

A Hybrid Structured-Neural Dialog System for Automated Counseling

Stefan Olafsson

Abstract

Many therapeutic counseling techniques require the counselor to provide appropriate responses to open-ended client talk. Contemporary approaches to automated counseling either constrain user input or respond using simple rule-based methods, leading to less personalized responses and inflexible conversation. In this work I explore the use of a hybrid dialogue management approach that combines the use of rule-based sequencing through a predefined counselor-driven therapeutic agenda with neural network-based natural language processing to respond to individual client utterances, allowing for automated responses to unconstrained client speech in well-circumscribed discourse contexts.

This architecture was implemented in an automated counselor that uses techniques from cognitive behavioral therapy (CBT) and motivational interviewing (MI) to help individuals with mild to moderate alcohol use disorder decrease their alcohol consumption. CBT and MI are effective client-centered counseling methods for motivating people to change. Unfortunately, many people avoid therapy, citing barriers that could be addressed by automated counselors that simulate face-to-face counseling, such as lack of time, feeling stigmatized, and financial constraints. However, key counseling techniques, such as reflective listening in MI, require making appropriate responses to unconstrained client speech.

I report the development and evaluation of a virtual agent counseling system that conducts CBT-MI sessions with college students with mild to moderate Alcohol Use Disorder. I conducted a feasibility study assessing the acceptance of a prototype virtual agent counselor among patients with substance use disorder, developed data-driven models that predict counseling behaviors and models that generate the language for counselor reflections trained from transcripts of counseling sessions using neural networks. Finally, I built a virtual counseling system that uses the hybrid structured-neural dialog manager and conducted a randomized experiment evaluating the virtual agent counseling system efficacy conducting CBT-MI sessions with college students facing alcohol use problems. After one session with the virtual counselor, there was an increase in participants' readiness, confidence, motivation, and commitment to changing their drinking habits.

Table of Contents

Chapter 1.	Introduction	5
Chapter 2.	Background and Related Work	8
2.1	Virtual Agents for Substance Use Counseling	8
2.2	Digital Interventions for Substance Use Disorder	11
2.3	Dialog Systems	12
2.3.1	Dialog Management.....	13
2.3.2	Natural Language Understanding	24
2.3.3	Natural Language Generation	27
2.3.4	Dialog System Safety.....	28
2.4	Alcohol and Substance Use Disorders.....	30
2.5	Alcohol and Substance Use Counseling	32
2.5.1	Cognitive Behavioral Therapy	33
2.5.2	Motivational Interviewing.....	34
2.6	Summary of Related Work.....	40
Chapter 3.	Feasibility and Acceptance of a Virtual Substance Use Counselor.....	41
3.1	Virtual Counselor System.....	42
3.1.1	Virtual Counselor Dialog Management.....	42
3.2	Study Design	44
3.3	Results.....	45
3.4	Conclusion.....	47
Chapter 4.	Technical Approach to Hybrid Dialog Management	66
4.1	System Architecture and Processes.....	66
4.2	Hybrid Dialog Management.....	69
4.3	Scripting Language and Extensions.....	73
Chapter 5.	Predicting Counselor Actions	48
5.1	Dataset and Experiments.....	48
5.1.1	Expert Evaluation	50
5.2	Results.....	51
5.2.1	Expert Evaluation Results.....	53
5.3	Conclusion.....	54
Chapter 6.	Automated Reflection Generation.....	56
6.1	Materials and Methods.....	56

6.2	Evaluation	58
6.2.1	Measures	59
6.2.2	Procedure.....	59
6.2.3	Results – Trained Judges	60
6.2.4	Results – MI Experts.....	61
6.3	Discussion	63
6.4	Conclusion.....	64
6.4.1	Limitations	64
Chapter 7.	Domain Specific Application	76
7.1	Overview of the Implemented Counseling Session	77
Chapter 8.	Evaluation of a Hybrid Dialog System for Virtual Alcohol Counseling.....	83
8.1	Measures	84
8.1.1	Agent and Conversation Measures	85
8.1.2	Attitudes Towards Behavior Change (Pre-Post).....	86
8.2	Participants and Method	87
8.3	Procedure.....	89
8.4	Results.....	92
8.4.1	Quantitative Analysis	92
8.4.2	Qualitative Analysis.....	103
8.5	Discussion	108
8.5.1	Limitations	110
8.5.2	Future work.....	111
Chapter 9.	Conclusions	111
9.1	Future Work.....	114
References	115

Chapter 1. Introduction

Client-centered therapies for substance use disorders allow clients to express themselves freely in non-authoritative discussions that lead to the discovery and resolution of conflicting thoughts and values. Permitting clients to speak and exercise autonomy results in stable and enduring change, as it allows them to see the importance of the change and its association with their core values and beliefs [114]. Autonomous (or self-determined) motivation has also been found to be associated with greater confidence towards making changes to one's behavior [108], greater treatment adherence, long-term maintenance of the change, and medical outcomes [80]. Additionally, counselor responses that are tailored to the individual's unique situation are more likely to result in greater internalization and processing of the counselor's suggestions, as well as being more likely to be understood and recalled [92, 106]. This style of client-centered counseling is positively correlated with therapeutic alliance, specifically trust, rapport, and communicative success [107].

Unfortunately, some of the known barriers to engaging in treatment ultimately impede recovery for many patients, for example, poor social support, privacy concerns, time conflicts, lack of treatment availability, and difficulty getting admitted [112]. A potential solution to addressing these barriers is automating counseling using software that provides the necessary services to support one-on-one sessions. However, using software to manage client-centered counseling sessions poses significant technical challenges, such as handling turn-taking between participants in the conversation [125], understanding client utterances, and responding appropriately to those utterances, while managing the therapeutic agenda (the goals and tasks of the counselor). As a result, most contemporary automated counseling systems use highly-

constrained user input resulting in interactions where the counselor's decisions cannot be made based on the users' natural expressions.

In this dissertation, I describe the development of a hybrid spoken dialog system for conducting Alcohol Use Disorder (AUD) counseling sessions that couples a dialog management approach that maintains control across multiple turns of dialog, with one that allows the client bouts of unconstrained speech. In practice, counseling sessions follow an underlying structure that places boundaries on the kinds of conversations that can be had. At the same time, client-centered counseling requires allowing clients to express themselves freely while gracefully and coherently carrying on with the conversation.

Modern approaches to natural language processing have the potential to meet these demands by modeling language based on existing data and give accurate predictions on unseen data [69]. Deep artificial neural networks (DNNs) are systems that can learn a task without the need to program the rules for how to do so. For example, DNNs can model aspects of the underlying complexity of counseling conversations, as revealed in experiments where such methods performed better than non-neural methods to automatically annotate session transcripts [48, 103]. It could therefore be feasible to model counselor behavior directly from annotated patient-provider sessions, for example, to predict a counselor's next action, given the dialog context.

This progress in the field of machine learning has made it possible to create dialog systems that can handle natural language input and has the potential to generate coherent output [1]. However, most contemporary neural-conversational systems struggle to maintain conversational cohesion across multiple dialog turns, resulting in incoherent dialog. This has implications for the effectiveness of health counseling systems and their safety [16].

My work focuses on cognitive behavioral therapy and motivational interviewing (MI) for people with alcohol use disorder (AUD), the second most common substance use disorder in the United States [131]. The standard of care for people with AUD is face-to-face client-centered counseling, where counselors support clients in exploring their thoughts, emotions, and behaviors and motivate them to make changes that increase the likelihood of reaching their goals. MI is an effective counseling method for motivating clients towards positive health behavior change [89]. It is a non-directive, client-centered approach to promoting healthy behaviors, originally developed for counseling patients with AUD but has since been adapted to a wide range of health behaviors [124]. The success of MI as a counseling method depends on the counselor's ability to use its set of techniques properly [135]. For example, *reflective listening* is a technique where the counselor listens intently to the client and paraphrases the client's remarks in a neutral form to demonstrate understanding and empathy. This goes beyond simple grounding [37] to draw attention to entities or actions in the discourse history that the counselor wants the client to focus on in order to advance the therapeutic agenda.

The user interface to the counseling system is an embodied conversational agent (ECA), designed to simulate face-to-face interactions by displaying a virtual human that exhibits non-verbal conversational behaviors and speaks using synthetic speech [31]. ECAs leverage human implicit knowledge of conversation to make dialog systems feel more natural, intuitive, and approachable. They can use facial expressions, gestures, and body posture, along with speech, to convey nuanced meaning, such as approval, disapproval, and concern. Furthermore, ECAs with relational skills can build rapport and increase the working alliance with users by engaging in small-talk and expressing empathy [14]. ECA-based counseling systems that use aspects of MI have been successfully implemented [76, 129, 153]; however, these systems have featured either

user menu options as input or allowed for a limited set of user utterances, as opposed to supporting unconstrained user speech.

I developed and evaluated a virtual agent spoken dialog system for conducting MI counseling sessions for individuals with moderate AUD. The system design is built on theories of health behavior change and motivation, research on embodied conversational agents for health counseling, and state-of-the-art approaches to natural language dialog management. Its evaluation involved implementing a fully automated speech-input ECA application for conducting counseling sessions with college students at risk for problematic drinking. This work contributes to the fields of substance use counseling, intelligent virtual agents, natural language processing, dialog systems, machine learning, and personal health informatics.

In this work, I address the following research questions:

- R1:** Will individuals with substance use disorder accept a virtual agent as a counselor?
- R2:** Can counselor behavior be modeled directly from annotated counseling sessions?
- R3:** Can neural network-based approaches to natural language generation produce coherent and appropriate language for MI reflections?
- R4:** Can an automated speech-based virtual counselor positively impact college students' attitudes towards decreasing their alcohol use?

Chapter 2. Background and Related Work

2.1 Virtual Agents for Substance Use Counseling

A growing area of research explores the use of virtual agent technologies to simulate face-to-face counseling and conduct counseling sessions with patients. Relational agents, ECAs designed specifically for longitudinal use [18], have been evaluated in a variety of health care settings, and

can successfully combine some of the benefits of face-to-face interaction, e.g., empathic facial expressions and conversational behavior [12], with those of a computer-based intervention, e.g., non-judgement, standardization, and scalability [78].

Zhou et al. (2017) developed prototype relational agents for substance use screening. The results from the agent administered screener in the first prototype had fair agreement with the one administered by a physician and the responses using the second prototype were found to be similar to those gathered by a research assistant ($\kappa=0.9$). This prototype was then expanded to a relational agent system that was deployed in a primary care setting, providing education, screening, and brief intervention to veterans with alcohol use problems. Participants who scored high on a questionnaire screening for AUD were recruited through the VA in Boston and conducted two sessions where the agent discussed guidelines for safe drinking and offers to help the participant change their behavior. The authors based the intervention design on a model that incorporates common elements of effective brief interventions (FRAMES) [20] and incorporated some techniques from MI. One month later, the participants had the second session where materials covered in the first session were reinforced. Participants interacted with the agent via multiple-choice selection of utterances displayed in a menu next to the agent at every dialog turn. Participants reported a therapeutic alliance¹ significantly greater than neutral and their alcohol consumption frequency and quantity was reduced, compared to baseline. The agent generated more referrals to AUD counseling than to primary care physicians; however, there were no significant differences in drinking behavior between the agent intervention group and a standard care control group. Additionally, participants reported being comfortable with the technology,

¹ “[T]he optimal therapeutic alliance is achieved when patient and therapist share beliefs with regard to the goals of the treatment and view the methods used to achieve these as efficacious and relevant.”[5]

found the agent non-judgmental and not intimidating, and were willing to disclose sensitive information to the agent [153].

Schulman et al. (2011) created a relational agent system that had constrained user input (menu options) and used MI for long-term behavior change [129]. They developed a counseling framework that allowed the agent's dialog system to use methods like MI in its planning of therapeutic actions. In a longitudinal evaluation study, participants rated the agent higher than neutral on ratings of satisfaction, empathy, MI spirit (a measure of MI fidelity), and relational closeness, and an expert on MI rated the agent's empathy levels and adherence to MI highly [15].

Lisetti and colleagues developed and evaluated spoken dialog agent systems for conducting brief health interventions, such as motivational interviewing for people with alcohol use problems [76, 77, 149]. These systems are based on a Markov decision process framework that uses reinforcement learning on data collected from user interactions to optimize its dialog policy. They found that a system that had been optimized over user interactions achieved a higher rate of task completion, user likeability, and perceived accuracy [149]. They also developed empathic virtual agents for delivering motivational interviewing that adapt their verbal and non-verbal behavior to that of the user during counseling sessions. They found that interactions with the empathic agent compared to the non-empathic one led to more positive attitudes, higher intention to use the system again, greater perceived enjoyment, higher perceived sociability, higher perceived usefulness, greater sense of social presence, higher levels of trust, a higher rating of anthropomorphism, greater likability, higher levels of animacy, greater perceived intelligence, and greater perceived safety [76].

My approach differs, in that users have open ended speech-input and the system retains the structure of the session. Moreover, it uses natural language processing to generate agent responses, based on the context.

2.2 Digital Interventions for Substance Use Disorder

Information and communication technologies (ICTs) have been leveraged to deliver validated, effective, tailored, and scalable interventions through personal digital devices, such as computers and smartphones. A meta-analysis of 14 studies (15 in comparison) found that personalized-feedback interventions via the internet without human therapeutic guidance are a viable option for reducing problematic drinking in student populations [120]. Additionally, a review of computer-based alcohol interventions in primary care suggests that they can be effective in reducing alcohol consumption and drinking related consequences in the general population [67]. Research also indicates that ICTs delivered via email and text messages are more effective for impacting adolescent's awareness of drugs and other substances than changing their consumption habits [9]. Though technology-based interventions for reducing alcohol use among adults are largely effective [44], similar efforts in a student population has had ambiguous results with respect to behavioral outcomes [132]. A review of technology-based self-help revealed that computer-based cognitive and behavioral interventions without any human contact are efficacious; however, face-to-face or telephone contact with a therapist contributes more to the reduction of addictive behaviors [91]. Additionally, based on this research, there is generally little knowledge available about the longitudinal effects of computer-based treatments for substance use disorders.

In 2011, a review of eHealth interventions included a few randomized clinical trials evaluating intervention effectiveness for managing addiction [56]. Two systems have been

designed for Substance Use Disorder (SUD) management, specifically to help patients prevent relapse. The Addiction-Comprehensive Health Enhancement Support System (A-CHESS) is a mobile application designed for alcohol treatment patients leaving residential treatment that provides behavior monitoring, communication capabilities, an open channel for patients and providers to stay in touch, as well as other support services. In a randomized clinical trial, patients using A-CHESS reported significantly fewer risky drinking days after a 4 and 8 month follow up, compared to patients receiving treatment as usual [55].

The Therapeutic Education System (TES) is designed to be used in a 12-week outpatient addiction treatment program setting as a substitute for two hours of treatment-as-usual per week. The intervention is delivered through a browser and contains 62 modules focusing on the skills needed for achieving and maintaining abstinence, as well as a reward-based incentive system contingent on abstinence and treatment adherence. In a randomized clinical trial, patients in the TES condition had a reduced dropout rate and an increased abstinence rate compared to patients in a treatment-as-usual condition [29].

Many patient-facing technologies that have been evaluated, focus primarily on connecting patients with others, e.g., counselors [143], recovering addicts [58], or friends and family [150]. Other applications include stimulating cognitive faculties through mobile and mixed-reality devices to ward against relapse [45, 99] and using smartphone sensors to detect drinking episodes [7]. Few digital interventions simulate the counseling session experience.

2.3 Dialog Systems

Researchers have developed and evaluated dialog systems capable of having conversations with people for over five decades. Dialog systems have input modalities ranging from constrained user input (e.g. menu options) to natural language input (text or speech) and have been

developed based on theory and practice from a variety of disciplines, e.g., statistical modeling, probability theory, linguistics, and discourse analysis. The common goal of researchers that work on these problems is to bring humans closer to being able to intuitively interact and collaborate with machines. An important facet to making it possible for machines to organize words syntactically or capture their semantics, in addition to general natural language processing, is maintaining some representation of the general discourse as the dialog between participants unfolds. For example, allowing interlocutors engaged in a conversational situation to take actions towards some abstract goal and jointly reach it or gracefully fail, whether it be socializing, ordering a pizza, or collaboratively solving a problem.

2.3.1 Dialog Management

Artificial intelligence researchers have since the 1950s worked towards creating computers that can engage in intelligent behaviors, including conversations and dialog management [85]. In a general system, a dialog manager has the methods and procedures that determine how information is exchanged between a user and the computer system, including input such as text fields, buttons, checkboxes, or speech. In a spoken dialog system (SDS), the basic components are speech recognition, language understanding, dialog management, and response generation.

Dialog managers (DMs) are components in dialog systems that handle the flow of the conversation. The DM is charged with tasks such as receiving input from natural language input, managing turn-taking (e.g., initiative, aggregation of one or more utterances into a speaking turn, etc.), maintaining a dialog history, predicting dialog acts, and making calls to external components like databases. The DM implementation approach depends on its purpose. Some require a great amount of manual task-specific authoring or a robust representation of the discourse structure, while others make decisions based on probabilistic models.

The research efforts made towards reaching this goal have been different and have often depended on the intended final use of the application [75]. For example, an interactive voice response (IVR) application on the telephone has an SDS as well as the capacity to receive user input from the numeric keypad. A modern example of these are the automated troubleshooting voice agents that guide customers through basic troubleshooting:

- (1) (a) A: How can I help you today? You can say “technical support” or “billing”.
- (b) U: I need help with my router
- (c) A: Okay, before we continue, could you tell your username please?
- (d) U: John Doe.
- (e) A: I heard “John Doe”, is that right? You can say “yes” or press 1.

Such systems are designed to allow users to perform a single main task [23]. IVR dialog systems have the basic components of an SDS: a speech recognition module, language understanding module, a dialog manager, and a response generator. Achieving perfect speech recognition – that is mapping acoustic signals to strings of words – and language understanding via various natural language processing methods does not mean that dialog management has been achieved. The only way a conversation can go on is if some aspect of the system takes charge, deliberates, makes decisions, and manages the interaction as it unfolds. This is the distinction between a system truly capable of maintaining a conversation and a speech-enabled search engine, such as a question answering chatbot.

An early example of a system resembling the IVR example was GUS (General Understander System). Developed in 1977, GUS was built to be a travel agent able to converse with a client about booking a trip [21]. It was an ambitious design that attempted to model natural dialog by using data structures called frames and slots that both defined the possible user inputs at any given time and allowed the system to maintain a representation of the dialog state. This approach essentially became the classic way to perform dialog management: the author

defines the frames and slots necessary for the interaction and defines what the proper responses should be, given the state of these objects at any given time. With this structure, a dialog manager knows what actions to take, given a state, for example whether to ask for more information because a slot is unfilled or move on to the next frame. An example of filling in missing information is given in (1c) and an example of how the system can frame in the conversation, in an attempt to constrain the input is shown in (1a).

Representing the state of the dialog in this way gives rise to the problem of mutual belief, that is believing whether the other participant holds the same representation as you. One method of coming to hold that belief is through grounding utterances as the conversation progresses. People's communication is based on mutual knowledge, beliefs, and assumptions, which is known as the *common ground*, and *grounding* is the process of contributing it [37]. This contribution requires that participants perform actions cooperatively [38] and they will assume that mutual understanding is taking place until they are provided negative evidence that this is the case. to the contrary. For example, utterances such as 'huh?' and 'what?' are common verbal indicators of confusion in English and show negative evidence of understanding, suggesting that the listener misheard or misunderstood the speaker's performance.

