

# Natural Language Processing for Mental Health

Dana Atzil-Slonim<sup>\*1</sup>, Iryna Gurevych<sup>\*2</sup>, Dirk Hovy<sup>\*3</sup>, and  
Diyi Yang<sup>\*4</sup>

- 1 Bar-Ilan University – Ramat-Gan, IL. [dana.slonim@gmail.com](mailto:dana.slonim@gmail.com)
- 2 Department of Computer Science, TU Darmstadt, DE.  
[iryna.gurevych@tu-darmstadt.de](mailto:iryna.gurevych@tu-darmstadt.de)
- 3 Bocconi University – Milan, IT. [dirk.hovy@unibocconi.it](mailto:dirk.hovy@unibocconi.it)
- 4 Stanford University, US. [diyi@stanford.edu](mailto:diyi@stanford.edu)

---

## Abstract

NLP has made remarkable progress in recent years, driven by breakthroughs in large language models (LLMs) and the availability of large-scale datasets such as social media posts, online forums, and patient records. These advances have made NLP models highly capable of extracting valuable insights from text data related to mental health. This development raises two natural questions: (1) How well, if at all, can NLP enable early detection, diagnosis, and intervention – not only for patients or support seekers but also for therapists or support providers? (2) Can NLP-driven solutions help bridge the gap between the escalating demand for mental health resources and the limited availability of mental health professionals, providing scalable and immediate support through chatbots, virtual therapists, and data-driven interventions? Both questions address the technical feasibility and the ethical concerns about using a developing technology in a sensitive application. This Dagstuhl Seminar brought together researchers across NLP, clinical science, human–computer interaction, and digital mental health to reflect on how NLP and AI can support mental health outcomes. Over the course of the week, we looked at key areas where NLP has the potential to transform mental health: understanding how mental states change and how therapeutic change occurs; exploring how NLP can help therapist training and feedback; identifying technological gaps and multilingual challenges in building reliable mental health models; and addressing pressing concerns around evaluation, validation, privacy, and ethics. Through vision talks, lightning sessions, and breakout groups, participants explored both the opportunities and limitations of deploying NLP for mental health, laying the groundwork for responsible, interdisciplinary research in this vital direction.

**Seminar** August 31 – September 5, 2025 – <https://www.dagstuhl.de/25361>

**2012 ACM Subject Classification** Computing methodologies → Natural language processing;  
Applied computing → Health care information systems; Human-centered computing →  
Empirical studies in HCI

**Keywords and phrases** Mental Health, NLP, Human-Centered AI, Large Language Models

**Digital Object Identifier** 10.4230/DagRep.15.8.62

---

\* Editor / Organizer



Except where otherwise noted, content of this report is licensed  
under a Creative Commons BY 4.0 International license

Natural Language Processing for Mental Health, *Dagstuhl Reports*, Vol. 15, Issue 8, pp. 62–79

Editors: Dana Atzil-Slonim, Iryna Gurevych, Dirk Hovy, and Diyi Yang



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary

*Dana Atzil-Slonim (Bar-Ilan University – Ramat-Gan, IL, [dana.slonim@gmail.com](mailto:dana.slonim@gmail.com))*

*Iryna Gurevych (TU Darmstadt, DE, [iryana.gurevych@tu-darmstadt.de](mailto:iryana.gurevych@tu-darmstadt.de))*

*Dirk Hovy (Bocconi University – Milan, IT, [dirk.hovy@unibocconi.it](mailto:dirk.hovy@unibocconi.it))*

*Diyi Yang (Stanford University, US, [diyi@stanford.edu](mailto:diyi@stanford.edu))*

**License** © Creative Commons BY 4.0 International license  
© Dana Atzil-Slonim, Iryna Gurevych, Dirk Hovy, and Diyi Yang

This document summarizes the outcomes of our Dagstuhl Seminar on “Natural Language Processing for Mental Health” (25361). The seminar was motivated by an urgent and growing global need: mental health issues are affecting millions of people worldwide, yet the majority of individuals in need of care do not receive any treatment, especially those from marginalized, low-income, or rural populations. While many mental health conditions are treatable, or even preventable with early detection and intervention, people often don’t receive support until their concerns have escalated. In parallel, NLP has made remarkable progress in recent years, largely due to advances in large language models (LLMs) and the availability of large-scale textual data from sources like social media, online forums, and clinical records. These technologies therefore offer a variety of opportunities in addressing the problems outlined above, both for patients and in therapist training. Indeed, NLP models are already being deployed in mental health applications. However, both the existing and envisioned use cases raise critical questions around feasibility, scalability, and responsibility.

To address these, the seminar brought together an interdisciplinary group of researchers from NLP, clinical psychology, HCI, and digital health to assess where, how, and under what conditions NLP can be responsibly used to support mental health needs. Concretely, the seminar focused on:

- Understanding the potential of NLP, and in particular LLMs, to support mental health diagnosis, intervention, and therapeutic processes;
- Identifying critical gaps in technical feasibility, evaluation, and deployment of these tools in real-world, high-stakes clinical settings;
- Exploring responsible and interdisciplinary solutions that bridge NLP, psychology, human-computer interaction, and ethics.

