitter in the period following their self-disclosure of their illness on Twitter. After verifying these changes with expert annotations, we found the post-disclosure period to be characterized by lowered stereotypy such as word repetitiveness (–24 percent) and linguistic complexity (–63 percent) and by increased readability (+47 percent) and topical coherence (+81 percent). Figure 3 illustrates these findings.

In a similar vein, in a different work (Chancellor et al. 2016), we employed feedback from clinical psychologists as gold standard information to develop an inference model for mental illness severity (MIS) in pro-eating disorder posts on Instagram. Instead of

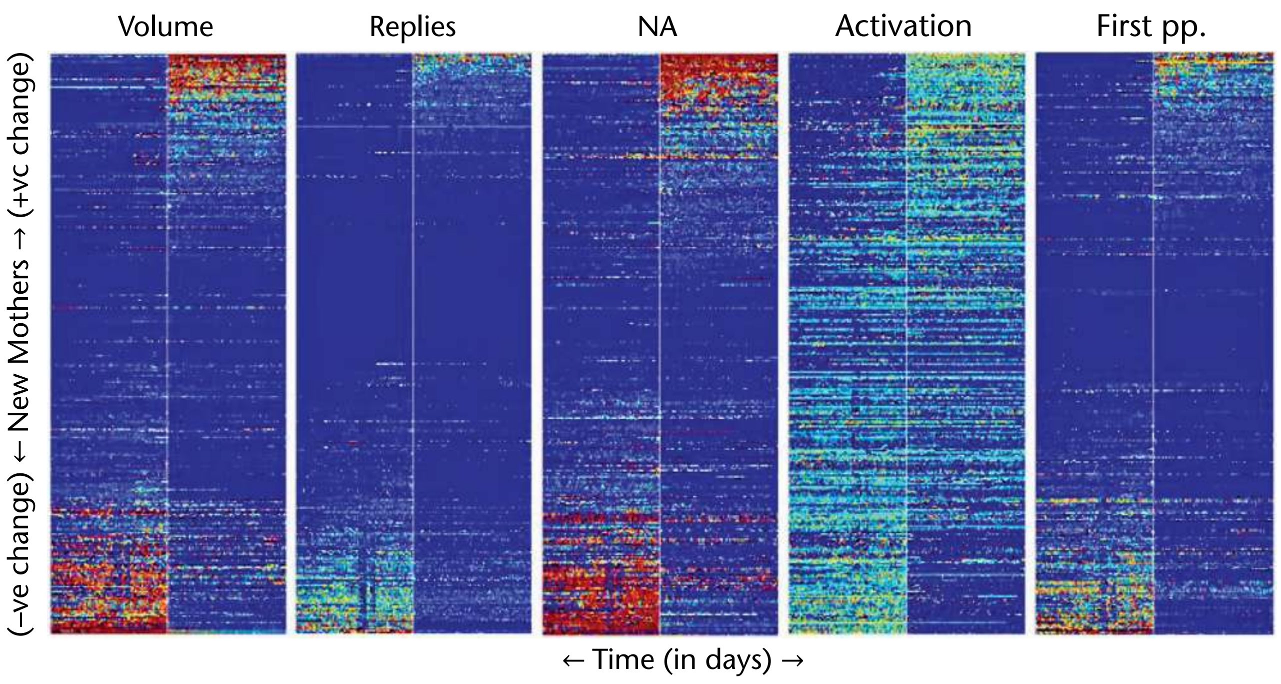

Figure 2. Facebook Data and Self-Reported Information to Predict Risk of Postpartum Depression.

Our prior work examining social media–based postpartum changes in activity, socialization, affect, and interpersonal attention of new mothers (De Choudhury et al. 2014). The heatmaps show individual-level changes in the postnatal period, compared to the prenatal phase. For 15 percent of mothers, these changes (for example, increase in NA and activation) are considerably higher following childbirth.

getting expert annotations on posts directly, a method that does not scale well to large datasets, we obtained them on outcomes of topic models. This strategy allowed us to scale our inference framework to a large corpus of Instagram posts, where we developed a semisupervised approach to map the labels on the topics to posts from users. Examples of high MIS content spans from expression of negative self-perceptions to disordered thoughts about eating to graphic illustration of acts that could lead to physical and emotional harm or death. This coincorporation of human feedback as gold standard information and of analytical AI-based data enabled deep explorations into the manifestation of MIS on the Instagram platform. We found that users who share pro-eating disorder content on Instagram exhibit a trend of increasing MIS in their content over time.

Interpreting Large-Scale Analysis

As we noted, AI approaches can be complemented with human feedback for interpreting the outcomes of an analysis or a computational model. Another way to combine AI methods with human intelligence is to have experts contextualize the AI findings in existing theory or theoretical/conceptual frameworks. By integrating knowledge from existing theories and frameworks, we can test our understanding of underlying mechanisms and our assumptions on unobserved factors that might be affecting our conclusions.