When people converse, they tend to minimize what is necessary to reach a mutual acceptance of the common ground and will make sure their contributions have what is necessary without adding more complexity [51]. Moreover, the purpose of the conversation and the medium will change the type of grounding that is being used. For example, conversational agents that only have an audio-only interface have the same constraints on grounding as the telephone [37]. Conversational agents are forced to use techniques for grounding that work within those constraints. For example, systems that are audio-only cannot provide grounding information

using non-verbal behaviors that are common in human face-to-face conversations, such as nodding and gaze cues.

A simple example of grounding is shown in (1e), which may be a robust strategy for making sure that everyone is on the same page but is perhaps not very natural. Traum and colleagues have worked on computational formulations of grounding and obligations in dialog in a variety of projects [122, 133, 134] and take the view that mutual belief is achieved through grounding the utterances in the dialog. They state that dialog systems should be designed around an agent that has its own mental state, rather than the way one's favorite IVR experience is designed. Such systems have goals, interaction capabilities, knowledge, and beliefs, and dedicated subsystems for processing those data. This formulation is often referred to as BDI (beliefs, desires, and intents) and is at the heart of many agent-driven dialog systems.

With the BDI formulation comes a host of interesting problems, one of which is mutual belief. This can get complicated very quickly and in fact, a propositional logic representation of mutual belief is necessarily infinite [27], thus reaching idyllic mutual belief in a finite dialog would be impossible. A computational model of grounding by Bunt et al. [27] shows how grounding can be achieved by reaching partial mutual belief. This is accomplished by strengthening weak mutual beliefs through feedback-chaining, i.e., cumulative feedback for one's verbal or nonverbal behavior. Their model shows that on the second positive feedback, sufficient weak mutual beliefs have accumulated to achieve informational grounding. This is shown in example (2) whereby the time we reach (2e), sufficient weak mutual belief has been established such that A believes that U wants the blue one and U believes that A believes the same.

- (2) (a) U: I would like the blue one, please.
(b) A: The blue one?

- (c) U: Yes.
- (d) A: Ok, here's the blue one.
- (e) U: Thanks.

Roque et al's work in this space from 2009 addresses the idea of feedback-chaining by implementing a grounding model that considers the type of evidence of understanding being presented and gives weight to the different types using a criterion, resulting in evidence having varying degrees of groundedness [122]. Implementing a dedicated module for grounding allows the agent to have a representation of mutual belief about the information in the dialog.

The 'intention' aspect of the BDI formulation is usually implemented as a goal structure of some kind. In a task-oriented dialog, for example, the goals of the interlocutors closely resemble the structure of the dialog itself [52] and an agent-based dialog systems could thus formulate plans in accordance with the tasks. The work of Grosz and Sidner (1986) paved the way for goal-planning dialog systems grounded in a theory of discourse structure [53]. Their work was based on insights gained from discourse analysis, namely that discourse structure is made up of three components: the structure of the sequence of utterances (linguistic structure), the structure of purposes (intentional structure), and the state of the focus of attention (attentional state). This theory allows for the description of the processing of utterances in a discourse, which calls for understanding how utterances come together into segments, how intentions are expressed and related, and how the mechanisms of the attentional state allow the tracking of discourse.

Grosz and Sidner viewed engaging in a dialog as a collaborative effort. Collaboration is defined as a coordination of actions between two or more participants, towards achieving shared goals, mostly involving communication. Discourse is defined as communication between two or more participants in a shared context. Collaborative discourse theory is thus the empirical and

computational research about how people communicate in the context of a collaboration [116]. A theory of SharedPlans was developed to remedy the fact that the prevailing AI planning mechanisms did not provide a basis for explaining collaborative behavior [54]. According to SharedPlans, the intentional structure is part of the discourse context. Thus, conversational participants need to recognize the discourse segment purposes and the relationships between them in order to process further utterances in the discourse.

Collagen is a collaboration management system, implemented based on SharedPlans theory [117]. It mediated the interaction between a software interface agent and a user, similar to a discourse manager. Its primary function is to provide a representation for recording decisions that the agent has made and communicated. A SharedPlan comes about through the interaction between the participants. When the participants come to hold the beliefs and intentions required for the collaboration, then the collaboration has been planned. The execution or acting on these intentions is interleaved among participants; however, Collagen was not created for deciding how to interleave them. The key contribution of the system is the implementation of the separate data structures for the goals and actions (the task at hand) and the order in which they're executed (implemented as an actual stack of discourse segments called a focus stack). The discourse state is a representation of the discourse segments, focus stack, and the plan tree (i.e., an approximate representation of a SharedPlan at any given time). Collagen interprets the discourse to evaluate how the current action contributes to the current discourse purpose (i.e., goal). There is a discourse generation algorithm that functions in the opposite manner of the interpreter, i.e., it looks at the current focus stack and associated SharedPlan and produces a prioritized agenda of actions that would contribute to the current discourse segment purpose.

The Collagen architecture has been used in various applications, for example as the foundation of a tutoring dialog system [118], to generate natural language from the discourse context [39], and partially automate the generation of dialog trees that otherwise are manually authored [115]. An updated version of the system, Disco, was implemented in an “always-on” relational agent and robot architecture designed for longitudinal use by elderly people suffering from loneliness. The novelty of this system was the development of a reusable engagement module for a robot operating system and an extension to SharedPlans that included a theory of relationships [116].

A real-time version of Disco called DiscoRT represents the biggest architectural change to this series of collaborative managers since the original Collagen. In this system the collaborative manager is a submodule in a larger system that can process input from multiple modalities, such as speech and motion, has representations of possible activities that are continuously suggested to the system, and a special loop that collects behavior proposals and schedules them for real-time generation. Disco is the dialog management module and has a focus stack and a plan tree. The tree represents the agent’s goals and the focus stack captures the pushing and popping of the topics of the conversation [93].

Allwood et al. [4] critiqued the SharedPlans approach, stating that it relied strongly on agents that cooperate without providing a clear definition of cooperation itself and that a representation of discourse obligations would specify conditions where the agent adopts intentions for ethical considerations. A representation of trust and obligation would lead to more comprehensive and direct abilities to engage in a range of dialog behaviors. Traum et al.’s work on discourse obligations [134] introduced the idea that social conventions play a role in conversational decision making and contests the typical “strong plan”-based agent approaches.

Obligations tend to delay our pursuit of personal goals and thus allows for navigating conversations in a way that more closely resembles what a humans might do.

Other researchers critiqued the idea that the stack model is the best way to predict when information is available. Walker conducted an analysis of utterances that contain information that is no longer relevant in a discourse and showed that people restate previously mutually believed propositions when enough information has passed since the proposition was originally stated. This means that information that was previously grounded has fallen out of focus and needs to be reintroduced, despite still being on the stack and thus should still focused and salient. The author proposes modeling this limited attentional capacity using the metaphor of a cache as an alternative to the stack representation. A cache would maintain a working set of discourse entities, properties and relations that are currently being used for some process. Due to the limited capacity, some information would be taken from the cache and placed in long-term memory so that it's not immediately available. A cache could be used to model attentional state and a stack for discourse intentions [138].

There are systems that have some of the attributes of Collagen or Disco but focus more on dialog management in the classical spoken dialog system sense. An example of this is RavenClaw, a dialog management framework that allows for rapid development of dialog management components and supports spoken dialog systems in complex goal-oriented domains. It can interpret user inputs with respect to tasks within the domain specification and maintain coherence during the conversation. The system separates the dialog task specification that captures all domain-specific dialog logic from the dialog engine, which is a domain-independent component that controls dialog by executing the dialog task specification and contributing basic

conversational strategies (e.g., timing and turn-taking, grounding, help, repeat, stop/resume, etc.) [23].

The development and evaluation of systems able to process and generate multimodal behavior has been the focus of researchers interested in creating software agents capable of conducting face-to-face conversation. ECAs have the ability to recognize and respond to verbal and nonverbal input, generate verbal and nonverbal output, deal with conversational functions, such as turn-taking, feedback, and repair, as well as give signals that indicate the state of the conversation and contribute new propositions to the discourse. In essence, an ECA is a virtual human that has social and linguistics intelligence and the ability to hold a face-to-face conversation [32].

The system architecture of REA includes a decision module that resembles the classic discourse and dialog manager component in dialog system architectures. However, it is more sophisticated, as it processes both propositional and interactional information and plans responses that are processed by a dedicated behavior generation module. Evaluations of REA in a “Wizard-of-Oz” setup – where the system is covertly controlled by a research confederate – compared the full version to one without the interactional capabilities showed that (a) users judged the full version as more collaborative and cooperative, and (b) that they could interact with the system without any prior training [8].

The development of ECAs has relied on studying how humans conduct face-to-face interactions. In a study of the relationship between turn-taking, discourse structure, and gaze behavior, researchers found that the co-occurrence of turn-initial and turn-final units in information structure units is very predictive of gaze behavior [34]. This data was subsequently used to build a model of gaze behavior that was implemented in REA’s decision module.

A more recent ECA architecture has a dialog manager and behavior realization system that incrementally adapts to utterances and behaviors produced by conversation participants. The Articulated Social Agents Platform (ASAP) Realizer is capable of switching turns smoothly, handling interruptions on-the-fly, and executing movements in synchrony. This system is a move away from the traditional turn-based systems, by allowing incremental understanding and the continuous processing of input and planning behaviors. It combines two previous realizers that focused on incremental multimodal utterance construction and interactional coordination, showing that this combination enables situations that were previously unattainable [32]. The adaptive dialog manager for conversational agents, flexdiam [33], is designed for people with a wide spectrum of cognitive capabilities and is used in spoken-dialog systems that aim to handle and perform conversational speech, incremental feedback, and provide information updates. It provides flexibility through a repair mechanism that fixes problems quickly and interactively, taking the interlocutor's capabilities into consideration [34].

The capabilities of systems like flexdiam and the ASAPRealizer are possible because of advances in spoken dialog system technologies. Allen et al.'s SDS had speech interpretation that used syntactic and semantic parsing, statistical error correction, and discourse history to make the system robust [3]. It fused three components that made this approach unique. The first was a statistical error correction post-processor, the second was a rule-based constituent parser that outputted a set of the most plausible speech acts, and the third component processed the speech acts using a dialog manager that maintains a discourse state akin to the attentional state found in Grosz and Sidner's theory of discourse structure [53].

At the turn of the millenium, Roy et al. introduced a method for modeling spoken dialog managers using a previously computationally intractable representation. Their paper claims that

spoken dialog managers using a Partially Observable Markov Decision Process (POMDP) for generating dialog strategies is an improvement over the conventional Markov Decision Process-based planners, which use exact values and are therefore ill suited for handling noisy input and ambiguous language. The method involves redefining what a dialog state is, that is a user's intent as opposed to the system state. The intention of the user with respect to the task is the underlying state, thus the state space is partially unobservable. Previous POMDP solvers had high complexity and were intractable for this task, therefore the authors used an algorithm that compresses the belief state, making belief-space planning feasible. The machine infers the state of the user from the noisy input and the framework provides a mechanism for modelling uncertainty [123].

The exploration of using POMDPs for spoken dialog management has since been reported in several papers [127, 152], and are summarized in a review from 2013. The review claims that progress in conversational speech systems has been slow in the past few decades and speech recognition errors have dramatically decreased, paving the way for various approaches to tackling the problem of handling spoken dialog in real-world situations. Traditional conversational systems are expensive to build and non-scalable, thus, statistical SDSs based on the mathematical framework of POMDPs represent an interesting avenue for continued exploration. POMDP-based SDSs combine belief state tracking and reinforcement learning. The belief state represents uncertainty and pursues all possible dialog paths in parallel, making for easy error correction. Rewards are associated with state-action pairs and the sum of the rewards is an objective measure of performance, allowing for using reinforcement learning to maximize system performance [151].

With large state spaces come large problems, thus the main efforts of researchers working on statistical SDSs have focused on creating efficient techniques for making this approach more tractable. POMDP systems should ideally be trained with real users, which is usually done in the lab or in the field. However, researchers have recently looked to online crowdsourcing platforms to recruit participants to train these dialog systems by interacting with them and providing turn-by-turn feedback on the model's performance. These interactions have also been used to build systems that can simulate user interactions in order to generate more dialog data for training. This method enables a wide space of possible dialogs and scenarios; however, it comes with the problem of there being a discrepancy between simulated and real user behavior. In addition, user simulation systems are quite complicated to build [151].

This overview accounts for a fraction of the research that has been conducted in the interdisciplinary field of dialog management. Though it seems like we have come a long way since the advent of systems like GUS, dialog systems built using hand-crafted policies and constrained user input are still the most robust. Advances in speech recognition and some aspects of natural language understanding have not happened at the same rate for dialog management. The impact that using neural networks for natural language processing has had, is yet to become clear for conversational systems. There is a genuine lack of high quality data in order to train probabilistic models using contemporary data-hungry methods [151], which becomes particularly apparent when researchers are spending significant time and effort building sophisticated user simulation systems just to generate the data for training.

2.3.2 Natural Language Understanding

Dialog managers use a dedicated natural language understanding (NLU) process to glean meaning from a user's utterance and to know what actions to take, given the state of the dialog at

any given time. For example, if a slot is unfilled the dialog manager can ask the user for more information, or if a slot is filled it can move on to another frame. There are a variety of approaches to NLU for spoken dialog systems [85]. Conventionally, an NLU process parses a user's utterance and uses grammars to decompose it into its constituent parts, i.e., a group of words that form a unit, such as noun and verb phrases. Traditionally, this was achieved using a lexicon and an ontology to interpret the "semantics" of the input, with respect to the dialog system's purpose.

NLU systems often use a combination of methods to achieve their task. One such method is dialog act classification, which involves determining the function of the user's utterance in the dialog [2]. This reduces the user's utterance to a concrete goal or intention. For example, asking a question reveals a user's intent to seek an answer and this knowledge affects how the dialog manager formulates its response. Similarly, intent identification is a task where a user's utterance is classified as one from a fixed set of intents (such as to make a purchase or request information), with the aim of finding the set of slots (or frame) that satisfy that intent [146].

Knowledge-based approaches to NLU, with handcrafted grammars and meaning representations, are ill-suited for handling open and unconstrained user speech that may include previously unseen words and irregular use of grammar. In the 1990s, statistical methods were increasingly used to learn grammars automatically from textual data [119] using machine learning, whereby machines build models by learning from experience (existing data) and give predictions on unseen examples (new data) [69]. Meaning representations shifted from being *symbolic* to *distributed* as distributional semantics became an area of research founded on the idea that "a word is characterized by the company it keeps" [43]. Under this paradigm, the information about linguistic items (words) is based on their distribution in large collections of

text. Specifically, this involves mapping discrete, categorical linguistic items to vectors of numbers in a process called *embedding*. Now concepts like the semantic similarity between words could be quantified and defined as the similarity between the words' vectors [119].

Word-vectors are well suited for computer algorithms and word-embeddings have become the backbone for contemporary language processing using neural networks (NNs) since they were first introduced in 2003 [11]. These methods have achieved state-of-the-art performances on many important tasks, such as automatically tagging words with their part of speech [22], translating from one language to another [41], converting speech to text [147], and identifying entities in text (like names, locations, and organizations) [8].

A NN is a collection of connected processing units called neurons that produce a signal, a single number called an *activation*, that is transmitted to other neurons [128]. The neurons in NNs are organized in layers. The first (input) layer neurons receive the data and activate neurons in the next layer through weighted connections. The last layer produces an output, for example, a network that tags words with parts of speech will output a part of speech for any given word as input. The network may have any number of hidden layers between the input and output layers, making it arbitrarily *deep*. Deep-learning involves teaching an NN with multiple hidden layers to perform a task by taking the following steps: (1) feeding in the input (such as a word); (2) propagating activations forward through the layers of neurons via the weighted connections; (3) getting an output from the last layer (such as a part of speech); (4) calculating the error (i.e., how different the predicted part of speech was from the ground truth); and (5) adjusting all the connection weights based on the calculated error by propagating backwards through the layers of the network. While the error rate continues to decline over time, the network is learning [128].

Once learning ceases to improve, the performance of the network (model) can be evaluated using previously unseen data that was held-out during training.

One of the drawbacks to using NNs for dialog systems is the amount of data needed to train high-quality models. For example, models for tasks such as dialog act classification are learned from processing hundreds of transcripts of dialogs where each utterance of interest has been manually annotated with dialog acts [103]. However, recent methods have made it possible to use language models that are pre-trained on immense datasets and then fine-tuned on the domain-specific dataset for the task of interest. This method of *transfer-learning* has been found to improve the state-of-the-art on a number of language processing tasks [104, 105] and can benefit the developers of dialog systems designed for domains that have a limited amount of data available.

2.3.3 Natural Language Generation

Modern natural language generation (NLG) methods have the potential to produce utterances for a virtual counselor tailored to users at runtime, based on the context of the conversation.

Traditionally, spoken dialog systems use NLG to transform structured data into natural language using one of two main approaches. The first is template-based generation, where the system uses pre-defined templates that either contain fully-formed text ready to be spoken by the system or have missing values that can be filled (like slots for NLU), given the state of the dialog. An example of the former is when a flight booking system asks the user where they would like to go: “What is your destination city?” An example of the latter is asking the user their departure time and slotting the name of the destination into the system’s question, such as “On what date would you like to go to <CITY>?”

The second main NLG approach is word-by-word generation. As with NLU, system creators may define a grammar and rules used in conjunction with information about the dialog state to produce the words to be spoken. For example, when a dialog system's NLU realizes that it should ask the user a question about their flight departure time, the NLG component may use a semantic representation of flight departure time and a grammar designed to produce a syntactically correct question. This method would yield results that are grammatically correct and predictable, in the sense that one can always discern how the responses were produced. However, this method lacks the flexibility needed for handling open-ended conversations and requires a great deal of manual authoring.

Modern state-of-the-art machine-learning-based language generation systems also produce language word-by-word or character-by-character [50]. These systems typically use models of language (such as English) trained on large datasets using deep-learning architectures, like transformers, that can be fine-tuned on smaller domain-specific datasets for the task at hand [111]. Approaches to NLG for dialog that combine the use of neural network architectures, like sequence-to-sequence and transformers [1, 136], with conditioning variables, like topics or goals, have been found to perform better than traditional methods in task-oriented domains [61, 63, 70].

2.3.4 Dialog System Safety

Using automated and autonomous methods to drive health counseling systems with humans has implications for their effectiveness of and safety [16]. These systems have various safety concerns, such as a system giving recommendations that could result in the user making a harmful decision, that may have been due to faults in system modules that are imperceivable to the user. For example, (1) the system's natural language understanding module may misinterpret user intentions, causing the language generation model to ultimately output an incorrect or

unhelpful utterance; (2) users may not know the full extent of the capabilities of the system and may lend more credibility to its recommendations than is warranted; and (3) lack of coherence across dialog turns can send mixed messages to the user, sow confusion, and ultimately lead the user to take advice that is potentially harmful.