As a major result from the seminar, we identified the following problems and future directions:

1. The need for systematic evaluation frameworks for NLP models used in psychotherapy and mental health support, including benchmarks, as well as longitudinal, multilingual, and real-world evaluation protocols.
2. The challenge of simulating clients or patients, and therapists using LLMs to support therapist training, research, and practice, particularly concerning modeling therapeutic authenticity, diversity, and change processes.
3. The persistent knowledge and communication gap between NLP and clinical psychology communities, and the urgent need to bridge this divide through interdisciplinary collaboration to ensure clinical relevance and real-world utility.
4. The lack of visibility and impact pathways for research in NLP and mental health across both technical and applied domains, and how to promote this work across venues, funding agencies, and policy-making spaces.

The seminar’s structure was designed to support both critical discussion and creative collaboration. Through a program of vision talks, lightning presentations, breakout groups, and informal exchanges, the seminar was organized around four thematic clusters: (1)

understanding how mental states change and how therapeutic change occurs; (2) how NLP can support therapist training and real-time feedback; (3) identifying technological, privacy, and multilingual challenges; and (4) addressing evaluation, validation, and ethical concerns.

This seminar has laid a solid foundation for a crucial research area. **Three perspective papers** are in development: one focused on simulating patients using LLMs, one on evaluation challenges and opportunities in psychotherapy applications, and one on strengthening interdisciplinary collaboration between NLP and mental health communities. Additionally, **a blog post** is being prepared to reflect on how to promote the broader impact of this work, and **a workshop submission** motivated by the seminar is currently under review for SIGCHI 2026.

These outcomes align with our initial goals: (1) to produce joint research publications and collaboration opportunities, such as position papers that map out the challenges and opportunities in building responsible and robust NLP systems for mental health; and (2) to form a cross-field community that continues to connect NLP and mental health researchers – both in technical venues and clinical practice contexts.

## 2 Table of Contents

### Executive Summary

*Dana Atzil-Slonim, Iryna Gurevych, Dirk Hovy, and Diyi Yang* . . . . . 63

### Overview of Talks

|                                                                                                                                                                 |    |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Integrating Innovations in Clinical Science and Artificial Intelligence to Study the Dynamics of Therapeutic Change<br><i>Dana Atzil-Slonim</i> . . . . .       | 67 |
| A (Humane) Vision for Digital Mental Health in a Post-AI World<br><i>Munmun De Choudhury</i> . . . . .                                                          | 67 |
| The Construct Mining Pipeline – a Computational Method to Reveal Psychological Constructs from Text Data<br><i>Jana Lasser</i> . . . . .                        | 68 |
| Can Language-based Assessments Outperform the Instruments on which They are Trained?<br><i>H. Andrew Schwartz</i> . . . . .                                     | 68 |
| Progress and challenges in NLP for mental health: Personalised longitudinal monitoring and beyond<br><i>Maria Liakata</i> . . . . .                             | 69 |
| Using AI in Measurement-Based and Data-Informed Psychological Therapy<br><i>Wolfgang Lutz</i> . . . . .                                                         | 69 |
| Using Large Language Models to Create Personalized Networks From Therapy Sessions<br><i>Hiba Arnaout</i> . . . . .                                              | 70 |
| Monitoring Patient Emotions at Scale to Assess Psychotherapy Outcomes Using Language Models<br><i>Matteo Malgaroli</i> . . . . .                                | 70 |
| Insights from Human-centered & HCI approaches for NLP/AI for Mental Health: Some Provocations<br><i>Stevie Chancellor</i> . . . . .                             | 70 |
| From Physics to Psychiatry: Dynamical Systems, Language, and Control<br><i>Hamidreza Jamalabadi</i> . . . . .                                                   | 71 |
| Supporting Mental-Health Communication: Towards a Proactive AI Support for (Human) Therapists<br><i>Cristian Danescu-Niculescu-Mizil</i> . . . . .              | 71 |
| Catching Disengagement Early: Development and Validation of an LLM Rating Scale for Client Engagement in Psychotherapy<br><i>Steffen T. Eberhardt</i> . . . . . | 72 |
| CARE: Training Counselors via LLMs<br><i>Ryan Louie</i> . . . . .                                                                                               | 72 |
| Helping the Helper: How AI Can Support Training of Peers in Delivery of Behavioral Health Care<br><i>Daniel Blonigen</i> . . . . .                              | 73 |