In prior joint work (De Choudhury et al. 2016), the authors developed a causal inference framework (Pearl 2009) to assess the likelihood that an individual will transition to discussions of suicidal ideation, given a history of mental health discourse on social media. This framework was developed on a large dataset of 880 users who shared more than 12K posts and 100K comments on the social media site Reddit. The output of the framework included words and phrases that indicated the likelihood of future suicidal ideation, given their usage in a post. Specifically, we applied a high-dimensional stratified propensity score method (Rosenbaum and Rubin 1983). This approach attempts to isolate the effects of a particular treatment from the effects of covariates by dividing the treatment (those who use a particular

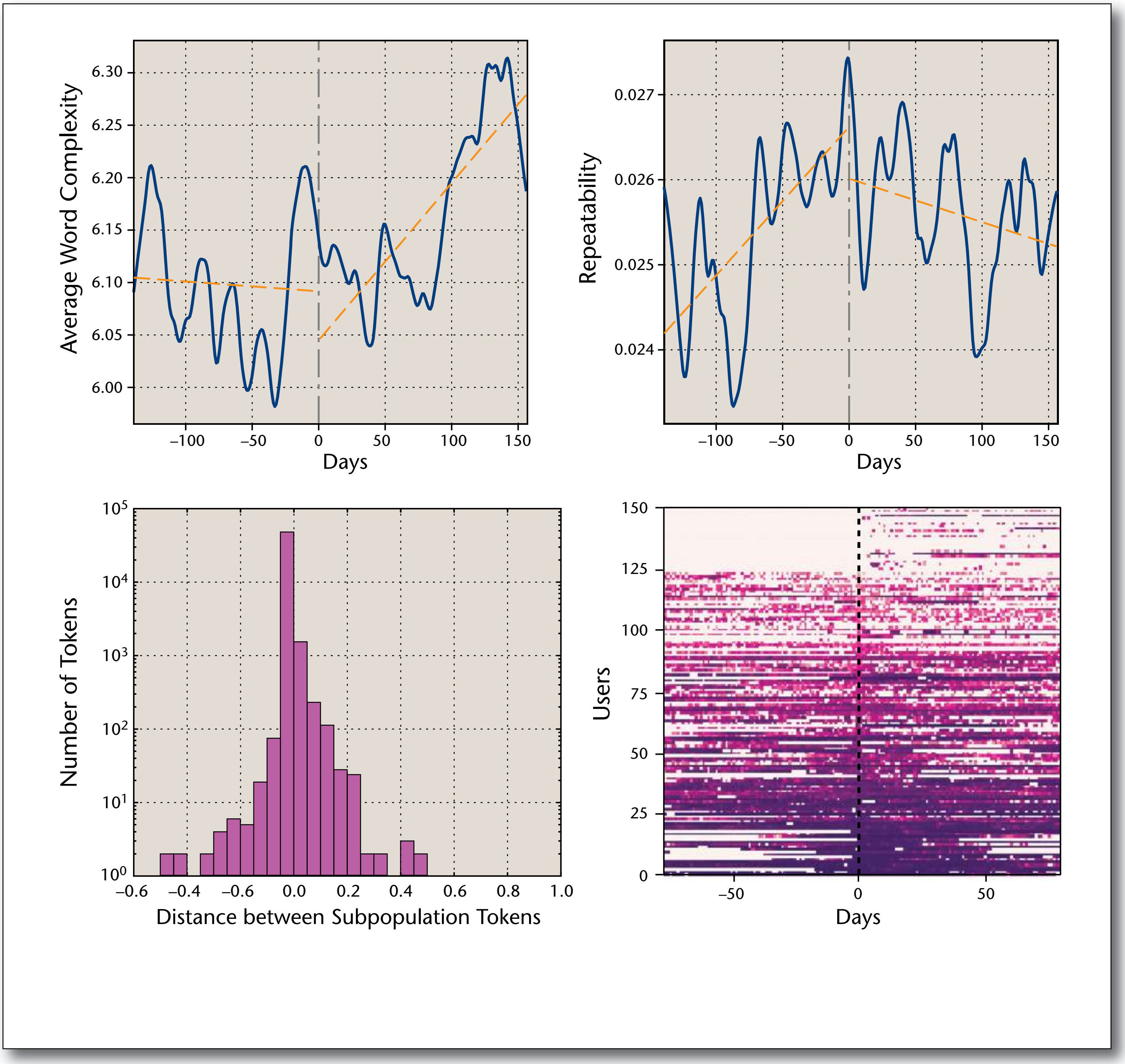

Figure 3. Post-Disclosure Period Changes.