Errors as a result from non-understanding and recovery strategies from them have been researched within the discourse and dialog system community. Bohus and Rudnicky investigated the main sources of conversational errors using a spoken dialog system for conference room booking and analyzed their impact on system performance [24]. They also compared how the strategies impacted user responses and if they lead to a successful recovery. They identified ten strategies the systems can use to recover from errors of misunderstanding and the strategies that had the top three highest dialog recovery rates were: (1) moving on to the next part of the task; (2) giving a full description of where they are in the dialog, what the problem is, and what the user can say at this point; and (3) telling the user what they can say at this point. The most successful dialog recovery strategy was to move on to the next part of the task without acknowledging the misunderstanding. This mirrors results from studies on how people often decide to recover from breakdowns in conversation, that is to not mention the problem and ask different questions related to the task [130].

A safety concern that is particular to large transformer language generation models are universal adversarial triggers. These are particular strings of characters that can be discovered automatically to force the model to consistently generate harmful language, such as racist utterances [142]. These universal triggers are often nonsensical from a natural language perspective, thus the probability for a user to accidentally include them in the input to a generation model is very low. Nevertheless, these adversarial triggers represent a vulnerability

since people with nefarious intent can certainly find and use them. Users of health counseling systems that provide recommendations should be counseled to consult a medical professional before acting on the advice these systems give.

Most robust dialog systems still use hand-crafted policies and constrained user input. Advances in natural language processing have not occurred at the same rate for dialog management, and the efficacy of using deep-learning to create conversational systems is not yet clear. There is a genuine lack of high quality data in order to train probabilistic models using contemporary data-intensive methods [151], which becomes apparent when researchers spend significant time and effort building sophisticated user simulation systems in order to generate data for training.

2.4 Alcohol and Substance Use Disorders

Substance use disorders (SUDs) are common illnesses affecting 1 in 12 American adults from all walks of life [110] and are characterized by changes in thought processes, mood, and/or behaviors. SUDs are prevalent in the United States, with tobacco use disorder in first place and alcohol use disorder (AUD) second place [131]. Every SUD has its own definition according to the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5). For example, AUD is defined as problems controlling the intake of alcohol, continued use of alcohol despite problems resulting from drinking, development of a tolerance, drinking that leads to risky situations, or the development of withdrawal symptoms [6].

SUDs are generally defined as “the recurrent use of alcohol and/or drugs causing clinically significant impairment, including health problems, disability, and failure to meet major

responsibilities at work, school, or home.”² In the United States, only about 19% of individuals with SUDs receive treatment [131]. Additionally, approximately 40-60% of people treated for SUDs relapse within a year [84]. Relapse to substance use is defined as a return to a disease state after a period of remission, where the individual is either ill or well, as with other chronic diseases, or defined as the failure to maintain newly formed behaviors that help an individual keep substance use at non-pathological levels [82]. For example, an individual diagnosed with Alcohol Use Disorder that has been meeting their responsibilities to their work, themselves, and their loved ones, can be considered to have relapsed if they fall back into patterns of drinking that negatively affect their ability to meet their responsibilities. The road to relapse may begin with subtle changes in established preventative behaviors long before their behavior reaches pathological levels, such as breaking an exercise routine or getting reacquainted with people that they associate with the problematic behavior. The compounding effect of breaking positive routines and behaviors, as well as lapses into further drinking episodes, may ultimately increase the odds of engaging in harmful behaviors again.

Alcohol use disorder (AUD) is the second most prominent SUD in the United States [131]. Alcohol misuse and its negative emotional, financial, and physical consequences disproportionately impact college students in the United States. For the general population, alcohol misuse is the third leading preventable cause of death and alcohol addiction is the second most common addiction in the United States, with 25.8% of adults in 2019 reporting engaging in binge drinking in the past month and an estimated 14 million adults (5.4%) being diagnosed with AUD. That same year, however, 8.7% of full-time college students ages 18-22 met the criteria for AUD, 33% reported binge drinking in the past month, 52.5% drank alcohol in the past month,

² <https://www.samhsa.gov/find-help/disorders>

8.2% reported heavy alcohol use in the past month, and 1,519 died from alcohol-related unintentional-injuries.³ Moreover, college students meeting the criteria for SUDs rarely seek traditional treatment [28].

College students are among the most technologically savvy populations in the United States, with high levels of adoption and satisfaction with modern technology and devices (95% laptop and 97% smartphone ownership) [26]. Being a young and highly educated adult, with a relatively high socio-economic status, is predictive of using the internet for seeking health information, as opposed to traditional resources and printed [65]. Additionally, Americans aged 18-29 are the most likely population to engage with virtual environments on a regular basis through video games, across a variety of platforms.⁴ Since college students with AUD are relatively unlikely to seek traditional treatment and given their proclivity for using technology, virtual or otherwise, in many aspects of their lives, including health, a solution involving virtual agent technology has the potential to be impactful.

2.5 Alcohol and Substance Use Counseling

Severe SUD, or addiction, is defined as a chronic relapsing brain disease [137] and The American Society of Addiction Medicine state that addiction is characterized by cycles of relapse and remission.⁵ Other characteristics of substance use disorders, including alcohol use disorder, include a physiological dependence to it, lack of behavioral control, craving, and use despite experiencing interpersonal, functioning, and emotional problems. The consequences of addiction are multi-dimensional, impacting one's emotional, social, and physiological wellness.

³ <https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-facts-and-statistics>

⁴ <https://www.pewresearch.org/fact-tank/2017/09/11/younger-men-play-video-games-but-so-do-a-diverse-group-of-other-americans/>

⁵ <https://www.asam.org/resources/definition-of-addiction>

Therefore, therapy for individuals with severe SUD requires a multi-faceted approach to successfully address these dimensions. Different methods are relevant for different treatment stages, for example, medications may be used for detoxification following admission to a treatment facility, while individual or group counseling sessions may occur during in- or out-patient treatment, and peer support programs may be relevant after discharge from treatment⁶.

Two of the most common approaches to substance use disorder treatment are Cognitive Behavioral Therapy and Motivational Interviewing. Both approaches are client-centered, whereby the patient and therapist work together to help the patient recover from their illness.

2.5.1 Cognitive Behavioral Therapy

Cognitive Behavioral Therapy (CBT) is a psychotherapeutic approach where the patient and therapist work collaboratively to resolve their illness. CBT is based on the cognitive model treatment approach [10] and aims to correct problematic behaviors, such as excessive substance use, using a particular set of skills. For example, identifying positive and negative consequences of substance use and identifying situations that put one at risk for use⁷. CBT has been shown to be more effective than minimal treatment for alcohol and other drug use disorders [79].

CBT clinicians work towards understanding the basis of client's beliefs and thoughts. The client is encouraged to collect evidence for and against their thoughts to help them form new ones by asking questions of themselves. The clinician guides the client step by step through this process and challenges them with questions that aid in the discovery of thoughts that are more helpful and does not directly come up with alternative thoughts for the client.

⁶ <https://www.samhsa.gov/disorders/substance-use>

⁷ <https://www.drugabuse.gov/publications/principles-drug-addiction-treatment-research-based-guide-third-edition/evidence-based-approaches-to-drug-addiction-treatment/behavioral-therapies/cognitive-behavioral-therapy>

CBT sessions are generally tailored to the target patient population and the specific psychological issue they are facing. For example, a treatment protocol combining CBT and Motivational Interviewing designed to address depressive symptoms and binge drinking among young adults was found to decrease symptoms and reduce alcohol consumption [102]. Sessions typically begin with a meeting agenda and at the end of the meeting the client is often asked to complete homework assignments relevant to the topics covered in the session. The completion of the assignments makes it easier for the client to implement the techniques learned in the sessions in practice. The clinician then begins each session reviewing the assigned homework and should make strides to reinforce their completion.

CBT typically has three phases: an initial, middle, and ending phase. The initial phase focuses on assessing the client's expectations for treatment, the middle phase on implementing the cognitive and behavioral strategies to help mitigate the client's unhelpful thoughts, and the ending phase emphasizes preventing relapse and includes a plan for ending the treatment.

Relapse in the context of SUDs and CBT can be defined as the failure to keep substance use at a non-pathological and relapse prevention as a therapy was created following the development of a CBT model of relapse [81, 82]. A meta-analysis of various SUD treatments using the RP model suggested that they are largely effective in improving psychosocial functioning, with strong treatment effects for alcohol use [64].

2.5.2 Motivational Interviewing

Motivational Interviewing (MI) is a counseling method for enhancing people's motivation for change through a variety of techniques. The main goal of MI is for counselors to help their clients – through conversation – explore and resolve ambivalence they may have about their current behavior and get them to consider engaging in behavior change or maintaining positive

behaviors they already show [89]. During typical brief MI counseling sessions to reduce alcohol use, counselors use self-reported patient information, such as drinking patterns, to provide personalized feedback. The sessions consist of several sections designed to address aspects of the client's alcohol use and its consequences and can therefore be tailored according to the self-reported information, since some topics will be more relevant to individual clients than others. However, motivating people to adopt or change their behavior is a difficult task and gracefully managing resistance to behavior change is crucial.

Counselors elicit *change-talk* from clients using a variety of strategies. Change-talk corresponds to statements by the client that reveal consideration of, motivation for, or commitment to change [89]. Strategies include listing the pros and cons of their behavior, talking about their current level of motivation and confidence to change, and encouraging them to speak freely without being judged. When clients are at the stage where they are maintaining previously acquired healthy behaviors, the counselor's goal becomes helping them manage substance use triggers, work on a coping plan for preventing relapse, and processing challenges or successes. Clients may resist working towards changing their unhealthy behaviors at any stage and MI is an effective tool for counselors to decrease this resistance and move towards willingness to change.

MI is often used in tandem with the transtheoretical model of behavior change (TTM) [109]. The TTM emerged out of psychotherapy in the mid-1980s and uses the concept of *stages of change* to track where individuals are in their behavior change trajectory (Table 1). This allows health behavior change intervention designers to tailor the content to individuals according to their stage and the stages can be used for assessing the intervention user's attitudes towards a target behavior, such as alcohol use [49]. MI has been found to be an effective method

to move individuals from the earlier stages, where people are more resistant to or not thinking about change, to the ones where people have moved further along the path to change.

Stage of Change	Definition
Precontemplation	Having <i>no</i> intention to take action towards behavior change
Contemplation	Having <i>some</i> intention to take action towards behavior change
Preparation	Having the <i>intention</i> to act within the next six months
Action	Having <i>taken action</i> within the last six months
Maintenance	<i>Continuing</i> to exhibit the intended behavior for more than six months

Table 1. The Transtheoretical model's stages of change and their definitions.

2.5.2.1 Techniques and Principles of MI

The key to achieving a successful motivational interview is using the appropriate style (e.g., empathy) and technique (e.g., reflective listening) to create an atmosphere of collaboration between the counselor and client. An MI counselor aims to increase a person's motivation for change by using four main techniques [89]:

- **Open questions.** The client is given the opportunity to talk about events in their own words without being led in a specific direction.
- **Affirmations.** Statements by the counselor that highlight the clients' strengths and acknowledges any positive behavior.
- **Reflective listening.** This involves listening intently to the client as they speak, showing them that they're being heard. Counselors signal this by being attentive, using verbal facilitations (e.g., 'mhhh'), and acknowledgements (e.g., 'okay' and 'I understand').
- **Summary reflections.** The counselor repeats, paraphrases, or summarizes part of what the client communicated during reflective listening. This gives counselors the opportunity to highlight what they want the client to focus on and a mechanism for controlling the conversation without seeming coercive. Reflections can be simple or complex, ranging

from a single word to a few sentences, and choosing one over the other at key moments has been found to impact the likelihood of the client expressing change-talk [74].

The counselor uses these techniques while adhering to the five principles of MI [89]:

1. Express empathy through reflective listening
2. Develop discrepancy between clients' goals and their current behavior
3. Avoid argument and direct confrontation
4. Adjust to client resistance rather than opposing it
5. Support self-efficacy and optimism

It is important for counselors to have the skills to balance these techniques and principles appropriately. Research into the relationship between MI skills and treatment outcomes has shown that counselors with better MI skills achieve better outcomes, following brief motivational interventions on patient alcohol use. This work also found that avoiding MI-inconsistent skills was more important than using MI-consistent skills at a particular rate or frequency during sessions [46].

2.5.2.2 Motivational Interviewing Skill Codes

The quality of MI counseling sessions can be measured by annotating session transcripts with a set of labels, such as the Motivational Interviewing Skill Codes (MISC), that codifies the principles and techniques of MI and then analyzing how well the counselor adheres to them [66].

Table 2 shows an example exchange between a counselor (Carl) and a patient (Pat) from a brief counseling session. The items listed in the "MI Action" column are MISCs and can be thought of as moves or actions taken by the during the conversation. With these strategies (or moves) the counselor can maintain a therapeutic agenda – covering topics and tasks they deem important – while allowing the client to speak freely.

In the example, counselor Carl asked patient Pat to talk about when he first started drinking, with a statement that's classified as an open question (line 1) but refers more generally to an invitation to speak freely. Carl invites Pat to continue expressing himself by using facilitation moves (lines 3, 5, 7, and 9), until Carl either feels the need to get Pat engaged in the conversation by grounding the mutual understanding of the information presented so far or he feels that enough information coinciding with his underlying therapeutic agenda has been divulged. Then Carl reflects and summarizes (line 11) the information of significance that has transpired across potentially multiple turns of talk. Pat adds more information and Carl reflects again (line 13), only this time it merely consists of the new information gained in the previous turn in order to acknowledge Pat's contribution. Carl does another complex reflection (line 14), referring to information from an earlier passage, since Carl wants Pat to focus on talking about negative alcohol related behaviors. Finally (line 15), after setting up the topic in the previous reflection, Carl invites Pat to speak freely about those behaviors.

	Speaker	MI Action (MISC)	Transcript
1	Carl	Open question	Tell me about [when you first started drinking]
2	Pat		I was about 14.
3	Carl	Facilitation	Mhmm.
4	Pat		And I had a lot of family problems and I ended up being removed from my home and group homes. That kind of led to everybody, drink this; do this; do that; do this.
5	Carl	Facilitation	Mhmm.
6	Pat		And there were a few times where I was a stupid kid and I didn't know any better and I'd abuse alcohol a little bit, like break up with a girlfriend, be hurt and go pound a bottle. But no, my alcohol use isn't very much. It used to be a little bit heavier. Not heavy. But then my last

			girlfriend was a raging alcoholic and it scared me so I backed off drinking.
7	Carl	Facilitation	Mhmm.
8	Pat		Her thing was vodka. So I don't even touch vodka anymore. I'm like, nah.
9	Carl	Facilitation	Mhmm.
10	Pat		I see what it did to her. But I don't really have any set typical drinking things. Just playing cards or playing pool. That's it.
11	Carl	Complex reflection	Okay. So alcohol, sounds like it used to be a way to either deal with some stuff whenever you were younger.
12	Pat		Escape from stuff, yes.
13	Carl	Simple reflection	So it was an escape for you.
14	Carl	Complex reflection	And then with your last girlfriend where it sounds like she used it heavily, some of her behaviors scared you.
15	Carl	Open question	Tell me about those behaviors.

Table 2. Example exchange between a counselor (Carl) and a patient (Pat) from a brief MI counseling session.

2.5.2.3 Automatic MI Annotation

Modern machine learning methods can be used to model patient-provider conversations, for example, to automatically annotate MI session transcripts, allowing researchers to analyze the quality of MI sessions at scale. Wallace et al. showed the feasibility of using machine learning to automatically classify utterances in transcripts of patient-provider communication. They used a conditional random field [73] to estimate the probabilities of six relatively high-level topics, achieving an inter-rater reliability between the model and human annotators of 0.49 and an average accuracy of 0.64 [140]. They also modeled topics and speech-acts in utterances

comprising patient-provider interactions jointly, which achieved better performance than a model in which topics and speech-acts were modeled independently [139]. They later extended this model to incorporate parameters representing individual doctors' speech acts, and clustered physicians based on these. The induced groupings were found to correlate significantly with the scores of patient ratings of physician communication [141].

Gibson and colleagues used a Recurrent Neural Network (RNN) – specifically a Long Short-term Memory (LSTM) [60] – to: (1) classify the Motivational Interviewing Skills Code labels per utterance, and; (2) predict counselor session level empathy ratings. They used these dialog turn level behavioral acts as an encoding for a session level empathy rating. This approach outperformed training the empathy predictor without these intermediary dialog acts [48].

Pérez-Rosas et al. provided further evidence that RNNs are a good fit for modeling MI sessions. In this work, the authors modeled motivational interviewing sessions to automatically identify certain counselor behaviors. They annotated 277 transcribed MI sessions using a standard coding scheme, which measures counselor MI proficiency by evaluating verbal behaviors, such as reflective listening. They showed that using a feature set combining semantic and syntactic features leads to higher model performance, as compared to using bag-of-word features, and that a Gated Recurrent Unit model [36] (a particular type of RNN) achieved the highest performance for annotating counselor reflections [103].

2.6 Summary of Related Work

Digital interventions for substance use disorder rarely focus on simulating the counseling session experience itself. ECAs simulate face-to-face conversations and have been used in automated counseling systems; however, these systems have not been designed to handle unconstrained

user input. Current ECA systems have not been designed to allow for unconstrained user speech of the kind that client-centered counseling calls for.

Deep-learning approaches have been found to aptly model the language of such sessions for the purpose of automatic transcript annotation. Dialog managers that combine impactful past approaches with recent advances in natural language processing may be well suited to conduct natural, safe, motivating, and engaging face-to-face conversations. For example, systems that implement a representation of discourse structure, processes for propositional and interactional information, handling incrementally and continuously processing user input, and making use of statistical models trained on human-human dialog data. Substance use disorder is a prevalent and debilitating condition and motivational interviewing is clinically proven to be an effective counseling method for substance use disorder treatment. I therefore explored recent advances in natural language processing to model patient-provider sessions to create an ECA-based CBT-MI automated counseling system that uses a dialog management approach where an end-to-end neural network dialog system is embedded within a state- and rule-based architecture.

Chapter 3. Feasibility and Acceptance of a Virtual Substance Use Counselor

Before beginning development of the counseling dialog management system, I addressed my first research question “Will individuals with substance use disorder accept a virtual agent as a counselor?” I first designed a prototype virtual counselor and evaluated its acceptance among patients in treatment for substance use disorder. Part of this work was presented at the GREATS workshop collocated with the international conference on Intelligent Virtual Agents in 2018 [98]

and published in the proceedings of the Autonomous Agents and Multiagent Systems in 2020 [96].

3.1 Virtual Counselor System

The virtual counselor was an ECA that shows relational behavior [13], such as engaging in small-talk and showing verbal and nonverbal displays of empathy. The relational agent spoke using a synthetic voice, was automatically driven by a dialogue system, and used template-based text generation (Figure 1). User inputs were made via constrained user selection of utterances. The agent program was built using the Unity3D game engine and the agent's speech was synthesized using the IVONA Sally voice for the English-speaking agent and IVONA Dora for an Icelandic version of the agent. To our knowledge, this was the first conversational agent evaluated in a clinical setting using Icelandic language technologies.