|                                                                                                                                                                              |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Designing Technologies for Digital Mental Health: an HCI Perspective<br><i>Gavin Doherty</i> . . . . .                                                                       | 74 |
| Practices of NLP for Mental Health in China<br><i>Minlie Huang</i> . . . . .                                                                                                 | 74 |
| Culture, Personalization and Mental Health<br><i>Monojit Choudhury</i> . . . . .                                                                                             | 74 |
| Artificial Intelligence, Affective Computing, and Health: Opportunities and Ethical<br>Considerations in Real-World Data Collections<br><i>Emily Mower Provost</i> . . . . . | 75 |
| Dear ChatGPT, Can You Keep My Secret? Privacy and Security in the Era of<br>LLMs<br><i>Anmol Goel</i> . . . . .                                                              | 75 |
| Computational Paralinguistics – In a Nutshell<br><i>Andreas Triantafyllopoulos</i> . . . . .                                                                                 | 76 |
| <b>Working Groups</b>                                                                                                                                                        |    |
| Responsible Evaluation of AI for Mental Health Systems . . . . .                                                                                                             | 76 |
| Simulating AI Patients for Psychotherapy: Challenges and Opportunities . . . . .                                                                                             | 76 |
| Identifying Intrapersonal and Interpersonal Dynamics Predictive of Change in<br>Mental Health . . . . .                                                                      | 77 |
| <b>Conclusion and Outlook</b> . . . . .                                                                                                                                      | 77 |
| <b>Participants</b> . . . . .                                                                                                                                                | 79 |

### 3 Overview of Talks

#### 3.1 Integrating Innovations in Clinical Science and Artificial Intelligence to Study the Dynamics of Therapeutic Change

*Dana Atzil-Slonim (Bar-Ilan University – Ramat-Gan, IL)*

License  Creative Commons BY 4.0 International license  
© Dana Atzil-Slonim

In psychological therapies, understanding what works for whom and when remains one of the most enduring challenges in mental health research and practice. Today, however, we are better equipped than ever to address this question – thanks to transformative advances in both clinical science and AI. In this talk, I demonstrate how our interdisciplinary team of clinicians and AI researchers is bridging top-down, theory-driven approaches with bottom-up, data-driven AI methods to uncover the dynamics that drive therapeutic change.

I begin by outlining key challenges in conceptualizing, treating, and researching mental health. I then discuss significant theoretical shifts in clinical science – such as the transition from general treatment models to transtheoretical processes and from one-person to two-person psychology – that have paved the way for addressing these challenges. Building on these foundations, I illustrate how our team has leveraged theoretical innovations and advancements in multimodal analysis and AI to explore the dynamic processes within clients (intrapersonal dynamics) and between clients and therapists (interpersonal dynamics) that are linked to improved treatment outcomes.

Central to this effort is the development of temporally aware, multimodal AI methods designed to address key limitations in modeling temporality, complex reasoning, situational awareness, personalization, and the integration of verbal and non-verbal data. I conclude by discussing how this integrative approach can enhance diagnostic precision, support personalized interventions, and improve the overall effectiveness of mental health treatments.

#### 3.2 A (Humane) Vision for Digital Mental Health in a Post-AI World

*Munmun De Choudhury (Georgia Institute of Technology – Atlanta, US)*

License  Creative Commons BY 4.0 International license  
© Munmun De Choudhury

Digital mental health has undergone three paradigm shifts: from the clinical gaze, to the quantified self, to predictive models built on social media and pervasive data. Today, we are on the cusp of a fourth shift – one defined by the rapid rise of large language models (LLMs) and generative AI. These technologies hold promise to expand access to support, reframe cognition, and facilitate empathic conversations. Yet, they also risk cultural misalignment, shallow validation, and the erasure of lived experience. Drawing on empirical studies spanning algorithmic prediction, online support, therapeutic alliance with AI, and cross-lingual evaluations of LLMs, this talk surfaces the tensions between correctness and care, agency and automation, identity and institutional power. I argue that the prevailing focus on efficiency and scale often neglects the ecological realities of people's lives and the invisible labor they contribute when their data fuel AI systems. A humane vision for digital mental health in a post-AI world requires moving from inference to interventions that center authenticity, inclusivity, and responsibility. I conclude by outlining design principles for an AI that connects without co-opting, empowers rather than replaces, and offers truth with care – not just comfort.

### 3.3 The Construct Mining Pipeline – a Computational Method to Reveal Psychological Constructs from Text Data

*Jana Lasser (University of Graz, AT)*

**License**  Creative Commons BY 4.0 International license  
© Jana Lasser

When starting to formalize psychological constructs, researchers traditionally rely on two distinct approaches: the quantitative approach, which defines constructs as part of a testable theory based on prior research and domain knowledge, often using self-report questionnaires, or the qualitative approach, which gathers data mainly in the form of text and bases construct definitions on exploratory analyses. We present a new computational method that combines the comprehensiveness of qualitative research and the scalability of quantitative analyses to define psychological constructs from semi-structured text data. Using structured questions, participants are prompted to generate sentences that reflect instances of the construct of interest. We apply computational methods to calculate embeddings as numerical representations of the sentences, which we then run through a clustering algorithm to arrive at groupings of sentences as psychologically relevant classes. The method includes steps for measuring and correcting bias introduced by data generation and for assessing cluster validity through human judgment. We demonstrate the applicability of our method on an example from emotion regulation. Based on short descriptions of emotion regulation attempts collected through an open-ended situational judgment test, we use our method to derive classes of emotion regulation strategies. Our approach shows how machine learning and psychology can be combined to provide new perspectives on the conceptualization of psychological processes.