Our work (Ernala et al. 2017) showing notable changes in linguistic organization in a clinically appraised psychotic population following diagnosis disclosures on Twitter (day 0 is day of disclosure).

word/phrase) and control groups (those who do not use the same word or phrase) into strata where the covariates of the treatment subgroup within a strata are statistically identical to the covariates of the control subgroup. Each strata is thus, in essence, artificially approximating a randomized controlled trial where the "assignment" of a treatment is statistically uncorrelated with covariates, allowing us to better distinguish the possible causal effects of a treatment on the outcome, in this case, being whether or not a specific Reddit user posted about suicidal ideation.

However, we noted that the linguistic cues, given by the aforementioned the causal framework (see table 1), did not allow us to examine how specific types of risk markers were associated with suicidal ideation, as illustrated in clinical psychology theories. To enable such comparison, we clustered these linguistic cues via spectral clustering to identify, via expert annotations, what themes led to increases or decreases in suicidal ideation. Then, we qualitatively, using reviews from the same experts, interpreted these themes with the sociocognitive model of suicide (Rudd 1990), to understand what risk markers of suicide are manifested in social media, and to what extent the linguistic cue clusters align with what is known from existing theories in the psychology domain to exacerbate or alleviate the risk of suicidal ideation.

For instance, we found themes containing words or phrases like “have nothing,” “no real,” “kill myself,” “abandoned,” and “die” that experts noted to relate to signals of hopelessness among individuals. The cognitive psychological integrative model of suicide (Dieserud et al. 2001) has identified hopelessness as a mediating variable between mental illness and suicidal ideation and there is ample evidence of the decisive role of hopelessness as an indicator both of current suicide intent and as a predictor of future suicidal behavior (Kashden et al. 1993; Glanz, Haas, and Sweeney 1995):

> But I want to di*e*. I feel so *abandoned*. I must be *an idiot*. I hope for some random event to kill me so that nobody has to be guilty. My loved ones would mourn me but they would move on. At least easier than if I actively killed myself.

We also observed manifestation of impulsive tones in a different theme given by the spectral clustering approach and labeled by the experts. The cognitive suicide model also suggests that impulsivity resulting from cognitive deficits (for example, cognitive rigidity, dichotomous thinking, inability to generate or act on alternative solutions) are prominent markers of suicide ideation (Beck 1979; Kashden et al. 1993):

> Theres a terrible feeling through my whole body every waking moment I have and theres only 2 ways to *ending it*. It hasnt been getting better only worse, I am *freaking out*. The only thing stopping me is I dont know about/have access to anything that would make it quick and clean

Next, the cognitive suicide model has further found lowered self-esteem and self-efficacy to be important attributes among those who are prone to suicide ideation (Schwarzer and Fuchs 1995). Feelings of social isolation and loneliness, conceptualized as a part of the cognitive vulnerability, have consistently been shown to be related to suicidal ideation, attempts, and completions (Bonner and Rich 1988). We found that tokens of one of the extracted themes contained a tone of decreased self-esteem, including that of guilt, self-loathing, and regret:

> I am too ugly to even make friends. I *hate it*. People do not want to be associated with me because of my image. I have tried talking to girls and they’ve all told me to go away and to just give up. So here I am, *giving up* and ending everything.

Together, these findings demonstrate that when human involvement is sought in interpreting the outcomes of large-scale AI approaches, we obtain a much richer, grounded understanding of the specific problem context. Moreover, interpreting results within the context of existing theories also provides a way to test the ability of our conclusions to generalize beyond our specific analysis, to generalize beyond specific datasets, or to focus on specific social media sites.

Improving Computational Models

In this subsection describing the coutilization of artificial and human intelligence for mental health, we describe joint work of the authors in which human insights were incorporated to revise the outcomes of a computational AI-based framework. Such approaches can be a way to fill in the gaps left behind by use of AI techniques alone, especially those gaps that are attributed to the limited “view” on human behaviors and mental health allowed by AI approaches.

We briefly summarize such an approach from our prior work. Utilizing comments received on posts shared in Reddit mental health communities as a proxy for social support, in recent research (De Choudhury and De 2014; De Choudhury and Kiciman 2017), we developed a human-machine hybrid statistical methodology that modeled and quantified the effects of the language of these comments in individuals who do and do not post on a suicide support community in Reddit. Applying stratified propensity score matching (Caliendo and Kopeinig 2008) in a iterative fashion, similar to the approach previously described, we first identified linguistic features (words/phrases) in comments that showed significant effects. We realized that, while the comparability of posts is conventionally judged through purely statistical measures, in our domain these statistics over low-level textual features may miss higher-level semantics of the text. Our contribution lay in realizing that because treatment assignment (that is, the provision of social support) is performed by human commenters who are replying to posts, we can augment our statistical analysis of balance with expert human assessments of balance.