3.1.1 Virtual Counselor Dialog Management

The virtual counselor system's dialog manager was implemented as a hierarchical transition network. This dialog manager is based on Grosz and Sidner's theory of discourse structure [53] that defines discourse as segments that have a particular purpose (see section 2.3.1). As discourse progresses, these segments are stacked, forcing the top segment into focus at any given time. Once a segment is completed, it is removed from the stack, activating the segment below. This allows participants to shift away from and return to topics of conversation, resulting in the dialog taking on a hierarchical structure.

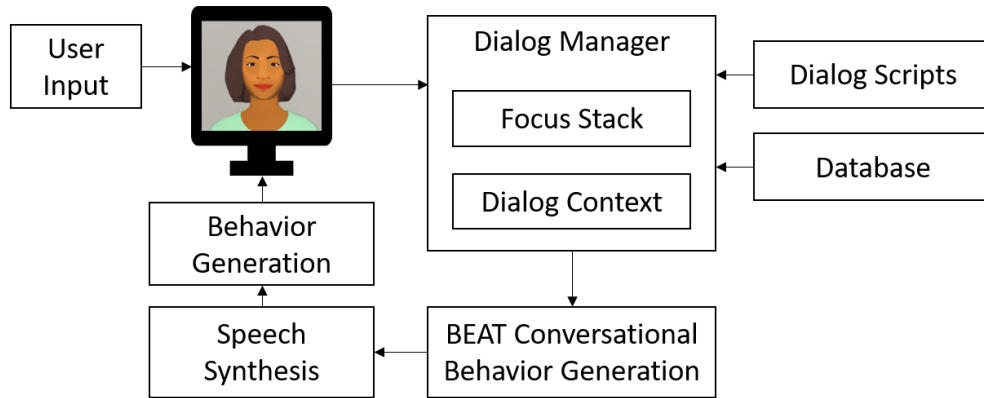


Figure 1. The virtual counselor system architecture used in the pilot study.

The dialog is defined in scripts written in a custom scripting language designed for this dialog management approach. Each script can be considered a discourse segment purpose that is pushed onto or removed from a list implemented as a stack, ensuring that the latest item to be added to the list will also be the next to be removed. Each script is defined as a set of one or more states. Each state must have a name and one or more (1) actions for the virtual agent to perform, (2) display various types of user interface elements, and (3) execute logic for storing variables or performing calculations.

Sub-dialogs (discourse segments) can be pushed into focus based on a variety of factors. For example, the dialog manager can reason about user input from an ongoing session, such as numeric inputs from questionnaire type responses or aggregated input across multiple states, as well as more complex information pulled from long-term storage, such as databases. Language and actions can be generated for the agent using a template-based approach, that is states can define templates where everything from whole utterances to single words can be inserted into a coherent response (set of actions) for the agent to perform. These actions are then sent downstream to a behavior realization module that synthesizes the agent's speech and synchronizes the playback with any nonverbal behavior that may be included in the action set.

The dialog manager is implemented in *C#* as part of the Unity3D client application that users interact with. The custom dialog scripts are converted to JSON files, which are then read and serialized into objects and variables by the dialog manager. Functions and variables implemented in dialog scripts are written in JavaScript, thus the client application includes a framework that maps the script-side calls to its relevant *C#* counterpart in the client application. All calls to external applications, such as databases, are handled by the Unity client.

3.2 Study Design

Participants were recruited at an addiction treatment hospital in Reykjavik (Iceland) and a treatment facility in Boston (Massachusetts), with eligibility criteria of being in medication assisted treatment for opioid use. Following consent, they filled out a socio-demographics questionnaire, conducted a 15 minute conversation with the agent on a laptop computer, filled out a questionnaire about the experience with the agent, and participated in a semi-structured interview. Participants used a constrained interface for input, selecting a response at every turn from a menu of options (Figure 2).

The agent led participants through activities common in CBT-based sessions for relapse prevention, namely emotional recognition and mindfulness with deep breathing [25]. The interaction included standard MI practices, such as reflecting participant choices back to them (based on their menu choice) and using techniques like the ‘readiness ruler’ used to engage clients in a discussion about their readiness to maintain abstinence.

The initial counseling conversation between the agent and patient includes a greeting and some social chat at the beginning of the interaction. After this, the agent leads the patient through two evidence-based skills shown to support recovery efforts among individuals with opioid use

disorder. The interaction closes with the agent asking a few details about the patient's opioid use, to assess comfort discussing this information. In this pilot, we focused on emotion recognition and mindfulness with deep-breathing, followed by an interactive session focusing on emotion regulation. The underlying concept was to have participants go through the motions of how to identify and cope with strong emotions, to fortify their resilience against relapse.



Figure 2. The agent used in the feasibility study, showing a menu of options for the user to choose from at every turn.

3.3 Results

A total of 23 participants successfully completed the study. Their average age was 40.22 years (SD 10.26), ranging from 23 years to 67. 22% were female, most were single, had stable housing, and had not graduated high school. The results showed that the participants were satisfied with the agent, wanted to continue working with her, trusted her, liked her, did not think that she was repetitive, felt it was easy to talk to her, and that she was interesting (Table 3).

Participants also felt that they and the agent understood one another, that the agent was honest about what she thought of them, that they had been honest towards the agent, and that they did not prefer speaking to a human over an agent (or vice versa) about this topic. Participants liked performing the deep breathing exercise with the agent (4 'no' vs. 19 'yes', $\chi^2(1)=9.78, p<.05$)

and most believed that this experience will help them in their recovery (4 ‘no’ vs. 19 ‘yes’, $\chi^2(1)=9.78, p<.05$). Additionally, all 23 participants were willing to self-disclose personal information about their drug use to the agent.

Item	Anchor 1	Anchor 2	Median (IQR) – Wilcoxon
How satisfied are you with the agent?	Not at all	Very satisfied	6.5 (2.5) W=270 p<.05
How willing are you to continue working with the agent?	Not at all	Very willing	5.5 (2) W=225 p<0.05
How much do you trust the agent?	Not at all	Very much	6.5 (2) W=261 p<.05
How much do you like the agent?	Not at all	Very much	7 (1) W=306 p<.05
How repetitive was the agent?	Not at all	Very repetitive	1 (1) W=54 p<.05
How easy was it to talk to the agent?	Not at all	Very easy	7 (0.75) W=297 p<.05
How interesting was the agent?	Not at all	Very interesting	7 (2) W=261 p<.05
How would you characterize your relationship with the agent?	Complete stranger	Close friend	3.5 (2.75) ns.
Do you feel like the agent cares about you?	Not at all	Very much	4.5 (3.25) ns.
Do you feel like you and the agent understand one another?	Not at all	Very much	4.5 (1.75) W=216 p<.05
Was the agent honest about what she thought of you?	Not at all	Very honest	5.5 (3) W=234 p<.05
How close do you feel you and the agent are?	Not at all	Very close	2.5 (2) W=99 p<.05
How honest were you with the agent?	Not at all	Very honest	7 (0) W=324 p<.05
Would you have preferred speaking to a person about this topic?	Preferably a person	Preferably the agent	4 (1.75) ns.

Table 3. The single item measures assessing general agent acceptance on a scale from 1-7. The last column shows whether participants’ ratings were significantly different compared to a neutral rating of 4.

In the semi-structured interview, 50% of participants indicated that they would like to use a virtual agent like ours to support them in their recovery. About 35% said they would definitely use some kind of technology for support. However, 15% said they would never use any kind of technology for treatment support. Patients also suggested that the language of the interactions should be dynamically tailored to how long they've been in treatment and that the system better simulate real conversations, such as being allowed to speak freely as opposed to using the dialog menu options.

3.4 Conclusion

Patients in medication assisted treatment had a generally positive reaction to a virtual agent discussing topics related to opioid use disorder therapy with them. They expressed high levels of trust in the agent and desired to work with her again. Patients participated in the activities the agent asked them to engage in and were willing to self-disclosed sensitive information to the agent. They were also largely satisfied with the overall experience; however, their perceived relational closeness with the agent were low. The interviews revealed that participants felt that one session was too early to talk about having any kind of relationship with the agent. Given the generally positive reactions, in conjunction with positive findings with the VA alcohol counseling agent [153], I conclude that patients in treatment for substance use disorder will accept a virtual agent as a counselor.

Chapter 4. Predicting Counselor Actions

To address my second research question, “Can counselor behavior be modeled directly from annotated counseling sessions?”, I conducted experiments evaluating the performance of models directly predicting the next action for a counselor to make, at any dialog turn, learned automatically from annotated MI counseling session transcripts [97]. Dialog systems typically have a defined set of possible user intents that are inferred by the system to progress the conversation. The model I developed receives counselor and client text as input and then directly outputs the next action the counselor should do. The model can make a prediction at any counselor dialog turn based on the dialog history and context, without explicit definition of intents other dialog state variables.

4.1 Dataset and Experiments

The dataset used in the experiments comprised 164 annotated counseling sessions collected during brief motivational interviewing sessions about alcohol use in an emergency room setting [68]. The annotation was conducted by researchers studying the mechanisms of behavior change and were coded using a variety of schemes, including the Generalized Behavioral Intervention Analysis System (GBIAS) and the Motivational Interviewing Skill Code (MISC) (see section 2.6.2.2) [68]. The MISC was created to measure clinicians’ adherence to using MI and the integrity of how it’s used [90]. Therefore, every counselor utterance in the data set was tagged with an MI-related action.

I conducted two sets of experiments. In the first set, the original MISC labels were adapted to include the seven most important actions for the purposes of conducting a session. These were *asking questions*, *reflecting*, *giving information*, *talking about the session structure*, *facilitation* (inviting the client to continue speaking), *acknowledging client utterances*,

affirmations, and *other* (any other MISC label) (Table 4). The next set included one experiment where the number of labels was reduced to five, to better represent the kinds of actions I wanted the virtual counselor to make, for example, by removing the ‘other’ category. For the 7-label prediction task, I compared the performance of six machine learning algorithms (Table 5) and then used the best performing model for the 5-label task (Figure 7).

Original MISC labels	7 label experiments	5 label experiments
Open question Closed question	Question [qu] - <i>What does a typical week of drinking look like for you?</i>	Question [qu]
Simple reflection Complex reflection	Reflection [ref] - <i>Sounds like you're typically drinking two beers</i>	Reflect [ref]
Giving information Structure	Giving information [gi] - <i>This number shows how many drinks you had</i> Structure [st] - <i>In this first part, we'll talk a little bit</i>	Inform [inf]
Facilitate Filler Acknowledgement	Facilitate [fa] – Mhmm Acknowledgement [ack] - Okay	Ground [gr]
Affirm Emphasize control Support	Affirm [af] – <i>That was a good thing you did</i>	NA
All other MISC labels	Other [o]	NA
NA	NA	Shift [sh]

Table 4. The labels that were used in the experiments, their corresponding MISC label, and example utterance.

The best performing model had two neural networks with long short-term memory cells (LSTM) [60] and a conditional random field (CRF) [73] to produce the final output (Figure 6). At every dialog turn, the counselor and client utterances were fed word-for-word into the model. First, each word was embedded into a 50 dimensional vector. These vectors were initialized by pre-training them on the data set of transcripts using the Word2Vec continuous bag of words (CBOW) approach [88, 113]. Then, each word embedding was passed through a ‘word’ LSTM with hidden layer size of 64. The hidden layer of this ‘word’ LSTM was then used as input to a

second ‘context’ LSTM with an input dimension of size 64 and hidden layer size of 64. The output of the context LSTM was then transformed into a vector of the same size as the number of labels. This vector was then used as features in the CRF. Finally, the CRF made the prediction for an entire sequence jointly (see Appendix B for more implementation details).

All the deep neural network models were implemented using the PyTorch library [100]. Each LSTM was unidirectional, had one layer, and no dropout. To fit the model, the negative log-likelihood loss function and Adam optimization algorithm [72] were used, set to the default hyper-parameters found in the PyTorch implementation: $\eta = 0.001$; $\beta = (0.9, 0.999)$; $\epsilon = 1e - 8$; $L2 = 0$.

4.1.1 Expert Evaluation

In addition to the automatic evaluation described above, MI experts were sought after to rate the best model’s output with respect to two important actions that counselors use to facilitate client involvement, namely *grounding* and *reflecting*. The experts rated actions that the final model predicted given the context leading up to them in spreadsheets containing held out patient-provider sessions that contained the dialog from original patient-provider sessions. They were asked to rate each action using the following 7-point Likert scale items to the degree they agreed or disagreed with the statement: (1) This is an appropriate action to take, with respect to MI; (2) This action is harmful in this counseling context; (3) In the context of MI counseling, this action makes sense; and (4) In the context of this session, this action makes sense.

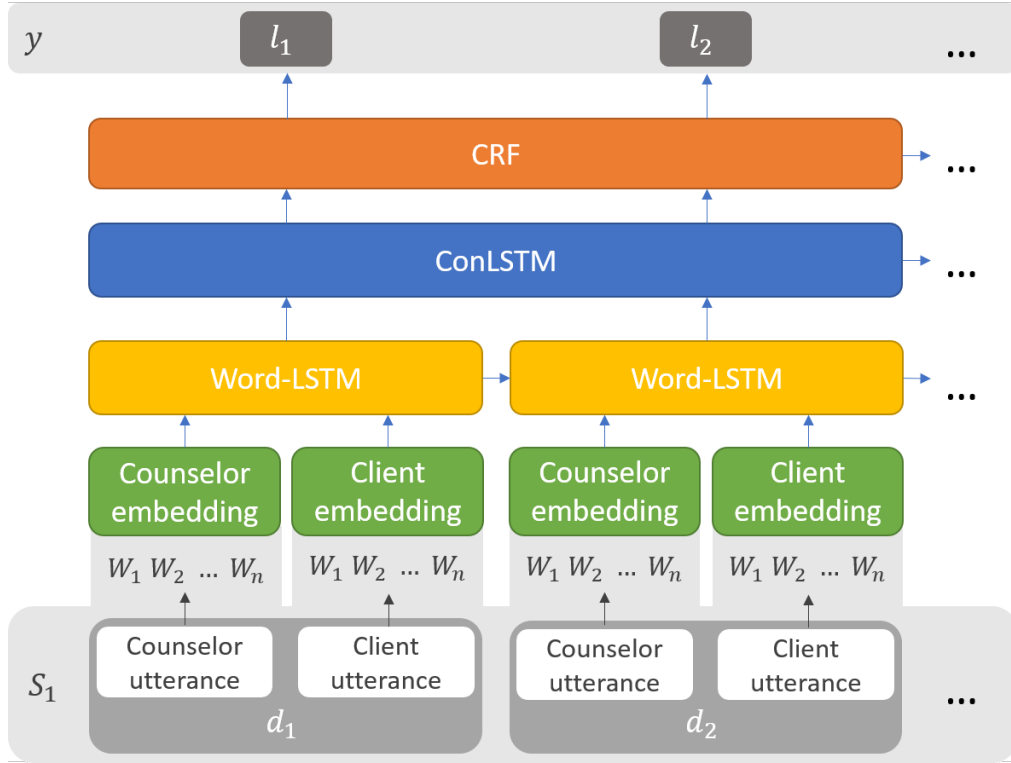


Figure 3. The ConLSTM-CRF model architecture. Each sequence (S) is a session component. Each component contains data samples (d) consisting of a label and patient-provider utterances. The model ultimately outputs a predicted label (l) for each d in the sequence.

4.2 Results

Results are reported as macro-averaged precision, recall, and F1 scores for each model (Table 5).

Precision is the number of true positives (TP) divided by the sum of TP and false positives, recall is TP divided by the sum of TP and false negatives, and F1 is defined as:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

For these metrics, a value of 1 is the best and 0 is the worst. The macro averages show the unweighted mean for each label, which does not take label imbalance into account. The

ConLSTM-CRF and the ConLSTM-CRF + BERT had the best results out of these modeling approaches. Given how closely they performed, the simpler model was chosen moving forward.

Model	F1	Precision	Recall	Accuracy
<i>7 Labels Task</i>				
Majority label only	0.07	0.05	0.14	0.34
Linear SVC + TF-IDF	0.26	0.25	0.28	0.37
CRF + Doc2Vec	0.28	0.26	0.29	0.46
ConLSTM	0.32	0.35	0.32	0.42
BERT-base w/fine-tuning	0.37	0.34	0.37	0.47
ConLSTM-CRF	0.41	0.44	0.4	0.5
ConLSTM-CRF + BERT	0.41	0.44	0.39	0.5
<i>5 Labels Task</i>				
Majority label only	0.11	0.07	0.2	0.37
ConLSTM-CRF	0.62	0.65	0.61	0.59

Table 5. Experimental results using various methods for predicting the next counselor move, given dialog context, and a comparison to predicting the majority label only. The columns show macro-averages.

For the final task, the dialog acts were reorganized to include the types of moves that the virtual counselor should be able to make. ‘Acknowledgments’, ‘fillers’, and ‘facilitations’ were combined into a *ground* action [5] and general ‘giving information’ and ‘session structure’ were combined into a new *inform* action. A new move called *shift* was created and added when the counselor shifts focus to a new component of the MI session. The ConLSTM-CRF architecture was used to train a new model, which had an overall F1 of 0.62, precision of 0.65, recall of 0.61, and average accuracy of 0.59 across the five labels (Table 5). Shifting had the highest F1 score, followed by grounding, then informing, reflecting, and finally questioning (Figure 7).

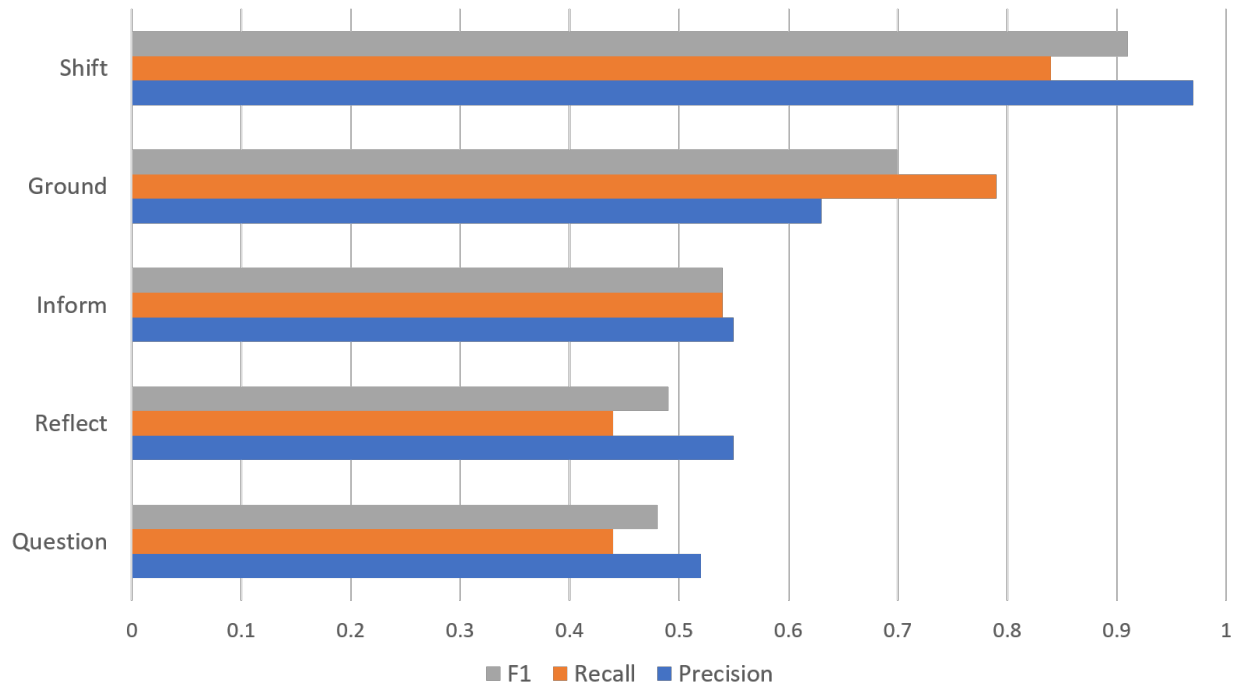


Figure 4. The results, per label, from the evaluation of the ConLSTM-CRF model. All scores are macro averages.