### 3.4 Can Language-based Assessments Outperform the Instruments on which They are Trained?

*H. Andrew Schwartz (Stony Brook University, US)*

**License**  Creative Commons BY 4.0 International license  
© H. Andrew Schwartz

**Joint work of** H. Andrew Schwartz, Brenda Curtis, Salvatore Giorgi, Lyle Ungar, Huy Vu, David Yaden, Tingting Liu, Kenna Yadeta, Gregory Park, Johannes Eichstaedt, Evelyn Bromet, Benjamin Luft, Roman Kotov, Sean Clouston, Youngseo Son, Martin Seligman, Varadarajan Varadarajan

This two-part talk begins with a vision for HLAB focusing on (a) a more accurate representation of people, (b) LLMs for safe mental health therapy, and (c) eudaimonix content recommendation from AI. It continues with a series of experiments addressing the assumption that models trained on standard instruments are bounded by their predictive validity. We demonstrate cases where language-based assessments (LBAs) predict psychological outcomes better than the instruments they were trained on. Evidence includes: a historical review, a simulated experiment, an experiment with pseudo-observed data, and another with fully observed data predicting external criteria. These findings provide theoretical and empirical evidence that language-based assessments can more closely approximate true psychological states. Mechanisms by which these assessments outperform traditional tools are explored, highlighting the potential for AI-based language analysis to reshape psychological measurement.

### 3.5 Progress and challenges in NLP for mental health: Personalised longitudinal monitoring and beyond

*Maria Liakata (Queen Mary University of London, GB)*

License  Creative Commons BY 4.0 International license  
© Maria Liakata

The first part of the talk provides an overview of the NLP landscape for mental health, discussing the range of applications and the challenges faced by systems based on large language models (LLMs), with examples from the literature, especially regarding how these factors impact their suitability and applicability to mental health. There are many unresolved challenges, among others, regarding appropriate generation, temporal robustness, temporal and other forms of reasoning, and privacy concerns, especially when working with sensitive content such as mental health data. The programme of work I have been leading in the past five years consists of three core research directions: (1) data representation and generation, (2) methods for personalised longitudinal models and temporal understanding, (3) evaluation in real-world settings, with use cases in mental health. I will give an overview of the work in my group on these topics and conclude with a presentation of the evaluation platform for LLM-based systems that we have been developing within the AdSoLve project.

### 3.6 Using AI in Measurement-Based and Data-Informed Psychological Therapy

*Wolfgang Lutz (Trier University, DE)*

License  Creative Commons BY 4.0 International license  
© Wolfgang Lutz

This presentation discusses a dynamic network model and research program supporting the delivery of personalized feedback to therapists at the start and throughout psychological therapy. The approach focuses on identifying the core individual elements of psychological distress and resources, while also enabling the quantification of multiple dimensions of distress (e.g., cognitive-behavioral, emotional, motivational, and interpersonal stressors) through the application of large language models. It addresses questions such as: How can psychological distress and resources be extracted, both qualitatively and quantitatively, from session transcripts, and how can they be integrated into clinical support tools and adaptive treatment planning for therapists? How can such models support the identification of patients at risk of treatment failure? And how can these models support clinical training? What are the technical and practical challenges that need to be addressed?

### 3.7 Using Large Language Models to Create Personalized Networks From Therapy Sessions

*Hiba Arnaout (TU Darmstadt, DE)*

License  Creative Commons BY 4.0 International license  
© Hiba Arnaout

Personalizing psychotherapy often relies on individual-level networks, but estimating these networks typically requires intensive longitudinal data, limiting scalability. Large Language Models (LLMs) offer a potential alternative by analyzing therapy transcripts directly. We introduce a pipeline that automatically generates client networks to support case conceptualization and treatment planning.

### 3.8 Monitoring Patient Emotions at Scale to Assess Psychotherapy Outcomes Using Language Models

*Matteo Malgaroli (NYU Grossman School of Medicine – New York, US)*

License  Creative Commons BY 4.0 International license  
© Matteo Malgaroli

Traditional methods of psychiatric evaluation face ongoing challenges regarding reliability, objectivity, and scalability. The integration of language models offers a potential solution for assessing psychiatric symptoms and informing theory through digital biomarkers. In this talk, I will discuss findings on using linguistic markers to capture mental health symptoms from clinical conversations. In particular, I will introduce VISTA, a scalable method for capturing temporal flows, and apply it to emotions expressed by a sample of over 10,000 patients receiving digital mental-health treatment. I show how the resulting clusters relate to clinical outcomes. These findings highlight the opportunity to monitor patient outcomes using only linguistically captured emotions, especially when direct measurement of mental-health outcomes is infeasible.

### 3.9 Insights from Human-centered & HCI approaches for NLP/AI for Mental Health: Some Provocations

*Stevie Chancellor (University of Minnesota – Minneapolis, US)*

License  Creative Commons BY 4.0 International license  
© Stevie Chancellor

I will discuss three provocations for the field of AI/NLP for mental health, guided by insights from empirical work in HCI and human-centered AI. My approach to this problem combines my disciplinary training in Media Studies and Computer Science. To unpack the transformative potential of human-centered AI, we'll look at my group's work in mental illness and online social systems (social media and generative AI) as examples. The goal of this is to inspire conversation among participants, encourage solutions, and spark interdisciplinary reflection.