Thus, we obtained expert assessments on the presence of suicidal ideation risk markers in posts associated with these features, for which the rater relied on their offline understanding and knowledge of the risk markers of suicide. Across different propensity strata, the raters specifically assessed “balance.” That is, if their expert assessment of risk markers of suicidal ideation aligned on pairs of posts in the same propensity strata, then we would infer the treatment and control groups for that particular linguistic feature and strata to be balanced. If not, we would assume that the groups in that strata are not comparable to each other and that our propensity score matching analysis needs further tuning to identify more accurately balanced treatment and control

Treatment Token	Count	Coverage	Treatment Effect	z	χ^2
Increased Change					
depression	318	0.901	0.3	8.04	7.78
useless	53	0.801	0.51	7.05	6.53
suicide	143	1	0.32	6.66	5.03
anxiety	216	1	0.24	6.56	4.11
suicidal	111	0.9	0.34	6.56	5.37
i_almost	40	0.901	0.52	6.44	4.22
and_an	45	0.7	0.51	6.4	6.15
medicine	41	0.8	0.52	6.38	4.86
unless_i	38	0.9	0.53	6.36	4.47
hug	42	0.8	0.52	6.36	4.9
money_i	35	0.801	0.52	5.89	3.96
out_as	34	0.701	0.53	5.89	4.76
this_happened	35	0.901	0.51	5.89	3.72
this_world	37	0.8	0.5	5.88	4.17
over_i	35	0.901	0.51	5.86	3.58
still_a	36	0.7	0.51	5.85	4.68
off_a	35	0.801	0.51	5.85	4.24
loneliness	37	0.8	0.5	5.84	3.99
class_and	34	0.901	0.52	5.84	3.39
alone_i	77	1	0.34	5.84	3.91
Decreased Change					
captain	11	0.4	-0.6	-4	4.24
differences	16	0.601	-0.57	-4.47	3.56
the_trip	11	0.601	-0.57	-3.76	3.2
intimate	11	0.501	-0.57	-3.73	2.93
to_in	20	0.701	-0.56	-4.92	4.1
too hard	16	0.601	-0.56	-4.4	3.56
suspect	16	0.701	-0.56	-4.4	3.04
always a	14	0.601	-0.56	-4.15	3.29
be_working	14	0.601	-0.56	-4.12	2.73
keep your	12	0.601	-0.56	-3.82	2.46
straight up	12	0.601	-0.56	-3.82	2.38
preferred	11	0.601	-0.56	-3.71	2.43
awesome_i	11	0.501	-0.56	-3.68	2.86
s_at	21	0.801	-0.55	-4.83	3.33
stated	20	0.801	-0.55	-4.8	3.66
slight	18	0.701	-0.55	-4.61	3.3
and_enjoy	17	0.601	-0.55	-4.44	3.48
gotten_to	16	0.7	-0.55	-4.35	2.77
it_work	15	0.501	-0.55	-4.22	4.17
came_from	15	0.701	-0.55	-4.21	2.76

Table 1. Linguistic Cues.

Statistically significant treatment tokens obtained via propensity score matching that contribute to increased as well as decreased change in likelihood of posting in SW.

Token	Strata	Treatment Post	Control Post
High Propensity Strata			
not easy	6	a reason behind my depression is how small by body frame is. i've never cared much about muscle but it's obviously one of the reasons i've been alone (friendships and relationships) for my whole life.	i'm aware there's no way to avoid pain 100%, which is why i'm attempting to go for the least painful way. we've talked in detail about exactly why our issues are troubling for each of us, so he knows that already
advice but	6	i don't even know what all i feel. ashamed, angry, at myself and at the family that never did a thing to support me before. i'm seriously thinking about just pulling out i'm tired of trying, and failing, over and over again.	feeling like shit but noone to talk to, just need a friend who can cheer me up. noones online on facebook that i can talk to so just alone right now …
Low Propensity Strata			
seek	2	i realize that i'm having depression. i have not showered for a week now, unable to sleep and always thinking negative about myself	i noticed during the livestream, even though that he wasn't using their (i'm assuming) condenser microphone, i felt that his volume and the tones of his voice sounded much more "comfortable" with the headset.
slow down	1	an american football fan but i am intrigued by the world cup. i remember watching 4 years ago and was fascinated. does anyone know of a quality app i can get on my phone that i can use to keep up with it?	greetings people, greetings people, i am a worthless nobody. i guess i want to take more of your time in the vain hopes that you'll somehow be able to make me feel better.

Table 2. Qualitatively Assessed Post Pairs and Associated Comment Tokens.

Post pairs and associated comment tokens qualitatively assessed to correspond to balanced and imbalanced treatment and control groups. Text has been slightly paraphrased to protect the identities of the users.

groups. Once the unbalanced strata were identified by the human raters, then we filtered the posts to the corresponding comparable subpopulations. We then modified our method to compute a local average treatment effect only over the strata deemed to be balanced by human assessments, so as to assess the effects of specific linguistic features of comments in future risk to suicidal ideation

With the help of these human assessments and as shown in table 2, we found that the effects of getting a token in a comment may not be homogeneous. Certain users may see little effect of getting a token (low-propensity strata), while others see a large effect (higher-propensity strata). By employing human raters in this task, we showed how the outputs of causal inference methods can be amalgamated with expert feedback to improve results.