4.2.1 Expert Evaluation Results

Three MI experts were recruited for this task, two having experience using MI for counseling individuals with substance use disorder and one on a variety of health behavior change domains in clinical settings. The experts provided ratings for an average of 57 actions each, predicted by the ConLSTM-CRF model, of two varieties that are key for facilitating MI sessions: *ground* and *reflect* (Figure 8). With respect to the appropriateness of the actions when it comes to MI, the median expert rating for the *ground* action was 5 (IQR=2) and 6 (IQR=2) for *reflect*. Regarding whether the action was found to be harmful in the counseling context, *ground* actions had a median of 1 (IQR=1), as did the and *reflect* actions. The *ground* actions were found to makes sense in the context of MI and the context of the session, each having a median of 6 (IQR=2),

and the *reflect* action was also found to make sense in the context of MI and the session, both having a median of 6 (IQR=2).

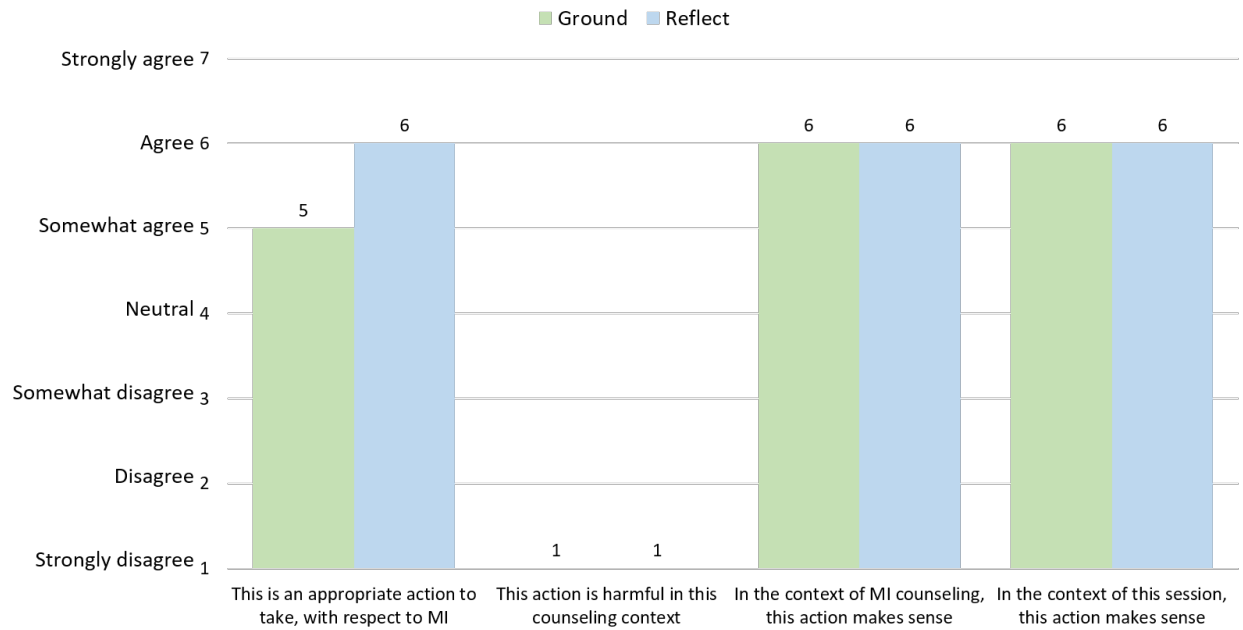


Figure 5. Median expert ratings of the predictions of the ConLSTM-CRF model in the context of original patient-provider counseling sessions.

4.3 Conclusion

To address my second research question, I evaluated models predicting the next counselor action, trained on an existing dataset of patient-provider MI sessions. The best model combined a deep learning algorithm (LSTMs) with sequence modeling (CRF) and managed a reasonable performance for predicting five high-level counseling moves, showing the potential of this approach. The model was trained on annotated patient-provider interaction session transcripts to predict a next counseling move for virtual counseling agents to make, given the counselor and client utterances at any given dialog turn. MI counseling experts rated the capability of the model to predict two important actions for facilitating MI sessions and found it made appropriate, sensible, and non-harmful predictions.

The limitations of this study are that the data used to build and evaluate the models were collected in an emergency room setting and may not generalize to all use cases. The methods used to train the models were limited in scope and may not be the best fitting models for the task. Thus, the models trained may not be transferrable to other domains, though the general approach has the potential to be applicable to other counseling domains, given the appropriate data. Additionally, pretraining BERT [40] on this data might be helpful.

Chapter 5. Automated Reflection Generation

To answer my third research question, “Can neural network-based approaches to natural language generation produce coherent and appropriate language for MI reflections?”, I created models for generating counselor reflections, which are statements that paraphrase a client’s remarks in a neutral form. A transformer language model (GPT-2) was fine-tuned on a dataset of reflections and evaluated conditionally generated statements. Reflections are one of the most fundamental MI counseling techniques and in this work, I trained a neural language generation model that automatically produces language for reflections, given a client utterance and dialog context.

Previous efforts have developed and evaluated natural language processing models for automatically tagging utterances within counseling session dialogs with specific speech act and/or topic codes [83, 97, 140]. Additionally, multiple virtual agent systems that use techniques from MI have been evaluated, most adopting a template-based language generation approach [129, 153]. By contrast, here the task is generating responses to unconstrained user input, to be used by a virtual counseling system. Finally, I conducted an evaluation of the final best-performing model by recruiting MI experts and individuals trained by an expert in MI and asking them to rate the generated utterances on several outcome measures.

5.1 Methods

To train the language generation model, I used transcripts of brief MI sessions in which a counselor discusses alcohol use with a patient in an emergency room setting [68]. These sessions were annotated with the Motivational Interviewing Skill Code, used by investigators to analyze behavior change processes and measure clinician adherence to MI [66]. Every counselor utterance was therefore coded with reflections (among other codes), and each session was split

into components, i.e., discourse segments with defined therapeutic goals and set of counseling techniques. This enabled to conditioning the model on the counselor's action (*reflect*), the user's language, and the current component (Figure 9).

The intuition for this task is that a language model learns the probability distribution of a sequence of a fixed set of symbols, e.g., words, given the previous symbols. With x denoting a sequence of words, this distribution can be represented as follows:

$$p(x) = \prod_{i=1}^n p(x_i | x_{<i})$$

In the conditional approach, the model learns the probability of the sequence of words given both the previous words and other variables called 'control codes' (e.g., a and c) [70]:

$$p(x|a, c) = \prod_{i=1}^n p(x_i | x_{<i}, a, c)$$

The pre-trained English GPT-2 model [111] was fine-tuned and the reflection generation model trained to be conditioned on the context, counselor action (a), and the session component (c). GPT-2 is a 12 layer transformer that induces 768-dimensional hidden token representations from 12 self-attention heads. In all, the model has 117 million parameters. The Transformers library [145], written in PyTorch [100] was used for this work.

Training data from 132 transcripts was created by extracting every client utterance leading up to a reflection, adding the conditioning elements (the counselor action and session component), and lastly the counselor's reflection (as a target). This yielded a total of 12,830 data samples for training. For every transcript there were an average of 97 data samples (reflections). The samples had an average context length of 54.66 (SD=89.4) words and reflection utterance length of 16.1 (SD=14.22) words.

The model was trained conditioned on the context, counselor action, and the session component. Training took place for three epochs without early stopping, fit using the Adam optimization algorithm [72] ($\eta=1e-5$; $\epsilon=1e-8$; $L2=0$), and the default parameters for fine-tuning in the Hugging Face implementation.⁸ Settings for language generation were top-k with $k=40$, temperature of 0.7, a repetition penalty of 1, and maximum output length of 40.

```
<|startoftext|> Mhm. That sounds fine. A typical week of drinking Well, I don't really have any typical weeks of drinking, but like today, I'm going to share a six pack with the downstairs neighbor That's about it I'm supposed to play cards on Thursday so I'll probably drink a 12pack. Other than that, I don't really plan around drinking or anything so I don't really have a typical week of drinking or anything.<reflect><2.2>So there's nothing ever really planned as far as how much you drink or what it's like<|endoftext|>
```

Figure 6. Example excerpt from the training data, conditioning on the dialog context, counselor action, and session component.

5.2 Evaluation

Two evaluations of the reflections generated by the system were conducted: first by judges trained by an expert in MI and in the second by MI experts. In the first evaluation, the trained judges received training from an individual with experience using MI in clinical research with patients diagnosed with substance use disorders. They were presented with example scenarios from the corpus of actual counseling sessions, and asked to rate (1) the generated reflections in that context (GENERATED), (2) the reflections used by the human counselor in the transcript (ORIGINAL), and (3) generic reflections generated from a simple rule-based system (RULE-BASED). RULE-BASED reflections were generated by concatenating an expression of

⁸ <https://github.com/huggingface/transformers>

understanding (e.g., “I see”, “I understand”) with a conversational directive (e.g., “let’s focus on that”, “please continue”), each selected at random.

For the second evaluation, individuals 18 years or older that had used MI in a professional capacity were sought after to rate reflections in the GENERATED and ORIGINAL contexts.

5.2.1 Measures

The trained judges rated utterance in context, determining: (1) whether the utterance met the criteria for being a reflection; (2) whether the utterance was grammatically correct; (3) whether the utterance was coherent within the context, and (4) shows an understanding of what the client said; (5) adds emphasis or marks an important or intense client emotion; (6) adds meaning to what the client said; and (7) facilitates the client-clinician interaction. Ratings were made using a 3-point Likert scale from “Disagree”, “Neutral” to “Agree”. Utterances that received a rating of “Agree” on items 4, 5, or 6 were considered reflections.

The experts rated the utterances in context using the following 7-point Likert scale items: (1) This is an appropriate utterance to say, with respect to MI; (2) This utterance is harmful in this counseling context; (3) In the context of MI counseling, this utterance makes sense; (4) In the context of this session, this utterance makes sense; (5) This utterance is coherent English; and (6) This utterance is coherent, given the context.

5.2.2 Procedure

An expert in MI trained two judges in recognizing reflections by given them a lecture on MI, its techniques, goals, how reflections are used, and what they look like, as well as giving them exercises to practice identifying them. The judges evaluated the counselor utterances in context and final inter-rater reliability on 48 samples was adequate ($\kappa=.78$). Test scenarios were

extracted from six counseling sessions in the corpus that were held out during model training, and in which the human counselor produced a reflection.

For each scenario, judges and experts were provided with the sequence of all counselor and client utterances preceding the reflection, followed by either a GENERATED, ORIGINAL, or RULE-BASED (judges only) reflection, selected at random. Judges then provided their ratings on the seven 3-point scales and experts on the 7-point scales. They were informed that the utterances they were rating had been generated in a variety of different ways but were not told that some of them were the original counselor reflections.

5.2.3 Results – Trained Judges

Two individuals were trained by a clinical researcher with experience using motivational interviewing for counseling with patients diagnosed with substance use disorders. The results from the ratings of these judges are summarized in Figure 10, which shows the proportion of items that were rated “Agree” on a scale from “Disagree”, “Neutral”, to “Agree”. 71.5% of the GENERATED utterances met the criteria for reflections, 91% of the ORIGINAL reflections, and only 3.6% of the RULE-BASED utterances. There was a significant difference between all three types of utterances with respect to demonstrating an understanding of what the client said (88% of the ORIGINAL reflections, 64.2% of the GENERATED type, and 3.6% of the RULE-BASED utterances; $2(2)=266.05, p<.05$).

Similarly, there were significant differences between the types of utterances regarding whether they added emphasis or marked an important or intense client emotion (34.3% of the ORIGINAL reflections, 24.2% of the GENERATED ones, and none of the RULE-BASED utterances; $2(2)=68.01, p<.05$). Regarding whether utterances added meaning to the client’s remarks, there were significant differences between the RULE-BASED reflections (0%) and the

other two ($2(1)=57.34$, $p<.05$), but not between the GENERATED (23.6%) and ORIGINALS (28.3%).

Finally, there were significant differences between all three types on whether they facilitated the client-clinician interaction ($2(2)=20.94$, $p<.05$). The RULE-BASED utterances had a ratio (99.4%) significantly higher than both the ORIGINAL (88.6%) and the GENERATED reflections (86%). The difference between the GENERATED and ORIGINAL reflections was not statistically significant.

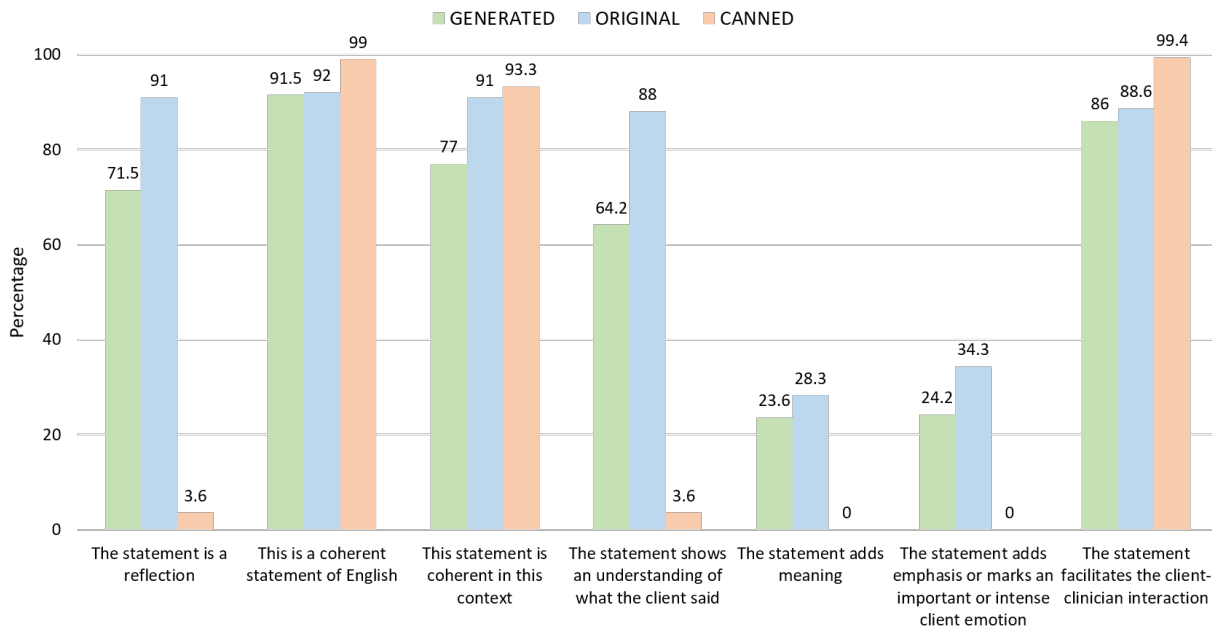


Figure 7. The items that the human judges used for rating. All items were on a 3-point scale with the anchors “Disagree” and “Agree”. The numbers indicate the proportion of utterances that received a rating of “Agree”.

5.2.4 Results – MI Experts

Three individuals with experience using MI in a professional capacity were recruited. Two had experience using MI when counseling individuals with substance use disorder and one on a variety of health behavior change domains in clinical settings. The experts rated an average of 40

reflections each and the total number of ratings were 120, with a 50/50 split between the GENERATED and ORIGINAL varieties.

Each reflection was rated on six 7-point Likert scale items. Figure 11 shows the values of the measures and median rating of each item, for the GENERATED and ORIGINAL reflections. For the statement “This is an appropriate utterance to say, with respect to MI”, the GENERATED reflections had a median rating of 4 (IQR=3.25) and the ORIGINAL reflections had a median rating of 6 (IQR=2). The statement “This utterance is harmful in this counseling context” the GENERATED reflections got a median rating of 2 (IQR=3) and the ORIGINAL ones had a median of 1 (IQR=1). For the third statement “In the context of MI counseling, this utterance makes sense”, the GENERATED reflections had a median of 4 (IQR=3) and the ORIGINAL reflections a median of 6 (IQR=2). For the fourth statement “In the context of this session, this utterance makes sense”, the GENERATED reflections had a median of 4 (IQR=4) and the ORIGINAL utterances a median of 6 (IQR=1.25). For the item “This utterance is coherent English” the GENERATED reflections were rated at a median of 6 (IQR=1), as were the ORIGINAL reflections (IQR=1). Lastly, for the item “This utterance is coherent, given the context” the GENERATED reflections had a median of 4 (IQR=4) and the ORIGINAL utterances had a median of 6 (IQR=6).

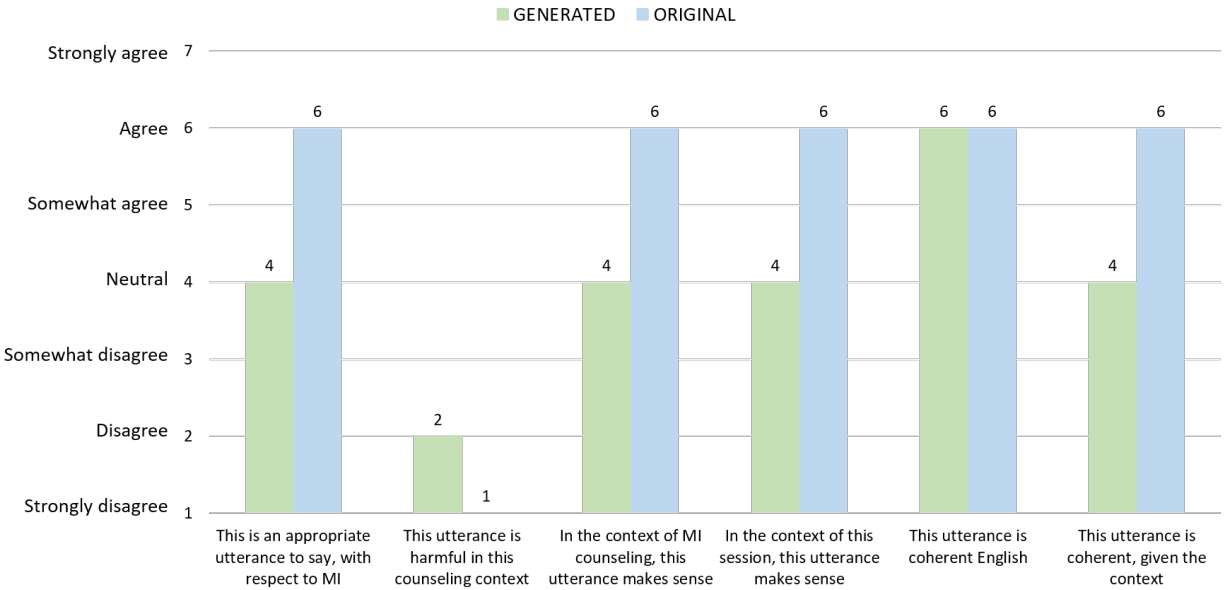


Figure 8. The median scores from the expert ratings of the GENERATED and ORIGINAL MI counseling reflections.