### 3.10 From Physics to Psychiatry: Dynamical Systems, Language, and Control

*Hamidreza Jamalabadi (Philipps-Universität Marburg, DE)*

License © Creative Commons BY 4.0 International license  
© Hamidreza Jamalabadi

Psychiatric interventions often focus on symptom alleviation, lacking the ability to fully capture the intricate dynamics of cognitive-affective states or tailor treatments to individual needs. From a dynamical systems perspective, though, the intervention can be observed in terms of an optimal control problem, where the ability to recover the dynamics based on continuous observation is key, as is the flexible implementation of control signals aimed at influencing cognitive-affective states – mirroring the severity and dynamics of mental disorders such as depression. Advances in natural language processing (NLP) and large language models (LLMs) offer transformative potential to address these limitations by enabling continuous, high-resolution monitoring of mental states through language, and further optimizing the processes of language-based therapies. This talk presents an integrative framework that combines physics-inspired dynamical systems theory with NLP. Language serves as a rich, temporally dynamic proxy for mental states, with LLMs facilitating the extraction of latent states and their temporal dynamics for predictive and interventional purposes. Optimized sampling rates, informed by ecological momentary assessment (EMA) and nonlinear modeling, enhance the detection of short- and long-term mood fluctuations. Furthermore, NLP-driven interventions, such as optimized psychotherapy and affective priming, can act as control inputs to reshape these trajectories, paralleling advances in neurostimulation that target pathological neural attractors. Early studies in our group show promising results in these directions, including those based on social media data (400,000 texts from more than 1,600 individuals over 6 years) and further experimental studies on optimizing interventions such as affective priming. This AI-informed neurocognitive dynamical systems framework paves the way for personalized interventions that stabilize resilient cognitive-affective trajectories, offering a paradigm shift from current approaches.

### 3.11 Supporting Mental-Health Communication: Towards a Proactive AI Support for (Human) Therapists

*Cristian Danescu-Niculescu-Mizil (Cornell University – Ithaca, US)*

License © Creative Commons BY 4.0 International license  
© Cristian Danescu-Niculescu-Mizil  
URL <https://convokit.cornell.edu>

Recent years have seen a gold rush to replace people with AI agents in communication: they can serve as your therapist, your tutor, your financial advisor, and your interviewer. In this talk, I propose a contrasting vision: one in which AI supports humans in their communication while preserving their agency. Achieving this vision requires moving beyond the current transactional paradigm embodied by current generative AI systems, which are designed to fulfill the immediate goals of a single person, such as answering a question, solving a math problem, booking a flight, or (repeatedly) replying in character. To meaningfully support human–human communication without disrupting or supplanting it, an AI system must instead follow a proactive paradigm: it needs to decide when to intervene to offer support

as the interaction unfolds, rather than wait to explicitly be prompted as AI agents and chatbots do today. In this talk, I present initial progress on AI technologies that enable such a proactive mode of operation and demonstrate them in the context of mental health crisis counseling. Data and code are available through ConvoKit.

### 3.12 Catching Disengagement Early: Development and Validation of an LLM Rating Scale for Client Engagement in Psychotherapy

*Steffen T. Eberhardt (Trier University, DE)*

**License**  Creative Commons BY 4.0 International license

© Steffen T. Eberhardt

**Main reference** Steffen T. Eberhardt, Antonia Vehlen, Jana Schaffrath, Brian Schwartz, Wolfgang Lutz, Tobias Baur, Dominik Schiller, Tobias Hallmen, Elisabeth André: “Development and validation of large language model rating scales for automatically transcribed psychological therapy sessions”, *Sci Rep* 15, 29541 (2025)

**URL** <https://doi.org/10.1038/s41598-025-14923-y>

Rating scales have shaped psychological research, but are resource-intensive and can burden participants. Large Language Models (LLMs) provide a tool for assessing latent constructs in text. This study introduces *LLM rating scales* that use LLM responses rather than human ratings. We demonstrate this approach using an LLM rating scale to measure patient engagement in therapy transcripts. Automatically transcribed videos of 1,131 sessions from 155 patients were analyzed using DISCOVER, a software framework for local multimodal human behavior analysis. Llama 3.1 8B rated 120 engagement items, averaging the top eight into a total score. Psychometric evaluation showed a normal distribution, strong reliability ( $\omega = 0.953$ ), and acceptable fit (CFI = 0.968, SRMR = 0.022), except RMSEA = 0.108. Validity was supported by significant correlations with engagement determinants (e.g., motivation,  $r = .413$ ), processes (e.g., between-session efforts,  $r = .390$ ), and outcomes (e.g., symptoms,  $r = -.304$ ). Results remained robust across bootstrap resampling and cross-validation, accounting for the nested data structure. The LLM rating scale exhibited strong psychometric properties, demonstrating the potential of the approach as an assessment tool. Importantly, this automated approach uses interpretable items, ensuring a clear understanding of measured constructs, while supporting local implementation and protecting confidential data.