The fact that we obtain better results by incorporating human feedback in an AI task like the one previously described is further clarified while investigating the context of use of specific linguistic tokens in comments, and situating those tokens in theoretical framework of social support. In the users who show reduced likelihood of suicidal ideation in our dataset, we found comments on their posts to contain greater expression of esteem (31 percent) and network support (23 percent), followed by emotional support (16 percent). Informational support (9 percent) and acknowledgments (5 percent) were relatively lower for comments containing tokens that decrease the likelihood of posting about suicidal thoughts. Overall, this distribution indicates the positive impact of esteem and network support in reducing one's future risk of suicidal ideation expression, a result which can be accurately inferred by filtering the outcomes of causal inference based on human feedback.

Implications

Our research shows that, while AI approaches have made and continue to make significant strides into domains like mental health, the involvement of natural intelligence in the form of human feedback is critical to the success of these efforts. Given the complexities and sensitivities of this domain, human insights and the integration of domain knowledge can situate the efforts in existing research, theory, and what is needed for further validation of insights with carefully designed experiments and empirical study designs. Importantly, for the same reasons of domain complexity and sensitivity, we caution against automatic deployment of the described AI approaches and emphasize that human involvement will help translate their potential benefits to real-world mental health context — similar arguments have been made earlier (Amershi et al. 2014) as well as more recently in domains outside of mental health. In the paragraphs that follow, we discuss some of these mixed-initiative, human-machine partnered implications of this research.

Clinical and Self-Reflection Interventions

The approaches that we discuss in this article can have widespread implications for the mental health clinician community. Currently, there is limited ability to aid chronic mental illness management (Simon and Ludman 2009). Patient-reported experiences in the form of clinical interviews and questionnaires have played a central role in management of these conditions for more than a century (Liberman 1988). These approaches do not include evidence-based assessments — behavioral, emotional, or cognitive symptoms must be recalled from a patient's memory — a method prone to retrospective recall bias. Time and budgetary constraints further limit psychiatrists from conducting more thorough and frequent in-person evaluations. These constraints preclude time-sensitive and objective monitoring of symptoms, and an ability to detect subtle and burgeoning changes that may not surface in patients' self-reports. With the human-machine mixed initiative approaches we presented here, technologies can be developed that allow clinicians to monitor patients' symptoms and identify patterns that may be harbingers of adverse health events in the future. This way, clinicians will be able to engage in evidence-based decision-making, beyond what is possible within the realms of in-person therapeutic settings. To reiterate, the involvement of humans, in this case stakeholders like the patient and the clinicians, is critical to ensure that the technologies function in a way that is accountable, interpretable, actionable, and transparent.

Interventions may also be designed, based on the human-machine approaches previously discussed, that promote self-reflection of one's (for example, a patient's) activity and behavior around mental health on various social media platforms. Our methods might be employed for the self-assessment of behavior, cognition, and affect, or might serve as an early warning mechanism for individuals struggling with mental health concerns. Reflective interventions, guided by an expert, such as a support network member or a clinician, could also be designed to reveal longitudinal trends relating to specific mental health markers, such as that of suicide ideation. Such an intervention might be used for instance, to identify time periods of anomalous patterns, which are known to be otherwise difficult for individuals to keep track of. The logging of these longitudinal trends can also serve as a diary-style data source to help caregivers or other trained professionals and clinicians gain a deeper understanding of an individual's risk for dangerous behaviors in the future.

Social Media Interventions

Individuals whose posted content contains phrases and other linguistic constructs relating to mental health risk, as revealed by our AI-based methods, may be flagged in the interfaces of moderators and other clinical experts for help and support, thereby involving a human "in the loop." Community moderators, support volunteers, and the social media platform creators and owners themselves may also be allowed to maintain a "risk list" in their interfaces that would include individuals forecasted by our AI methods to exhibit signs of increased risk in the future. Such a list may sort or rank individuals based on their forecasted risk score. This approach would allow improved preparedness on the part of the moderators, platform owners, and experts to bring timely and appropriate help to those in need. Further, on being informed that an individual could be prone to risk in the future, moderators and experts may make provisions to connect them with appropriate mental health resources (for example, a web-based hotline like Crisis Text Line, or a community like 7 Cups of Tea2), encouraging peers or trusted friends and family, or field private messages with relevant information on help seeking or therapy.

Finally, our work also includes implications for volunteers intending to provide social support to vulnerable groups on social media. Applications could be developed, leveraging the human-machine collaborative techniques we discussed, that continually educate such volunteers to be self-aware and learn about what kind of information is perceived to be beneficial to social media users seeking help and support around mental health challenges.