5.3 Discussion

The trained judges found that a majority of the GENERATED utterances could be considered reflections, most were rated as being coherent English and contextually coherent, as well as facilitating the interaction. Further, a majority of the GENERATED reflections were deemed to show an understanding of the client’s remarks, meeting the minimum requirements for “simple reflections” in MI [89]. The ORIGINAL counselor reflections were rated significantly higher than the ones that were GENERATED, except on whether they were coherent English, added meaning to the client utterance, and facilitated the interaction between the client and clinician.

The MI experts did not find the GENERATED reflections harmful and felt that they were coherent sentences of English. They were neutral with respect to the GENERATED reflections being appropriate MI utterances and whether they made sense in the context of MI and the session. Conversely, the experts agree that the ORIGINAL counselor’s reflections are appropriate for MI and are coherent and sensible given the context of MI and the session.

5.4 Conclusion

Seeking to answer my third research question, I proposed and evaluated an approach to automatically generating counselor reflections in MI-based substance use counseling sessions, given unconstrained client input and dialogue context.

Judges trained by an expert in MI found that a majority of the statements generated by the fine-tuned transformer language model met the criteria for being reflections; however, the MI experts were ambivalent regarding the appropriateness and contextual coherence of the generated reflections, with respect to MI. Nevertheless, experts found the generated reflections to be harmless and both the judges and experts felt that the language generated was coherent English.

More complex reflections add meaning and emphasize important emotions in the client's remarks that may change a person's understanding or feeling about their situation [89]. The judges' ratings indicate that the generated reflections seem to do this at a similar (though lower) rate as the original counselor's reflections. However, the model's ability to generate complex reflections needs further investigation. Performing simple or complex reflections in a particular sequence and at key moments is known to promote client behavior change [74]. Therefore, the ability to generate either type deliberately could be a powerful ability for a virtual counselor. Reflections that are generated using this method seem to be an improvement over the rule-based approach, where statements are generated without taking the dialog context into account.

5.4.1 Limitations

This work has two main limitations. First, the dataset is relatively small and future work would benefit from a larger variety of counselor reflections. Second, the evaluation involved only five raters and would be improved by a larger, controlled experiment.

Another issue that was not addressed here is the inherent risk that using language generation in a clinical setting poses. For example, one would want to be careful that certain patient utterances ('triggers') do not induce inappropriate responses [142]. This issue could be explored in future work.

Chapter 6. Technical Approach to Hybrid Dialog Management

I developed and evaluated a hybrid spoken dialog system for conducting CBT-MI counseling sessions, capable of simulating face-to-face conversations using an Embodied Conversational Agent (ECA) that provides the auditory and visual cues conveying intelligence and human-likeness to users [33]. The system supports speech input to allow users to express themselves and influence the flow of the conversation, as clients do in client-centered counseling, such as MI and CBT sessions. To support unconstrained client speech, I used modern machine learning-based NLU techniques, specifically neural network models, on a corpus of annotated MI sessions between patients and providers used for a variety of purposes, such as predicting the counselor's next action or using the dialog history and context to generate the counselor's next utterance.

From the client's perspective, the counseling session is open-ended and unstructured; however, in practice, brief MI and CBT interventions are typically manualized, and counselors move through sessions one section at a time covering prescribed material while allowing clients space for unconstrained talk. From a systems perspective, each of these sections can thus be implemented as objects that can be reasoned about.

In order to develop an automated counselor that is able to manage both the overall agenda-driven structure of a CBT-MI session, as well as unconstrained client speech, I used a hybrid dialog management approach that sequences through a structured session agenda, while giving users opportunities to freely express themselves. The following sections describe this general architecture in the context of a specific application, namely, one that has the goal of getting college students with mild to moderate Alcohol Use Disorder to consider decreasing their alcohol consumption.

6.1 System Architecture and Processes

The dialog manager conducts counseling sessions and promotes behavior change by using techniques from CBT and MI, while adhering to a well-defined session structure (Figure 3). Though these techniques are general, the sessions are designed based on a manualized intervention protocol [101] that includes pre-defined sections covering topics such as understanding the client situation, giving the client personalized feedback, assessing client motivation, and creating a plan for change.

The **Input Layer** manages the user's input by capturing their speech audio and relaying it to the dialog manager for further real-time processing. It also manages additional information that might be available from or about the user, such as questionnaire responses collected prior to the session. The raw audio is processed continuously for automatic speech recognition and pause detection, while the survey data is used at various points in the session, such as in the 'PersonalizedFeedback' component (Figure 4).

The **Interaction Layer** serves to maintain the channel of communication between the agent and user. For example, the general rule for turn-taking [125] is to limit the gaps and overlap between user and agent by taking speech pauses into account.

In the **Counseling Layer** are the system's core components for dialog management. The dialog manager reads dialog scripts, maintains the dialog history and current context, handles natural language understanding, and generates the language and nonverbal behavior for the virtual agent counselor to perform. This involves three interconnected processes: (1) natural language understanding, (2) action prediction, and (3) language generation. These processes require the current user input, the dialog history and context, and the survey data to reason and take action. Additionally, this process automatically generates the virtual agent's conversational

nonverbal behaviors using the Behavior Expression Annotation Toolkit [30], such as eyebrow movement, gazes, posture shifts, and hand gestures, as well as lip-syncing.

Lastly, in the **Output Layer** the generated language is sent to a speech synthesizer and the on-screen virtual agent performs the nonverbal behaviors in sync with the speech audio.

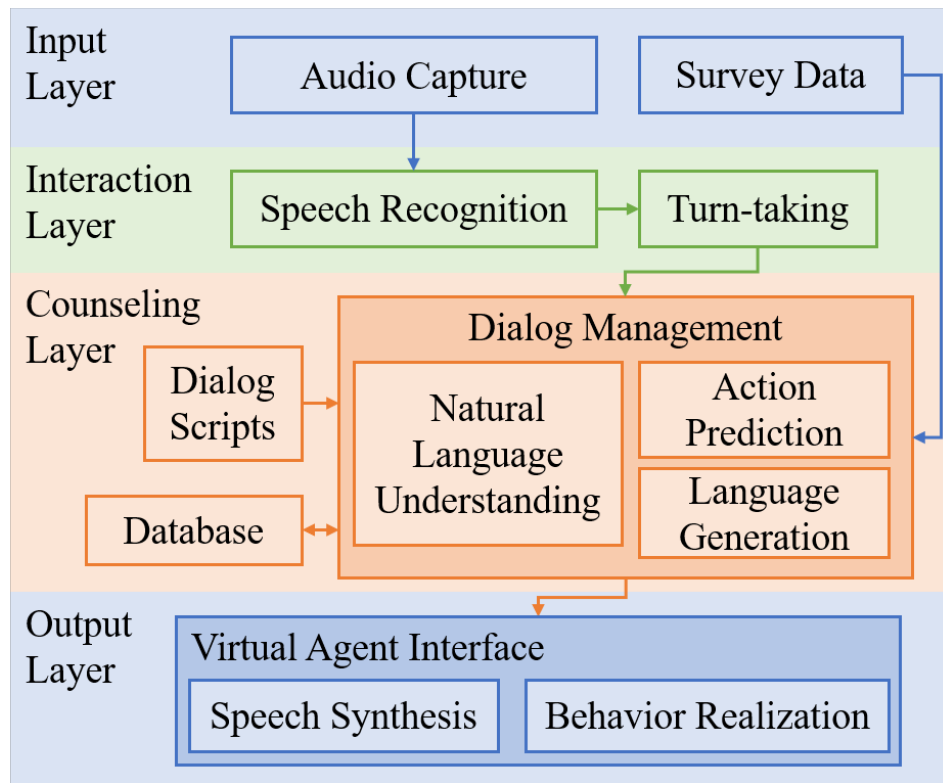


Figure 9. The counseling agent system architecture has four layers, each responsible for distinct processes and tasks.

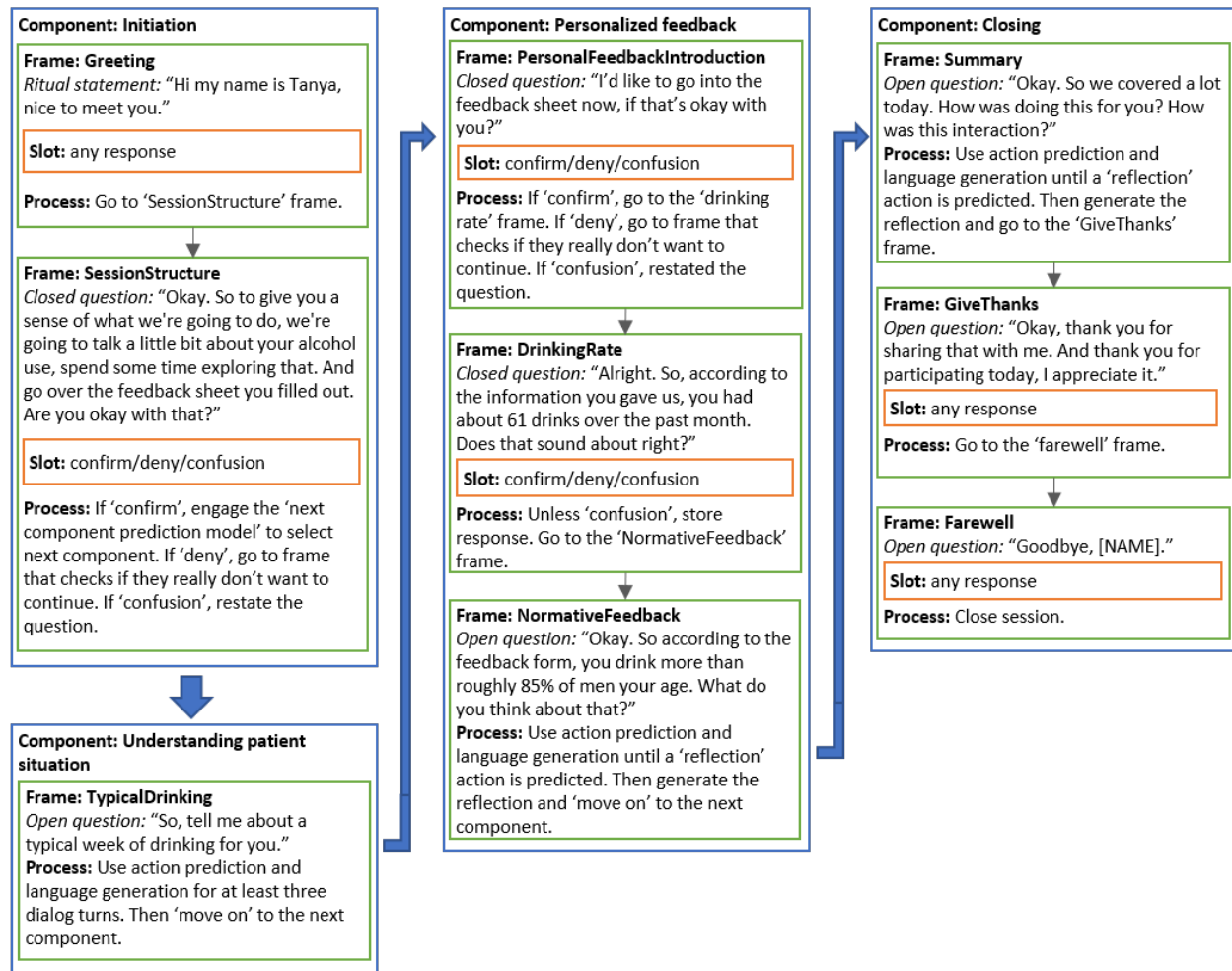


Figure 10. An example completed session comprised of a series of components and frames that specify the structure and context for input processing. The flow from one component or frame to the next is indicated with arrows.

6.2 Hybrid Dialog Management

The hybrid spoken dialog system uses a dialog manager (DM) that merges a frame-based dialog system approach with using neural network models to drive the dialog. The DM has two modes: structured and neural. In the structured mode, the system uses a frame-based dialog engine architecture commonly used for task-based dialog systems that trace their origins to the GUS system [21], which underlies most contemporary commercial dialog systems. Frame-based systems define one or more objects called **frames** that have **slots**, or inputs, that can take

particular **values**. Dialog flow control is also specified by the frame and are determined by conditions and rules. For example, a transition to a new frame could occur once all the slots are filled or when a slot has a particular value. Natural language understanding involves extracting the **intent** and fillers for slots from the user's utterance. An intent is the task or goal that a user is trying to accomplish with their words, such as *greet*, *request clarification*, or *buy a movie ticket*. In Figure 4, for example, the SessionStructure frame is expecting user input that can be mapped to one of three intents (*confirm*, *deny*, and *confusion*) and has a dialog flow rule where the transition to another frame is conditioned on the matched intent.

This is similar to the frame-based dialog design formalism VoiceXML⁹; however, instead of using non-probabilistic grammar-based language models and rule-based semantic parsers to map user speech to an intent, the current system uses machine learning to train probabilistic models. For example, the intent to *confirm* is defined as a set of {'yeah', 'yes', 'okay', 'sure', 'fine'} and probabilistic intent matching, using neural networks with word embeddings as input features. With this method, user utterances such as "sure, alright", "that's fine", or "yeah sure" will all match the *confirm* intent. The current implementation of the system interfaces with the IBM Watson Assistant¹⁰ by sending the user utterance and receiving a list of intents with associated confidence values. The dialog manager then selects the most appropriate intent based on the confidence level and contextual factors, such as the session component, dialog state, and previous intents.

In the neural mode, the behavior of the counselor is emergent, in that it is determined at runtime by neural network models that predict actions and language based on dialog context. The dialog flow is directly influenced by the user's utterances, as opposed to conditions and rules.

⁹ <http://www.voicexml.org/>

¹⁰ <https://www.ibm.com/cloud/watson-assistant>

For example, in the TypicalDrinking frame in Figure 4, the agent makes an open elicitation such as “Tell me about a typical week of drinking”. Then, for every subsequent turn of talk, instead of relying on mapping user utterances to expected intents, the DM uses a model of counselor behaviors to predict the next action the virtual counselor should take. There is no limit to the number of actions that the system could predict and act on, it only depends on the nature of the action prediction model. The current implementation of the system uses two actions: *reflecting* and *grounding*, which influence the language and non-verbal behavior generation modules in different ways. For example, predicting a *reflect* action informs the language generation module to generate a reflection, given the current user utterance, session component, and dialog context; a *ground* action tells the language and behavior generation module to produce a simple grounding move, such as saying the agent saying “okay”, doing a head nod, and inviting the user to elaborate on their previous utterance with a phrase such as “tell me more about that”. Both the action prediction and language generation processes use models trained on patient-provider MI session data using deep neural networks (see chapters 4 and 5).

The language and nonverbal behavior of the virtual counselor is generated based on the current action, session component, and dialog context. The NLG models are created by fine-tuning large language models on our patient-provider alcohol counseling session data [40] (details in chapter 5). The NLG model could be trained to generate utterances conditioned on several counselor actions and contextual variables. For example, a similar approach could be taken to create an *open question* generator that produces a question that the counselor should ask the user whenever the action model predicts an *open question* action. In the current implementation of the system, the main task of the language generator is to produce reflections for the agent whenever the action prediction process outputs a *reflect* action.

The DM switches between the two modes depending on how the flow of the conversation should be controlled. This can be specified a priori by the author of the counseling session or controlled at runtime by other system components based arbitrary logic. In the structured mode the frames have full control and the DM solely relies on the frame mechanism described in the previous paragraphs. In Figure 4, the SessionComponents module contains the structure of all sessions the DM can manage, which in this case is alcohol counseling for college students [101]. A session **component** is a predefined segment of the counseling session that consists of one or more frames and is analogous to a discourse segment depicting a stretch of dialog associated with a task goal [53]. Therefore, a session is defined as a set of components, four of which are depicted in Figure 4. In the neural mode, the neural networks have full control and make decisions about the actions of the counselor based on dialog context. The decision to switch back to the structured mode can be left up to the neural networks themselves, for example, by training them to predict when to switch to a new session component, or it can be specified by rules, such as switching after a set number of dialog turns or when a particular keyword is detected.

This merger of a frame-based dialog system and neural network models enables script writers to design interactions that have an overall structure, such as in counseling sessions, and embed sections where conversation emerges that does not have a prescribed flow or content without worrying about the potential safety concerns that come with losing the general context. If the models that drive the unstructured sections of dialog are found to be not harmful and fit the context of the discussion, the likelihood of the emergence of unsafe discussions is low. The current implementation has a strategies for handling negative evidence of grounding, or misunderstandings, between the user and agent. In the case where users express confusion, namely when the NLU module detects a *confusion* intent, the virtual counselor can repeat what

was said. However, the dialog script author can decide exactly how the repetition is implemented, for example, they may opt to repeat verbatim the agent's last utterance, have variants for particular dialog states, or call a separate custom repetition language generation system. In the case where the agent cannot discern the user's intention, the current system implements Bohus and Rudnicky's most successful dialog recovery strategy: *move on* [24]. Any dialog state can specify a *move on* state for the DM goes to whenever the NLU module cannot successfully infer a user's intent. In the current implemented system, the agent apologizes to the user and asks them to repeat themselves. The DM then increments a counter and if the system fails to discern the user's intent a second time, the DM goes to the designated *move on* state. The precise implementation of this strategy in the hands of the dialog script authors.

The hybrid spoken dialog system architecture combines neural networks for natural language processing and generation with a structured dialog management approach, allowing users to influence the direction of the dialog using unconstrained speech, while ensuring that the system maintains the trajectory of the counseling session as a whole.

6.3 Scripting Language and Extensions

The dialog content and flow are defined using the custom scripting language described in section 3.1. All possible conversations are defined by a set of scripts that are themselves defined by a set of states. A state may contain the atomic actions performed by the agent, commands to instantiate user interface (UI) elements, and any logic that should be executed (Figure 5).

The custom scripting language was extended to enable hybrid spoken dialog management by introducing the ability to script customizable UI elements, such as *intents* and specifying the mode as either *structured* or *neural*. A markup language based on the Extensible Markup

Language (XML)¹¹ text format was developed to allow script writers to specify layouts and widgets in a straightforward manner using text, as opposed to building all UI panel objects in the client application. At runtime, the client reads the custom UI specification, parses the elements (tags and attributes), and instantiates the relevant objects. This includes conventional UI objects such as full screen panels, buttons, images, and forms, as well as intents for the speech-input interface.

In the structured mode, the DM relays user speech to the NLU API which returns a set of matched intents and confidence scores. The DM reasons about the intents it receives and attempts to match them to one of the expected intents for that particular dialog state (Figure 5). If no match is found, the system engages the *move on* strategy described in the previous section to continue the session. For example, in the first and last states in Figure 5, the DM is set to *move on* to a specified state if it cannot match an intent. If a match is found, the DM executes the functions specified in the intent's *onmatch* attribute. The last state in Figure 5 shows how the DM is switched to the neural mode. The current session component is set to make sure the models have the proper context and the models then predict counselor actions and generate counselor language. In this particular implementation, the DM was configured to switch back to the structured mode after three turns of dialog.