### 3.13 CARE: Training Counselors via LLMs

*Ryan Louie (Stanford University, US)*

**License**  Creative Commons BY 4.0 International license

© Ryan Louie

The global mental health crisis demands innovative approaches to scale high-quality care. While AI chatbots for therapeutic use are on the rise, I argue that research should develop scalable solutions for contexts where human support remains essential. I will present our work developing CARE, a Large Language Model (LLM)-based training system that empowers human counselors through practice with AI-simulated patients and feedback from AI mentors. In this talk, I highlight two research thrusts: technical challenges in training counselors with LLMs and evaluating the impact of LLM-based training with novice counselors. Advancing

these research thrusts requires answering interdisciplinary questions at the intersection of natural language processing, mental health, and human AI interaction. How might we develop realistic LLM simulations of patients when privacy concerns restrict data access and domain-expert feedback is expensive? How can we mitigate the chance of generic or clinically-inappropriate LLM feedback? How can we conduct stage-appropriate user studies of LLM training systems that yield actionable insights to improve design? Our work advancing CARE contributes to responsible AI in mental health by developing new tools, algorithms, and empirical evidence for scaling counselor education via LLMs, thereby increasing the supply of well-trained human therapists to meet society's growing demands.

### 3.14 Helping the Helper: How AI Can Support Training of Peers in Delivery of Behavioral Health Care

*Daniel Blonigen (VA Palo Alto Health Care System, US & Stanford University, US)*

**License** © Creative Commons BY 4.0 International license  
© Daniel Blonigen

Globally, there is a demand–capacity problem in mental health care. One in four individuals has a behavioral health disorder (substance use and/or mental illness), yet more than half of all behavioral health providers report no openings for new patients, and many report burnout with existing caseloads. Peer Recovery Workers (PRWs) have been cited as critical to improving access and engagement in behavioral healthcare and mitigating staffing shortages. However, high rates of burnout and turnover are well-documented challenges to implementing PRWs, often stemming from inadequate training and supervision. Artificial Intelligence (AI) could be an ideal solution for ensuring high-quality training and ongoing supervision of PRWs in a strained healthcare system with limited resources. In this talk, we review how peer competencies include many of the interpersonal skills foundational to effective psychotherapy (e.g., rapport-building, active listening, reflecting, validating, empathizing), and the value of scaling peer training in these competencies using AI. CARE is an AI-powered web-based platform to train and empower individuals in basic counseling skills. In CARE, individuals select and/or create AI patients to roleplay emotional support conversations. Using a Large Language Model trained by senior psychotherapy supervisors, CARE provides counselors with feedback on their responses to the AI patients using a multi-level structure that mirrors how supervisors provide feedback to trainees. Although lab experiments have established CARE's proof-of-concept to improve the feedback quality provided to novice counselors, the program was not designed for PRWs whose scope of practice and training needs differ from those of traditional counselors (e.g., sharing lived experiences). Consequently, there may be a need to customize and pilot the use of CARE with PRWs. Mixed-method designs are suggested to collect systematic feedback from relevant stakeholders to guide customization of CARE for peers and then evaluate the feasibility and acceptability of the customized version with novice peers working in real-world clinical settings. This research program has the potential to address the demand–capacity problem by increasing the adoption of the PRW workforce in behavioral health settings and improving the quality of care they provide to patients.

### 3.15 Designing Technologies for Digital Mental Health: an HCI Perspective

*Gavin Doherty (Trinity College Dublin, IE)*

**License**  Creative Commons BY 4.0 International license  
© Gavin Doherty

In this talk, I present a Human–Computer Interaction perspective on the design of digital health interventions for mental health. Drawing on experience developing and evaluating a range of novel systems to support the delivery of mental health care, the talk considers the critical yet often ill-defined concepts of acceptability and engagement, before examining in more detail the HCI issues surrounding the use of machine learning technologies in this context, looking at a specific example relating to outcome prediction. The talk concludes with a brief consideration of more personalised interventions, including the use of LLM-based capabilities within a broader design space for Ecological Momentary Interventions, along with the associated design issues.

### 3.16 Practices of NLP for Mental Health in China

*Minlie Huang (Tsinghua University – Beijing, China)*

**License**  Creative Commons BY 4.0 International license  
© Minlie Huang  
**URL** <http://coai.cs.tsinghua.edu.cn/hml/>

In this talk, the speaker discusses his NLP practices for mental health applications in China. Specifically, he developed tools with LLMs for mental state assessment, built chatbots to provide effective emotional support, and developed LLM models for simulating clients and therapists for the purpose of training and evaluation. Some of these tools have been deployed in real-world applications.

### 3.17 Culture, Personalization and Mental Health

*Monojit Choudhury (Mohamed Bin Zayed University of Artificial Intelligence – Abu Dhabi, AE)*

**License**  Creative Commons BY 4.0 International license  
© Monojit Choudhury

Mental health conditions exhibit extreme behavioral variability. A stunning example is Autism Spectrum Disorder (ASD), a catchall term covering a wide range of symptoms and varying degrees of communication difficulties. Only recently have we begun to understand the types and causes of ASD, and specialized interventions that work for every individual are still a far-off dream. AI, particularly large language models (LLMs), offers a unique opportunity to continuously learn from a user's behavior, along with demographic information, patient history, and assessment results, which could serve as sandboxes for testing different (potentially novel) interventions before delivering the optimal one. In my talk, I discuss recent work that simulates user behavior in terms of what a user knows and does not know, and how to explain unfamiliar concepts in a lucid and personalized way from a cross-cultural communication perspective. These studies reveal that LLMs can learn to personalize, but

they also tend to produce stereotypical, less variable responses than real users. LLMs often agree with responses from other LLMs, but not as much with human users, who align more strongly with each other. This insight suggests that culture can serve as an excellent prior for personalization, but systems must also continuously learn from unfolding behavior to achieve the best results. Such methods provide a promising direction toward individual mental health–relevant behaviors for intervention sandboxing and innovation.