AI Implications

Our work also raises questions about the challenges that AI as a field faces in realizing these domain implications. Such questions largely pertain to how the AI tools are used and in what ways human intelligence can alleviate some of the challenges that arise in real-world deployment of these tools and approaches. We discuss two such aspects.

Guarding Against Errors and Negative Outcomes

Many AI approaches, including some of the ones described in this article, have been considered as inscrutable black boxes of decision-making (Horvitz 2017). Improvements in computational power coupled with the availability of large volumes of training data (such as from social media) — data used to train AI models that infer mental health state — are driving advancements in machine learning, which are reinforcing the black-box phenomenon. In fact, in the context of mental health, neural networks are becoming an increasingly popular way of making predictions of a variety of symptoms and attributes (Chancellor et al. 2017; Manikonda and De Choudhury 2017; Reece and Danforth 2017). Neural networks, or largely, deep learning, are so opaque that it is practically impossible to understand what they deduce from training data and how they reach their conclusions — making it hard to judge their correctness in a domain where accuracy is critical to human life. Thus, these advances in machine learning techniques are enabling the creation of black-box AI approaches that, although they have better predictive power, are significantly more complex, especially to a layperson like a patient or a clinician, and are also less interpretable or explainable, again to the same layperson, who is likely to benefit the most from their outcomes.

Black boxes are also vulnerable to risks, such as accidental or intentional biases, errors, and frauds, thus raising the question of how to "trust" these systems and tools that make important and sensitive inferences about an individual's mental health state. Incorrect interpretation of the output of these systems (for example, what does mental health risk really mean), inappropriate use of the output (for example, using them directly in diagnosis or treatment), and disregard of the underlying assumptions (for example, that every individual is different, and so is their social media use and mental health state) can have drastic consequences. Involving humans can help correct some of these biases, and a "human" face to AI systems that make predictions about a person's behavior and mental health is likely to be more trustworthy to stakeholders.

Privacy and Ethics

Our work also raises important questions relating to privacy and ethics, questions that pose vexing complexities to the variety of stakeholders who are likely to be impacted by this research. Again, the involvement of experts and other individuals "in the loop" or toward the general functioning of AI systems will be helpful in tackling and addressing some of these challenges.

Mental Health Counselors and Clinicians. While our work provides new opportunities for mental health clinicians and counselors to learn what factors and attributes might precipitate risk for states such as depression and suicidal ideation, it also raises important ethical obligations. In a typical therapeutic setting, a clinician has control of the information that is sought, gathered, and used. The inferences and assessments about a patient's mental health are also made by the clinician themselves, by incorporating their understanding of a patient's state as well as other types of relevant collateral information. However, when these inferences are made by an algorithm, what should be the clinician's response, how should they act on this information? How can they navigate the therapeutic relationship with a patient, in the face of information delivered through an AI tool, while respecting the patient's privacy needs and therapeutic expectations?

In essence, AI-based technology provides an unprecedented opportunity to engage people both outside traditional mental healthcare settings and far earlier in the course of illness (Baumel et al. 2018). Capitalizing on this opportunity, however, requires stakeholders like clinicians and counselors to challenge underlying assumptions about traditional pathways to mental health treatment and care. Further, AI approaches to identifying illness and tracking symptoms will need human feedback with respect to redefining existing clinical rules and regulations. Although the potential beneficial impact of AI technology integration could be transformative, new critical questions regarding clinical expectations and responsibilities will require resolution.

Social Media Platform Owners, Designers, Moderators, and Participants. The techniques we presented could allow moderators, support volunteers, and owners and designers of social media platforms to make improved decisions and choices based on forecasted likelihood of risk. However, when inferences and assessments are made by an AI system instead of solely a human, what are the obligations for the moderators, the volunteers, the platform creators, or the community as a whole when they discover an individual to be at a higher likelihood of risk for behavior such as suicidal ideation? How can social media sites reap the benefits of our method and gain from the design opportunities outlined, while at the same time protect their ethical obligation to act upon situations that may need an intervention? We also envision ethical questions regarding revealing to social media users the implications of the use of certain type of language or certain patterns of activity, or surfacing to them inferred risk measures. To this end, interventions would require careful consideration because there is a delicate line between overintrusiveness and concern (in AI terms, balancing false positive and false negative rates). As noted in our prior work, further research is needed to better define the trajectory between online activity and making first clinical contact to explore opportunities for digital intervention (Birnbaum et al. 2017). In fact, ethical challenges go beyond the uses of the outcomes of AI technology. Without appropriate human involvement, due to the Hawthrone effect (McCarney et al. 2007), stigma, or other self-censorship reasons, over time individuals may eventually refrain from offering cues that might reveal their risk. How can social media sites, then, continue to be platforms of authentic expression and a means that enable disclosure of deep-seated mental health concerns? How can AI tools leverage human feedback in ways that ameliorate these self-censorship challenges?