¹¹ <https://www.w3.org/XML/>

STATE: PURPOSE2

AGENT: Okay. I am not going to tell you what to do about your drinking. Instead, I will provide you with some information, and some suggestions for you to consider. But what you decide to do with it, is entirely up to you. You know yourself best, and only you are responsible for the decisions you make. How does that sound?

ACTION: \$SET("MOVEON_STATE", "TYPICAL_DRINKING_TRANSIT");\$

CUSTOMUI:

```
<intents>
  <dialog type="structured"/>
  <intent onmatch="GO('TYPICAL_DRINKING_TRANSIT');">acknowledgement</intent>
  <intent onmatch="GO('NOT_OK');">negative</intent>
  <intent onmatch="GO('PURPOSE2');">confusion</intent>
</intents>
```

STATE: TYPICAL_DRINKING_TRANSIT

AGENT: Alright [NAME].

ACTION: \$GO("TYPICAL_DRINKING");\$

STATE: TYPICAL_DRINKING

AGENT: So, tell me about a typical week of drinking for you.

ACTION: \$SET("COMP", "<2.2>"); SET("MOVEON_STATE", "DRINKING_PATTERNS");\$

CUSTOMUI:

```
<intents>
  <intent onmatch="GO('TYPICAL_DRINKING_CLARIFY');">confusion</intent>
  <dialog type="neural"/>
</intents>
```

Figure 11. An example dialog script that can be run by the current implemented virtual counseling system.

Chapter 7. Domain Specific Application

The virtual counseling system was designed to conduct a single session based on a manualized intervention for college students with comorbid depressive and binge drinking symptoms [101].

The intervention is an eight week protocol that combines CBT and BMI and designed to meet the needs of this population. The treatment manual places an emphasis on the importance of addressing depression and heavy drinking together, since depressive symptoms have been found to exacerbate heavy drinking [47]; however, in the scope of this project, only content related to alcohol use was implemented.

The treatment protocol has eight sessions [102]. The first two sessions include brief motivational interviewing and focus on addressing heavy alcohol use followed by six sessions of CBT for depression that focus on strategies to reduce depressive symptoms. The protocol begins with MI techniques to address the ambivalence some patients may feel about changing their alcohol use and about being in treatment. The treatment also begins with reviewing feedback that patients have provided on their alcohol use, comparing this to college students' drinking norms, and discussing negative consequences related to alcohol use. Then issues of safety are reviewed and how the negative consequences with respect to drinking may be reduced. The depressive symptoms associated with drinking alcohol are then discussed. The first session ends with asking the patient if they are willing to consider drinking less alcohol in the coming week and how much. The next session takes place a week later and begins by reviewing the patient's self-reported alcohol use over the past week and discuss strategies that may help them reach their goal. The next six sessions address depressive symptoms using CBT strategies developed for the treatment of depression and MI techniques are continually used throughout these sessions when the patient expresses ambivalent attitudes to changing their behavior. The sixth session

emphasizes engaging in pleasant activities to mitigate negative mood and teaches techniques for managing anxiety. In the final session, the clinician and patient review the progress that has been made towards the patient's goals and strategies for maintaining the newly formed behaviors are put in place.

7.1 Overview of the Implemented Counseling Session

The script used by the virtual counselor in the final implemented system was based on content from the first and second session in the treatment protocol. Prior to starting the session, users fill out a personal feedback form that asks them questions regarding their drinking patterns that the virtual counselor will use during the session (see section 8.3 for details). The session begins with the virtual counselor introducing herself and asking the user how they are doing. Then she engages in small-talk to build rapport [13] asking what they think about their university or college. The agent also mentions the COVID-19 global pandemic and asks whether they were negatively impacted by it. Following negative responses in this discussion, the virtual counselor gives empathic feedback by showing a facial expression of concern and saying an utterance such as "I'm sorry to hear that".

Following this initial introduction and rapport building, the virtual counselor sets the agenda for the session. She explains that the session is based on a manualized intervention for alcohol counseling and asks if they are willing to continue the session. If users are not willing to continue, she would ask them to reconsider and if they decline the session would come to an end. She explains that she is not there to tell them what to do but only show them information that may be useful to them. Then she asks them to tell her about a typical week of drinking for them, giving the users an opportunity to express themselves freely. During this portion of the dialog, the virtual counseling system activates the neural mode of the hybrid dialog system. It uses the

action prediction model to decide whether to reflect or ground and facilitate the conversation. In the case of a reflection, the language generation model produces the response of the agent given the dialog context.

Next, the virtual counselor moves on to discuss the information provided in the personalized feedback form. With their permission, the agent presents them with the information they provided before starting the session and based on that calculates how many drinks they consumed over the course of a typical week. This gives users the opportunity to reflect on their drinking patterns using information tailored specifically to them.


Following these reflections, the agent calculates the user's percentile rank in terms of alcohol consumption and compares them to national percentiles based on numbers from a study of the alcohol and drug use of college students across the United States [87]. The agent then asks the user about their thoughts on hearing their rank and mentions that their level of drinking may be due to being normalized by their situation or social circle and uses the neural mode to process the open-ended user input.

Next, the agent talks about heavy drinking days and talks about the information that the user supplied about the amount of drinking they do at those times. Then she asks them an open ended question on what their thoughts are about that and activates the neural mode to handle the open-ended responses.

In the next section the agent covers the topic of blood alcohol concentration (BAC). She provides a definition of the term and shows and talks about a graph with different levels of BAC depending on a variety of factors. She makes the point that the more someone drinks, the higher their BAC and that this may impact their behavior. She then shows charts listing behaviors, from benign to troublesome, and the association between them and BAC (Figure 12). Then the user is

asked whether they have experienced any of the negative effects of alcohol listed on the chart.

The agent then talks about tolerance to alcohol, provides a definition and what some of the effects of tolerance are on behavior over time.



Progressive Effects of Alcohol		
BAC (%)	Behavior	Impairment
0.01–0.029	<ul style="list-style-type: none"> Average individual appears normal 	<ul style="list-style-type: none"> Subtle effects that can be detected with special tests
0.03–0.059	<ul style="list-style-type: none"> Mild euphoria Sense of well-being Relaxation Talkativeness Joyous Decreased inhibition 	<ul style="list-style-type: none"> Alertness Judgment Coordination Concentration
0.06–0.10	<ul style="list-style-type: none"> Blunted Feelings Disinhibition Extroversion Impaired Sexual Pleasure 	<ul style="list-style-type: none"> Reflexes Reasoning Depth Perception Distance Acuity Peripheral Vision Glare Recovery
0.11–0.20	<ul style="list-style-type: none"> Over-Expression Emotional Swings Angry or Sad Boisterous 	<ul style="list-style-type: none"> Reaction Time Gross Motor Control Staggering Slurred Speech
0.21–0.29	<ul style="list-style-type: none"> Stupor Loss of Understanding Impaired Sensations 	<ul style="list-style-type: none"> Severe Motor Impairment Loss of Consciousness Memory Blackout
0.30–0.39	<ul style="list-style-type: none"> Severe Depression Unconsciousness Death Possible 	<ul style="list-style-type: none"> Bladder Function Breathing Heart Rate
>0.40	<ul style="list-style-type: none"> Unconsciousness Death 	<ul style="list-style-type: none"> Breathing Heart Rate

Figure 12. The negative effects of alcohol on behavior increase and physical and mental impairment occurs as blood alcohol content rises.

Next, the agent introduces the biphasic effect of alcohol, that is the high and low that alcohol produces and how the depressant effects can deepen with heavier use. She shows charts

accompanying this topic and highlights the difference in the effect on the high positive feelings compared to the low depressant feelings when tolerance to alcohol builds up (Figure 13). The agent then makes a point of saying that it does not take long for the body to lose its tolerance to alcohol.

The Biphasic Response to Alcohol

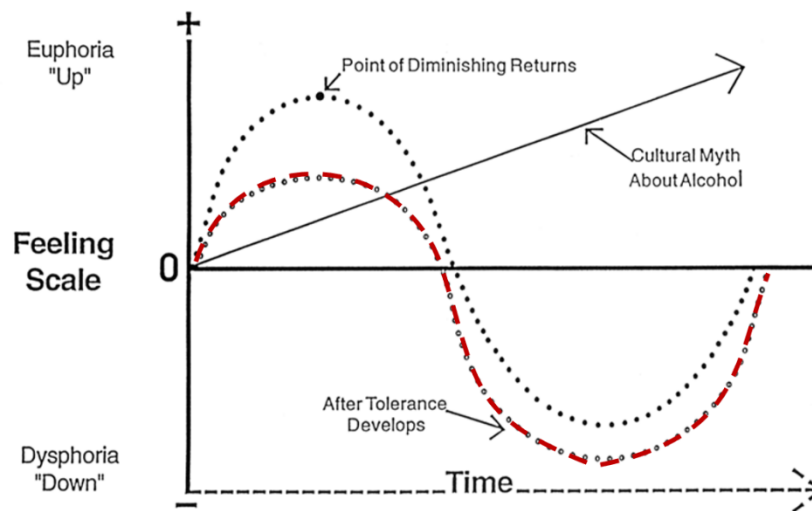


Figure 13. Positive feelings are experienced early on in a drinking session. As more is consumed, the drinker experiences more negative feelings. Tolerance to alcohol develops over time and dampens positive feelings, while deepening the depressant feelings

Next, the agent reviews the negative effects of alcohol that the user selected on the personal feedback form before starting the session. The user is asked to select the one negative effect that stands out the most to them and the agent activates the neural mode of the dialog system to ask them to share their thoughts on the negative consequence they selected. Then she shows a triangle shaped scale chart, where the top scale represents alcohol consumption (none to heavy) and the bottom one represents alcohol related consequences (mild to severe) (Figure 14). She makes the point that the majority of people fill the area on the left (larger) side of the triangle, closest to the mild consequences and ‘none’ or ‘little’ end of the scales, and that relatively few

people are at the severe and 'heavy' end of the scales. Users are then asked to place themselves on that scale from 1-10, where 1 is no consumption and 10 is heavy. Then she points out that no matter where they are on the scale, they can always move more towards the lighter end of the scale and if they do so they'll decrease the risk of coming to harm while drinking.

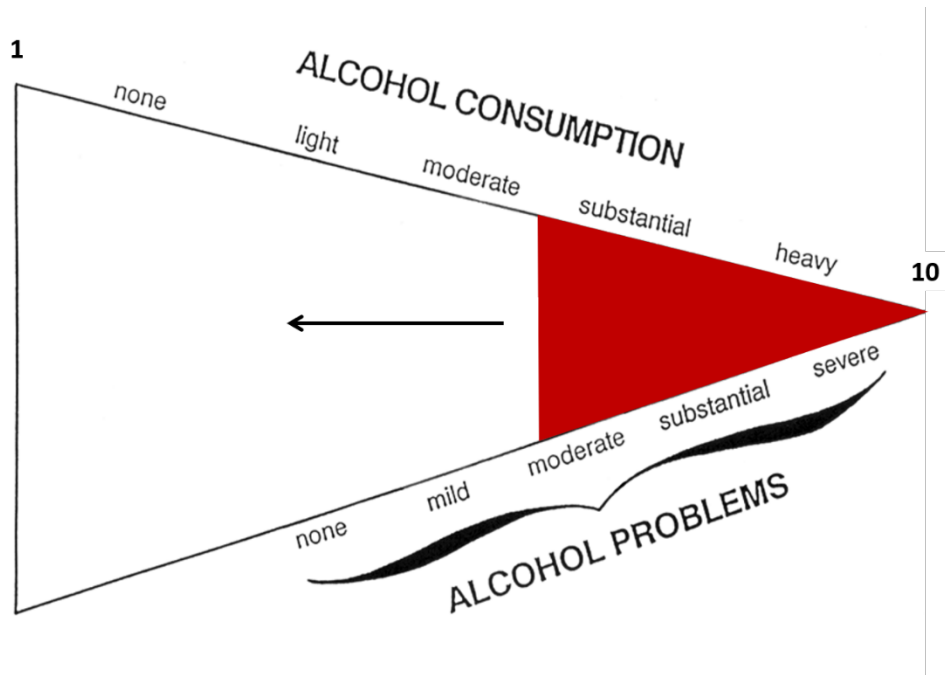


Figure 14. Shows the relationship between increased alcohol consumption and more alcohol problems. The agent highlights that most people are in the larger white area of the triangle and emphasizes that steps can be taken to move in that direction.

Next, the agent asks them, given all they have talked about so far, whether they would be willing to cut back on the number of drinks they normally have. If they are uncertain, she will attempt to persuade them by presenting possible things they might do to decrease their risk of coming to harm and then asks them to select one thing they might consider doing. Then she asks again if they are willing to cut back on their number of drinks. If they say yes, she will ask them how many drinks they plan on having in the next week and does a calculation to make sure they chose a number that's actually lower than what they currently drink. Once satisfied with a lower

number, she asks them to say their goal out loud: I'm only going to drink X drinks for the next week.

Then the agent presents a few tips for safe drinking and asks them to read them over. She reminds them that it's up to them to do what they want with the information that was presented to them and using the neural mode of the hybrid dialog system asks an open question regarding any final thoughts they have on the information that was discussed with the agent during the session. Finally, before ending the session, the agent asks the user if they would consider keeping track of their drinking over the next week, then thanks them for talking with her, wishes them luck, and bids farewell.

Chapter 8. Evaluation of a Hybrid Dialog System for Virtual Alcohol Counseling

Client-centered counseling, such as CBT and MI, are effective methods to change people's attitudes towards their behavior, such as drinking less alcohol [59]. These methods place the client at the center of the discussions, allowing them to express themselves and influence the course of the interaction [80, 108, 114]. Theoretically, virtual counselors that implement client-centered counseling methods as part of their dialog manager should impact clients' attitudes towards their drinking behavior in a more positive direction than an agent using more restrictive technology.

The goal of this study was to measure the effect of a virtual alcohol use counselor on attitudes towards drinking alcohol among college students with mild to moderate AUD, as measured by the Alcohol Use Disorders Identification Test [126]. The negative consequences of alcohol use disproportionately impact college students in the United States. In 2019, an estimated 8.7% of full-time college students ages 18-22 met the criteria for AUD and 1,538 died from alcohol-related injuries. Moreover, college students rarely seek traditional SUD treatment [28]. College students have high levels of adoption and satisfaction with modern technology and devices [26] and being a highly educated young adult is predictive of seeking out non-traditional treatment for health conditions [65]. Therefore, a virtual agent may be an impactful medium for alcohol counseling.

I implemented a virtual agent system that used the hybrid structured-neural approach. The system had speech user input and used natural language understanding to map user utterances to dialog intents. During five selected sections of the counseling session, the system switched from its structured dialog mode to the neural mode and drove the conversation by

predicting facilitation moves using the counselor action prediction model and produced reflections using the reflection generation model. I evaluated the efficacy of this system, in a randomized, between-subjects experiment. Participants were randomized to interact with a virtual agent counselor under one of the following conditions:

1. By speaking with the agent implemented using the speech-based hybrid dialog management system (SPEECH)
2. By using multiple-choice menu options appearing on screen at every dialog turn (MENU)

Using client-centered counseling styles where the counselor demonstrates cooperation in the conversation and recognition of clients' lived experience, for example through reflective listening, are positively correlated with the trust, rapport, and communicative success constructs of therapeutic alliance [107]. Additionally, self-determined or autonomous motivation is associated with greater confidence towards changing one's behavior [108] and counselor responses that are tailored to the individual's circumstances result in greater internalization of the counselor's suggestions [92].

Therefore, the study aims to address the following hypotheses:

H1. Participants in the SPEECH conditions are hypothesized to have higher level of satisfaction, trust, and working alliance with the agent compared to participants in the MENU condition.

H2. Participants who talk to an agent in the SPEECH condition, as opposed to the MENU condition, are hypothesized to have significantly higher readiness, motivation, confidence, and commitment towards drinking less alcohol.

8.1 Measures

A variety of measures were used to evaluate the evaluation study hypotheses. The general study design is depicted in Figure 15, showing the independent and dependent variables.

8.1.1 Agent and Conversation Measures

Perceptions of the virtual counselor were assessed using single 7-point Likert scale items following the session with the agent. The first of these are satisfaction, liking, trusting, and wanting to continue working with the agent, as well as how knowledgeable the agent was. These items measure underlying factors that drive long-term engagement with a relational agents [12, 18]. Participants were also asked how similar they felt they were to the agent and how they would characterize their relationship with her to determine interpersonal closeness.

Conversational naturalness was measured by asking participants to rate how natural they felt the conversation had been, the degree to which they felt like they could express themselves, and if they felt like they had agency during the conversation. These factors are theories to positively influence attitudes towards behavior change [92, 108].

Participants filled out measures of interpersonal trust [144] and the Bond subscale from the Working Alliance Inventory [62]. Establishing trust and therapeutic alliance to build a relationship with patients is important in the treatment of substance use disorders. Early alliance during treatment has been found to influence early improvements and plays an important role in predicting drug treatment outcomes [86]. Virtual agents that show relational behaviors, such as small-talk and displays of empathy as the virtual counselor did, have been shown to have a higher degree of alliance and trust with people than non-relational agents [19]. Higher levels of trust between individuals is also correlated with a higher willingness to self-disclose during conversations [144]. The trust scale is an 8-point 15-item semantic differential scale, where each item has a first anchor that is a word associated with trust and a second anchor that has the opposing meaning. For example, the first three items on the scale are *benevolent - exploitive*, *trustworthy - untrustworthy*, and *confidential - divulging*.

8.1.2 Attitudes Towards Behavior Change (Pre-Post)

Measures of attitudes towards behavior change were assessed before and after interacting with the agent. The Readiness to Change Questionnaire – Treatment Version (RCQ-TV) [35, 121] was designed to identify how individuals currently feel about their drinking and asks participants to think about their current drinking habits on thirty 7-point Likert scale questions. Each question is associated with one of three dimensions that correspond to three of the transtheoretical model's *stages of change* (see section 2.5.2) [109]. The dimension of the RCQ-TV that has the highest score determines the participant's stage of change [57]. The stages can be ranked from low readiness to high, from precontemplation to action, therefore the stage of change measure in this study (SOC) is an ordinal measure from 1 to 3, where 1 = precontemplation, 2 = contemplation, and 3 = action.

Additional measures of change in attitude towards behavior change were three single scale items with anchors from 1 to 10, where 1 is the lowest and 10 is the highest. The items addressed *motivation* (How motivated are you to drinking less alcohol in the future?), *confidence* (How confident are you that you could drink less alcohol in the future?), and *commitment* (How committed are you to drinking less alcohol in the future?). These items have been used in previous studies to measure participants' attitudes towards behavior change before and after interacting with a virtual agent [94, 95].

The final measure of attitudes towards drinking was the change in number of standard alcoholic drinks¹² that participants said they were going to have over the next week compared to the average number of drinks they had per week before starting the study (Drink Change). The participants' average current number of drinks per week was calculated using their answers on

¹² <https://www.niaaa.nih.gov/alcohols-effects-health/overview-alcohol-consumption/what-standard-drink>

the personal feedback form they fill out before starting the interaction with the agent (Table 6), specifically the number of times they drank in the last month and the number of drinks they had per occasion.