### 3.18 Artificial Intelligence, Affective Computing, and Health: Opportunities and Ethical Considerations in Real-World Data Collections

*Emily Mower Provost (University of Michigan – Ann Arbor, US)*

**License** © Creative Commons BY 4.0 International license  
© Emily Mower Provost

Emotions provide critical cues into our health and well-being. They are particularly important in the context of mental health, where changes in emotion may signify changes in symptom severity. However, information about emotion and its temporal variation is often accessible only through survey methodologies (e.g., ecological momentary assessment, EMA), which can become burdensome over time. Affective computing technologies, such as automated speech emotion recognition systems, could provide an alternative, namely offering quantitative measures of emotion using acoustic data captured passively from a consented individual’s environment. However, these technologies are not without risk and can pose a potential for harm. There are critical ethical issues that must be thoughtfully considered. In this talk, I discuss our journey in affective computing for health modeling, presenting the design of these technologies alongside the ethical considerations that have shaped their development.

### 3.19 Dear ChatGPT, Can You Keep My Secret? Privacy and Security in the Era of LLMs

*Anmol Goel (TU Darmstadt, DE & University of Copenhagen, DK)*

**License** © Creative Commons BY 4.0 International license  
© Anmol Goel

In this talk, I explore the critical privacy and security challenges of Large Language Models (LLMs). Users are increasingly turning to LLMs for sensitive and personal interactions, including advice, companionship, and counseling. This trend creates a “Personalization–Privacy Paradox,” in which the utility of personalized AI is in direct conflict with the need to protect user data.

The talk outlines concrete privacy threats, including membership inference, data leakage, and prompt extraction attacks. To combat these risks, a dual approach is proposed: Proactive Privacy (privacy by design) and Reactive Privacy (unlearning). I discuss recent results showing that differential privacy can reliably work for steering vectors, and that current unlearning evaluations are suboptimal – providing only a false sense of privacy. Finally, I present ongoing work on data poisoning and attribution methods for language models.

### 3.20 Computational Paralinguistics – In a Nutshell

*Andreas Triantafyllopoulos (Technical University of Munich, DE)*

License  Creative Commons BY 4.0 International license  
© Andreas Triantafyllopoulos

Speech is about more than content – it’s about what you say, when, and how. The latter two facets of spoken language – the “when” and the “how” – fall under the purview of computational paralinguistics. The use of *para* emphasizes contrast to classic linguistics. Paralinguistics concerns phenomena often neglected by traditional computational linguistics. In this talk, I provide a brief overview of the field, starting with motivational examples and a taxonomy of the relevant phenomena. I outline methodological approaches used in computational paralinguistics, from traditional feature-based analysis to modern foundation model-based methods that integrate speech and language processing. I also present a recent case study on using speech in the mental health context.

## 4 Working Groups

In addition to vision talks and breakout discussions, the seminar gave rise to three ongoing working groups. Each group brings together participants from NLP and clinical psychology to develop joint position or perspective papers that extend the conversations initiated at Dagstuhl into concrete research agendas and community guidelines. Together, these efforts reflect the seminar’s overarching themes: understanding dynamic mental health processes, supporting therapist training, and ensuring that AI systems for mental health are evaluated and deployed responsibly.

### 4.1 Responsible Evaluation of AI for Mental Health Systems

This working group, led by Iryna Gurevych (Technical University of Darmstadt, DE), focuses on how AI systems for mental health should be evaluated before they can be trusted in sensitive clinical and community settings. Building on analysis of recent NLP work on mental health, the group identified recurring gaps in current practice, including over-reliance on automated metrics, limited involvement of clinicians or people with lived experience, and insufficient attention to safety, equity, and context.

In response, the group is developing a taxonomy that distinguishes assessment, intervention, and support systems and links each aspect to appropriate evaluation dimensions. The resulting position paper will provide a shared, clinically grounded framework that allows AI researchers, clinicians, implementation scientists, and other stakeholders to speak a common evaluative language and move toward more reliable, interpretable, and socially responsible AI for mental health.

### 4.2 Simulating AI Patients for Psychotherapy: Challenges and Opportunities

A second working group, led by Diyi Yang (Stanford University, US), examines the emerging use of AI-simulated patients as training tools for psychotherapy and counseling. The group starts from the long tradition of actor-based roleplay in clinical training and asks what it

would take to extend these practices into scalable, AI-driven simulation environments. Key questions include how to balance conversational realism with pedagogical value and safety; how to represent long-term therapeutic trajectories versus short, skills-focused exercises; and how to ensure cultural, linguistic, and demographic diversity, including rare and edge-case presentations.