One way to tackle these challenges could be to thoroughly assess the acceptability of our method or the technologies it enables to different stakeholders, thus incorporating human feedback into the design and functioning of the AI systems. This strategy constitutes a promising direction for future research. Collaborations between AI researchers, mental health experts, community moderators, designers, developers, social media companies, and ethicists can also help develop protocols and guidelines that facilitate the use of our work in practical contexts in the future.

Conclusion

In this article, we presented a discussion of the role of human involvement in deriving meaningful value out of AI techniques and approaches. We highlighted work from several threads of our prior research to describe this agenda, particularly focusing on the domain of mental health.

To conclude, the underlying impetus for investigating these types of problems of societal significance with AI is, of course, the desire to help people improve their (here, mental health) outcomes, whether through early identification of people at risk, better personalization of treatments, or discovery of new treatment strategies. Bridging the gap between insights derived from AI approaches and real-world action will require combining the outcomes of the approaches with human feedback, interventions, and simultaneous human/empirical observations to provide strong validations of benefits. The challenges posed in moving from AI outcomes to intervention in social media platforms are particularly exacerbated in sensitive domains — for example, how to get informed consent from very large populations when it comes to mental health assessments or how to ensure interventions that avoid real-world harm while respecting privacy of individuals online. It will be a significant challenge to develop new protocols that safely translate insights from observational studies of AI methods/tools, to active experimentation involving expert feedback, and then to large-scale deployments involving real people, while simultaneously respecting principles of individual autonomy, minimizing risk of harm, and ensuring that benefits and risks are distributed across all parties who are directly or indirectly, positively or less beneficially, affected by the underlying AI.

References

Amershi, S.; Cakmak, M.; Knox, W. B.; and Kulesza, T. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35(4): 105–20. doi.org/10.1609/aimag.v35i4.2513.

Baumel, A.; Baker, J.; Birnbaum, M. L.; Christensen, H.; De Choudhury, M.; Mohr, D. C.; Muench, F.; Schlosser, D.; Titov, N.; and Kane, J. M. 2018. Summary of Key Issues Raised in the Technology for Early Awareness of Addiction and Mental Illness (TEAAM-I) Meeting. *Psychiatric Services* 69(5): 590-92. doi.org/10.1176/appi.ps.201700270.

Beck, A. T. 1979. *Cognitive Therapy of Depression.* New York: Guilford Press.

Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research* 19(8): e289. doi.org/10.2196/jmir.7956.

Bonner, R. L., and Rich, A. 1988. Negative Life Stress, Social Problem-Solving Self-Appraisal, and Hopelessness: Implications for Suicide Research. *Cognitive Therapy and Research* 12(6): 549–56. doi.org/10.1007/BF01205009.

Boyd, D., and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, and Society* 15(5): 662–679. doi.org/10.1080/1369118X.2012.678878.

Caliendo, M., and Kopeinig, S. 2008. Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys* 22(1): 31–72. doi.org/10.1111/j.1467-6419.2007.00527.x.

Chancellor, S.; Kalantidis, Y.; Pater, J. A.; De Choudhury, M.; and Shamma, D. A. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 3213–3226. New York: Association for Computing Machinery. doi.org/10.1145/3025453.3025985.

Chancellor, S.; Lin, Z. J.; Goodman, E.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work, and Social Computing*, 626–38. New York: Association for Computing Machinery. doi.org/10.1145/2818048.2819973.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the Fifth Annual ACM Web Science Conference,* 47–56. New York: Association for Computing Machinery. doi.org/10.1145/2464464.2464480.

De Choudhury, M.; Counts, S.; Horvitz, E.; and Hoff, A. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing,* 626-38. New York: Association for Computing Machinery. doi.org/10.1145/2531602.2531675.

De Choudhury, M., and De, S. 2014. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In *Proceedings of the Eighth International Confer-*

ence on Weblogs and Social Media, 71–80. Palo Alto, CA: AAAI Press.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting Depression via Social Media. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*, 128–37. Palo Alto, CA: AAAI Press.

De Choudhury, M., and Kiciman, E. 2017. The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk. In *Proceedings of the 11th International Conference on Web and Social Media*, 32–41. Palo Alto, CA: AAAI Press.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 2098–110. New York: Association for Computing Machinery. doi.org/10.1145/2858036.2858207.

Dieserud, G.; Røysamb, E.; Ekeberg, Ø.; and Kraft, P. 2001. Toward an Integrative Model of Suicide Attempt: A Cognitive Psychological Approach. *Suicide and Life-Threatening Behavior* 31(2): 153–68. doi.org/10.1521/suli.31.2.153.21511.

Dietterich, T. G. 2009. Machine Learning in Ecosystem Informatics and Sustainability. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 8–13. Palo Alto, CA: AAAI Press.