The group also considers data governance, risks of “overfitting” trainees or models to simulators, and methods for benchmarking educational utility across therapeutic orientations. The planned perspective paper will articulate design principles, evaluation standards, and a research roadmap for AI-simulated patients that are not only technically sophisticated but also ethically sound and transferable to real-world clinical practice.

### 4.3 Identifying Intrapersonal and Interpersonal Dynamics Predictive of Change in Mental Health

The third working group, led by Dana Atzil-Slonim (Bar-Ilan University, IL), focuses on intrapersonal and interpersonal dynamics as central mechanisms of change in mental health. The group brings together clinical and AI researchers to ask which aspects of within-person experience and between-person interaction – such as emotional trajectories, synchrony, co-regulation, or conversational redirection – are most predictive of positive (or negative) outcomes, and how these phenomena can be modeled using modern AI methods.

The planned position paper will first synthesize clinical and methodological work on verbal and non-verbal dynamics at multiple temporal scales (from utterances to sessions and longer-term courses of care), and then review AI techniques for capturing temporal and multimodal patterns. Building on this, the group will identify promising intersections between the two literatures, highlighting opportunities to leverage advances in NLP and multimodal models to analyze longitudinal mental health datasets to identify intrapersonal and interpersonal dynamics predictive of positive outcomes.

## 5 Conclusion and Outlook

This Dagstuhl Seminar on “Natural Language Processing for Mental Health” brought together researchers from NLP, clinical psychology, human–computer interaction, and digital mental health to examine how language technologies can responsibly support mental health needs. Across vision talks, lightning presentations, and breakout discussions, participants highlighted both the transformative potential of NLP and large language models for understanding mental states, supporting therapeutic processes, and scaling access to care, as well as the substantial challenges around evaluation, temporality, multilinguality, privacy, and ethics that must be addressed before these systems can be trusted in real-world, high-stakes settings.

The seminar has laid the necessary groundwork for tackling these challenges through concrete, interdisciplinary collaborations. The three working groups initiated at Dagstuhl are developing perspective and position papers on responsible evaluation of AI for mental health systems, the design and use of AI-simulated patients for psychotherapy training, and the identification of intrapersonal and interpersonal dynamics predictive of change in mental health. Alongside planned community-building activities such as blogs, workshops, and joint projects, these efforts aim to articulate shared frameworks, datasets, and research

agendas that bridge technical and clinical expertise. In doing so, the seminar advances a growing cross-field community committed to developing NLP and AI for mental health that is clinically meaningful, empirically grounded, and ethically responsible.

## Participants

- Tim Althoff  
University of Washington – Seattle, US
- Hiba Arnaout  
TU Darmstadt, DE
- Dana Atzil-Slonim  
Bar-Ilan University – Ramat-Gan, IL
- Daniel Blonigen  
Stanford University, US
- Stevie Chancellor  
University of Minnesota – Minneapolis, US
- Monojit Choudhury  
MBZUAI – Abu Dhabi, AE
- Torrey Creed  
University of Pennsylvania, US
- Cristian  
Danescu-Niculescu-Mizil  
Cornell University – Ithaca, US
- Munmun De Choudhury  
Georgia Institute of Technology – Atlanta, US
- Gavin Doherty  
Trinity College Dublin, IE
- Steffen Eberhardt  
Universität Trier, DE
- Anmol Goel  
TU Darmstadt, DE
- Philipp Graffe  
Universität Stuttgart, DE
- Iryna Gurevych  
TU Darmstadt, DE
- Nick Haber  
Stanford University, US
- Dirk Hovy  
Bocconi University – Milan, IT
- Darya Hryhoryeva  
Charles University – Prague, CZ
- Minlie Huang  
Tsinghua University – Beijing, CN
- Zac Imel  
University of Utah, US
- Hamidreza Jamalabadi  
Universität Marburg, DE
- Christopher Landau  
Universitätsklinikum Frankfurt, DE
- Jana Lasser  
Universität Graz, AT
- Maria Liakata  
Queen Mary University of London, GB
- Ryan Louie  
Stanford University, US
- Wolfgang Lutz  
Universität Trier, DE
- Matteo Malgaroli  
NYU School of Medicine – New York, US
- Emily Mower Provost  
University of Michigan – Ann Arbor, US
- Clarissa Ong  
University of Louisville, US
- Flor Miriam Plaza del Arco  
Leiden University, NL
- Julia R Pozuelo  
Harvard Medical School, US
- Alla Rozovskaya  
City University of New York, US
- Brian Schwartz  
Universität Trier, DE
- H. Andrew Schwartz  
Vanderbilt University – Nashville, US
- Raj Sanjay Shah  
Georgia Institute of Technology – Atlanta, US
- Bhavyajeet Singh  
TU Darmstadt, DE
- Thamar Solorio  
MBZUAI – Abu Dhabi, AE
- Aseem Srivastava  
MBZUAI – Abu Dhabi, AE
- Jina Suh  
Microsoft Research – Redmond, US
- Andreas Triantafyllopoulos  
Klinikum rechts der Isar der TU München, DE
- Diyi Yang  
Stanford University, US