Eaton, W. W.; Smith, C.; Ybarra, M.; Muntaner, C.; and Tien, A. 2004. Center for Epidemiologic Studies Depression Scale: Review and Revision (CESD and CESD-R). In *Instruments for Adults*. Vol. 3 of *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment,* 3rd ed., edited by M. E. Maruish, 363–77. New York: Routledge.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. In *Proceedings of the ACM Human Computer Interaction* 1(2): article 41. New York: Association for Computing Machinery. doi.org/10.1145/3134678.

Glanz, L. M.; Haas, G. L.; and Sweeney, J. A. 1995. Assessment of Hopelessness in Suicidal Patients. *Clinical Psychology Review* 15(1): 49–64. doi.org/10.1016/0272-7358(94)00040-9.

He, J.; Bailey, J.; Rubinstein, B. I.; and Zhang, R. 2015. Identifying At-Risk Students in Massive Open Online Courses. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence,* 1749–55. Palo Alto, CA: AAAI Press.

Horvitz, E. 2017. AI, People, and Society. *Science* 357(6346): 7. doi.org/10.1126/science.aao2466.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science* 353(6301): 790–94. doi.org/10.1126/science.aaf7894.

Kashden, J.; Fremouw, W. J.; Callahan, T. S.; and Franzen, M. D. 1993. Impulsivity in Suicidal and Nonsuicidal Adolescents. *Journal of Abnormal Child Psychology* 21(3): 339–53. doi.org/10.1007/BF00917538.

Liberman, R. P. 1988. *Psychiatric Rehabilitation of Chronic Mental Patients*. Washington, DC: American Psychiatric Press. doi.org/10.1176/ps.39.8.893.

Manikonda, L., and De Choudhury, M. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 170–81. New York: Association for Computing Machinery. doi.org/10.1145/3025453.3025932.

McCarney, R.; Warner, J.; Iliffe, S.; Van Haselen, R.; Griffin, M.; and Fisher, P. 2007. The Hawthorne Effect: A Randomised, Controlled Trial. *BMC Medical Research Methodology* 7(1): 30. doi.org/10.1186/1471-2288-7-30.

Pearl, J. 2009. Causal Inference in Statistics: An Overview. *Statistics Surveys* 3: 96–146. doi.org/10.1214/09-SS057.

Pearl, J. 2018. Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution. arXiv preprint arXiv:1801.04016 [cs.LG]. Ithaca, NY: Cornell University Press.

Reece, A. G., and Danforth, C. M. 2017. Instagram Photos Reveal Predictive Markers of Depression. *EPJ Data Science* 6(1): 15. doi.org/10.1140/epjds/s13688-017-0110-z.

Rosenbaum, P. R., and Rubin, D. B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1): 41–55. doi.org/10.1093/biomet/70.1.41.

Rudd, M. D. 1990. An Integrative Model of Suicidal Ideation. *Suicide and Life-Threatening Behavior* 20(1): 16–30.

Schwarzer, R., and Fuchs, R. 1995. Changing Risk Behaviors and Adopting Health Behaviors: The Role of Self-Efficacy Beliefs. In *Self-Efficacy in Changing Societies,* edited by A. Bandura, 259–88. Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9780511527692.

Simon, G. E., and Ludman, E. J. 2009. It's Time for Disruptive Innovation in Psychotherapy. *The Lancet* 374(9690): 594–95. doi.org/10.1016/S0140-6736(09)61415-X.

Vasisht, D.; Kapetanovic, Z.; Won, J.; Jin, X.; Chandra, R.; Sinha, S. N.; Kapoor, A.; Sudarshan, M.; and Stratman, S. 2017. Farmbeats: An IOT Platform for Data-Driven Agriculture. In *Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation,* 515–29. Berkeley, CA: Advanced Computing Systems Association.

Munmun De Choudhury is an assistant professor in the School of Interactive Computing at the Georgia Institute of Technology where she directs the Social Dynamics and Wellbeing Lab. Prior to joining Georgia Tech, De Choudhury was a faculty associate with the Berkman Klein Center for Internet and Society at Harvard and a postdoc at Microsoft Research, following obtaining her PhD in computer science from Arizona State University. De Choudhury's research interests are in computational social science. With her students and collaborators, De Choudhury focuses on developing computational methods to assess, understand, and improve personal and societal mental health from online social interactions.

Emre Kiciman is a principal researcher at Microsoft Research. Kiciman's research interests include social computing and computational social science, causal inference methods, and information retrieval. His current work focuses on causal analysis of large-scale social media timelines, using social data to support individuals and policymakers across a variety of domains, and more broadly on the implications of AI on people and society. Kiciman's past research includes entity-linking methods for social media and the web, deployed in the Bing search engine; and foundational work on applying machine learning to fault management in large-scale internet services.