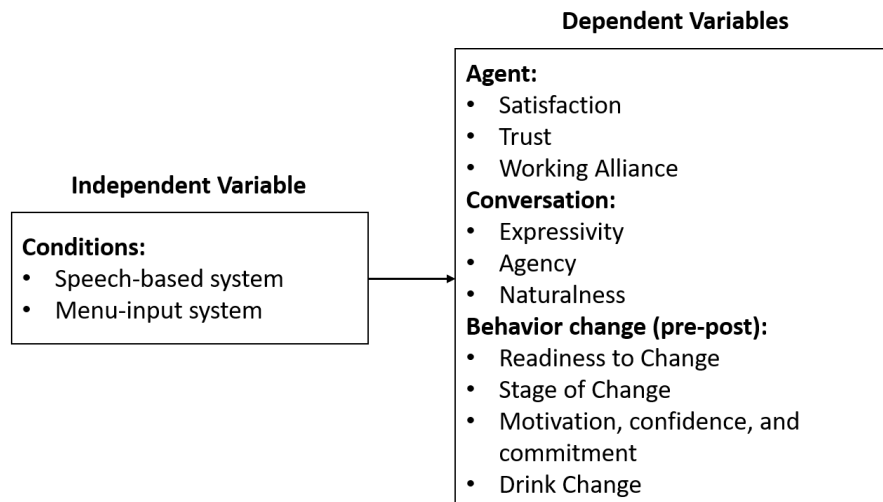


Figure 15. Evaluation study design.

8.2 Participants and Method

College or university students (18 years or older) who are native speakers of English and score a 6-15 on the 10-item Alcohol Use Disorders Identification Test (AUDIT) [126] were recruited. A score in this range on the AUDIT indicates a moderate risk for problematic drinking and the appropriate level of risk to warrant a brief intervention using MI, with a low likelihood of having severe alcohol use disorder.

Participants were randomized to the SPEECH or MENU condition using a blocked randomization method [42]. In every block of four participants, two individuals were randomly assigned to the SPEECH condition and two to the MENU condition. This ensured that there was an equal number of participants in each study arm contains by the end of every block.

The information presented and the overall structure of the session was the same in both conditions, with one exception: the agent in the SPEECH condition prompted the participants to

speak freely using an open elicitation five times during the session. These parts of the session varied in length and content, since they depended on the participants' utterances; however, since these sections of dialog were nested within the larger session structure, all participants experienced every section of the counseling session. Images of the virtual agent and interface in each condition are shown in Figures 16 and 17.

The study was conducted remotely using the Zoom video conferencing software. Participants ran the virtual agent software in the Chrome web browser on their personal computers and answered survey questions on Qualtrics, therefore additional eligibility criteria were having access to the internet, owning a computer with a web-cam and microphone running Windows or MacOS. An online questionnaire was created to screen participants into the study, asking questions to assess each of these eligibility requirements. Eligible participants were then contacted for scheduling a time for the study session to take place.

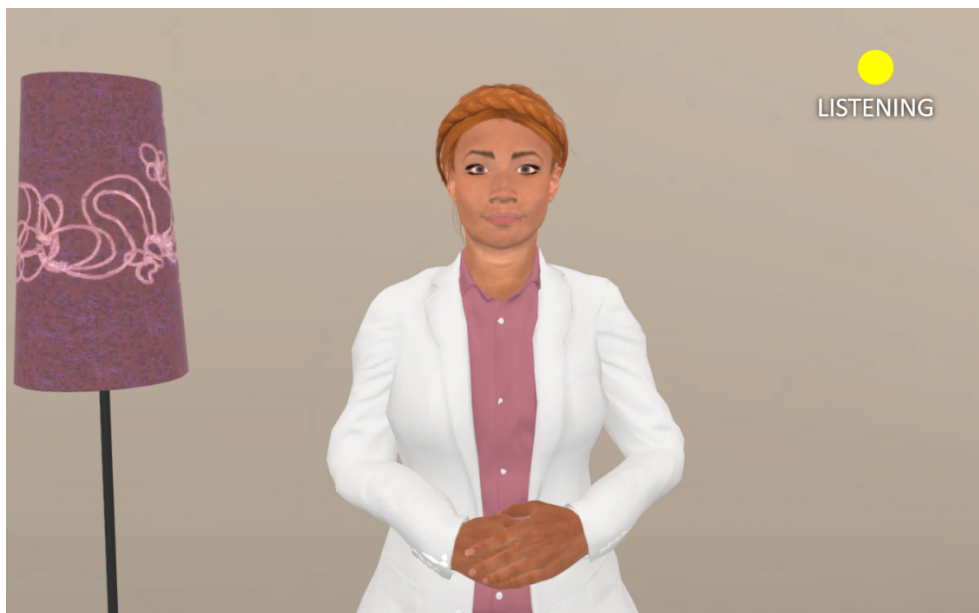


Figure 16. The virtual agent and user interface used in the SPEECH arm of the study.



Figure 17. The virtual agent and user interface used in the MENU arm of the study.

8.3 Procedure

Following informed consent and prior to starting the session, eligible participants followed the flow depicted in the subject flow diagram in Figure 18. First, participants filled out questionnaires on socio-demographics and their readiness, motivation, confidence, and commitment to reducing their alcohol consumption. Participants were also asked to complete a personalized feedback form (Table 6). The form included the definition of a standard drink and a ‘heavy drinking day’ (4 or more drinks for men and 3 or more drinks for women), and the virtual counselor used the information from this questionnaire to tailor the content to the individual during the counseling session. Participants were randomized to the SPEECH or MENU condition and interacted with the virtual counselor for approximately 30 minutes. The research assistant monitored the outputs of the system during the participants’ interactions to intervene in case of system malfunction.

Question	Type
1. Age	Numeric input
2. Gender	Multiple choice
3. How many times in the last month did you drink alcohol?	Numeric input
4. When you drank, approximately how many standard drinks did you have?	Numeric input
5. In the past six months, during your heaviest drinking week, how many times did you drink?	Numeric input
6. During that week, approximately how many standard drinks did you have per occasion?	Numeric input
7. How many heavy drinking days did you have in the past month?	Numeric input
8. In the past year, have you experienced any of the following while drinking or as a result of your drinking? (Please check all that apply)	Checkbox - <i>Getting a headache, Feeling sick, Passing out, Being a passenger of drunk driver, Blacking out, Driving while intoxicated, Getting hurt, Drinking more than intended, Needing to drink larger amounts, No longer getting drunk on the same amount, Acting obnoxiously, Regretting sexual situation, Missing class, Getting behind in schoolwork</i>

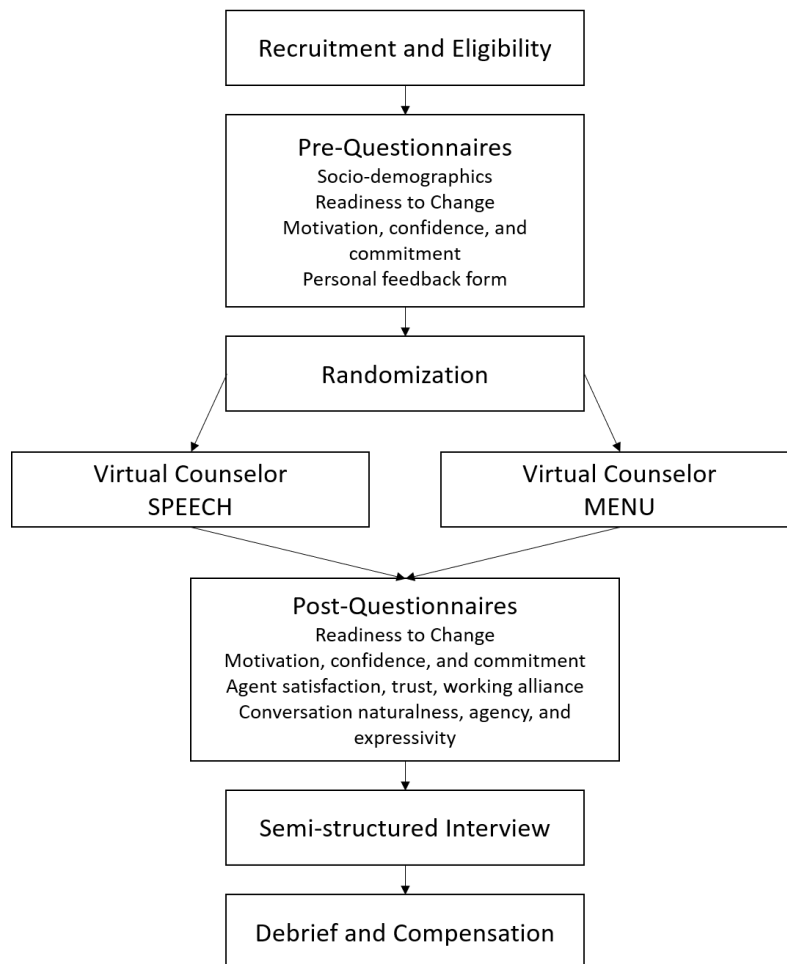
Table 6. The personal feedback form that participants filled out prior to the session. The virtual counselor used this information to tailor the content to the individual.

Following the session with the agent, the participants again filled out the readiness to change questionnaire and motivation, confidence, and commitment to changing their drinking behavior. Next, they rated the agent on several 7-point Likert scale items assessing and then asked if they felt like they could express themselves, had any agency, how natural the conversation felt to them (Table 10). Then, participants answered the questionnaires assessing interpersonal trust and working alliance. Finally, they were asked a series of questions in a semi-structured interview to provide further context to the qualitative data and give participants the opportunity to express their views in their own words:

1. Please tell me about your overall experience using the system.

2. What are your thoughts on [being able to speak | using the menu options] to interact with the agent compared to other input modalities?
 - a. What were the pros and cons of interacting with the agent in that manner?
3. How did the agent make you feel during the session?
 - a. Did the method of interaction impact how you felt?
 - b. Did the topic of discussion impact how you felt?
 - c. Did the agent's behavior and/or utterances impact how you felt?

Before ending the session, participants were informed about the purpose of the study and its conditions, were provided with resources for getting help facing alcohol problems,¹³ and compensated for their time with a \$20 Amazon gift card.



¹³ <https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/treatment-alcohol-problems-finding-and-getting-help>

Figure 18. The flow of participants through the study.

8.4 Results

317 college students responded to the study advertisement, posted on physical and online bulletin boards and sent through email to undergraduate and graduate students at Northeastern University in Boston. The ad included a link to a survey designed to screen for eligible participants and 63 individuals met the study inclusion criteria. 27 responded to an email sent asking their availability for scheduling a session and 26 eligible participants were scheduled. Two participants had to be disqualified due to the virtual agent application malfunctioning during their session, leaving 24 participants who successfully completed the study, with 12 in each condition.

8.4.1 Quantitative Analysis

The average age of participants was 21.88 (SD=3.11) and was similar in each condition (Table 7). Participants' ethnic background was also similar in both conditions, with most identifying as Asian or White. One identified as Middle Eastern or North African and one as being Hispanic, Latinx or Spanish Origin. Some participants selected more than one ethnic background. A majority identified as Female (54%), many as Male (42%), and one as Non-binary. These ratios were similar in each condition, except for Non-binary gender, only present in the SPEECH condition. Most participants were not in a relationship (75%), 16% had an advanced college degree, 42% were graduate students, and the remaining 42% were undergraduate students. 75% were regular or expert users of conversational assistants, such as chat-bots, Siri, and Alexa, and 30% were regular or expert users of virtual environments, such as video games and virtual reality.

All = 24

SPEECH = 12

MENU = 12

Age – Mean (SD)	21.88 (3.11)	21.83 (3.07)	21.92 (3.29)
Background (non-exclusive):			
White	12 (50%)	7 (58%)	5 (42%)
Asian	13 (54%)	5 (42%)	7 (58%)
Middle Eastern or North African	1 (4%)	1 (8%)	0
Hispanic, Latinx or Spanish Origin	1 (4%)	0	1 (8%)
Gender (count):			
Male	10 (42%)	6 (50%)	7 (58%)
Female	13 (54%)	5 (42%)	5 (42%)
Non-binary	1 (4%)	1 (8%)	0
Relationship status:			
Single	18 (75%)	9 (75%)	9 (75%)
In a serious relationship	6 (25%)	3 (25%)	3 (25%)
Education (count):			
Advanced degree	4 (16%)	2 (16%)	2 (16%)
College graduate	10 (42%)	5 (42%)	5 (42%)
Some college	10 (42%)	5 (42%)	5 (42%)
Experience with conversational assistants:			
I've tried one a few times	6 (25%)	2 (17%)	4 (33%)
I use one regularly	17 (71%)	9 (75%)	8 (67%)
I am an expert	1 (4%)	1 (8%)	0
Experience with virtual environments:			
I've never used that	1 (4%)	0	1 (8.3%)
I've tried that a few times	16 (67%)	7 (58%)	9 (75%)
I use that regularly	5 (21%)	4 (33%)	1 (8.3%)
I am an expert	2 (8%)	1 (8%)	1 (8.4%)

Table 7. Descriptive statistics on demographics and background information.

Readiness and Stages of Change

There were significant findings regarding participants' attitudes towards changing their drinking behavior. Readiness to change is measured by the RCQ-TV by allocating participants to one of three stages of change [109, 121]: precontemplation, contemplation, and action. An individual in the precontemplation stage has a low readiness to change while someone in the actions stage has a high readiness. Participants were allocated to stages by summing the scores of the items on the

RCQ that are associated with each stage. The stage that has the highest score is the individual's allocated stage.

There were significantly fewer participants in the precontemplation stage after the interaction with the agent compared to before, 5 (21%) vs. 15 (63%), $X^2(1)=5$, $p<.05$. Additionally, the difference in the number of participants in the contemplation stage was not significantly different before and after interacting with the agent, 6 (24%) vs. 8 (33%), $X^2(1)=0.28$, ns. Finally, the difference between the number of participants in the action stage was significantly higher after interacting with the agent compared to before, 11 (46%) vs. 3 (13%), $X^2(1)=4.57$, $p<.05$.

RCQ-TV Dimension	Before – Count (%)		After – Count (%)	Statistical test
Precontemplation	15 (63)	>	5 (21)	$X^2(1)=5$, $p<.05$
Contemplation	6 (24)	<	8 (33)	$X^2(1)=0.28$, ns.
Action	3 (13)	<	11 (46)	$X^2(1)=4.57$, $p<.05$

Table 8. Counts of participants allocated to a dimension of the RCQ-TV questionnaire before and after the session with the virtual counselor, the direction of change, and results from a test determining if these counts are significantly different.

Motivation, Confidence and Commitment

There were significant pre-post differences in participants' self-reported motivation, confidence, and commitment to behavior change (Table 9). Since these outcomes were measured on single item ordinal scales, a paired-samples Wilcoxon signed-rank test for independent means was used in the following analyses and the median as measure of center with inter-quartile range as a measure of variability.

Motivation to drinking less alcohol in the future was significantly higher after interacting with the agent (Med=8, IQR=3.25), compared to before the session (Med=6.5, IQR=4), $W=15$, $p<.01$. Confidence in being able to drink less in the future before (Med=9, IQR=2.25) and after

the session (Med=9, IQR=1.25) was also statistically significant, $W=16.5$, $p<.05$. Similarly, the difference in commitment to drinking less alcohol in the future was significant comparing scores before (Med=6.5, IQR=4.25) and after (Med=7, IQR=3.25) interacting with the agent, $W=9$, $p<.01$.

Item	Scale	Before – Med (IQR)		After – Med (IQR)	Statistical test
How motivated are you to drinking less alcohol in the future?	1-10	6.5 (4)	<	8 (3.25)	$W=15$, $p<.01$
How confident are you that you could drink less alcohol in the future?	1-10	9 (2.25)	<	9 (1.25)	$W=16.5$, $p<.05$
How committed are you to drinking less alcohol in the future?	1-10	6.5 (4.25)	<	7 (3.25)	$W=9$, $p<.01$

Table 9. Differences in motivation, confidence, and commitment towards changing one’s drinking behavior, measured before and after the session with the agent, and the direction of the change.

Number of Alcoholic Drinks

The number of standard alcoholic drinks was measured before starting the session, in the personalized feedback form, and by the agent towards the end of the session. Neither of these measures were normal (before $W=0.77$, $p<.01$; after $W=0.62$, $p<.01$), therefore, the paired-samples Wilcoxon signed-rank test was used to assess statistical significance. The average number of drinks participants said they were going to have in the next week ($M=3.62$, $SD=5.49$) was significantly lower than the average number of drinks that participants said they were currently drinking ($M=8.42$, $SD=9.33$), $W=231$, $p<.01$.

Agent Ratings

Following the counseling session, participants rated the virtual counselor on several 7-point single scale items (Table 10). The median and inter-quartile range for each item per condition is reported and a comparison of those ratings with a neutral score of 4.

The median *satisfaction* with the agent in the SPEECH condition was 5 (2) and 5 (1) in the MENU condition. Both were found to be significantly higher than neutral: SPEECH $W=102$, $p<.05$; MENU $W=144$ $p<.05$. Wanting to *continue working* with the agent had a median score of 4.5 (3.25) in the SPEECH condition and 4.5 (2) among participants in the MENU condition. Neither were significantly different from neutral: SPEECH $W=78$, ns.; MENU $W=96$, ns. A single scale item of *trust* had a median of 5 (1.25) in the SPEECH condition and 6 (1) in the MENU condition. These ratings were both significantly higher than neutral: SPEECH $W=102$, $p<.05$; MENU $W=126$, $p<.05$. *Liking* the agent had a median rating of 5(2) following the SPEECH condition and 6 (1.25) among participants in the MENU condition. These ratings were significantly higher than neutral in both conditions: SPEECH $W=108$, $p<.05$; MENU $W=126$, $p<.05$. Participants in the SPEECH condition gave the agent a median rating of 6 (1.25) on how *knowledgeable* they found her and a 7 (1) in the MENU condition. These ratings are significantly higher than neutral: SPEECH $W=144$, $p<.05$; MENU $W=177$, $p<.05$.

Relational closeness in the SPEECH condition had a median rating of 4 (3.25) and a 3 (2.25) in the MENU condition. These ratings were not significantly different neutral: SPEECH $W=66$, ns.; MENU $W=54$, ns. Participants rated how similar they felt they are to the agent and those in the SPEECH condition gave a median rating of 5.5 (3) and those in the MENU condition gave a 5.5 (2.25). These ratings were significantly higher than neutral: SPEECH $W=108$, $p<.05$; MENU $W=108$, $p<.05$.

The naturalness of the conversation was measured using three single scale items. The first item asked participants how natural they thought the conversation with the agent had been. Among participants in the SPEECH condition, this item had a median rating of 3 (3.25) and a 4 (2) among those in the MENU condition. These ratings were not significantly different from

neutral: SPEECH W=60, ns.; MENU W=66, ns. The second item asks participants to rate the degree to which they felt like they could express themselves during the conversation. This item had a median rating of 4 (3) among participants in the SPEECH condition and a 4 (2.25) among those in the MENU condition. These ratings are not significantly different from neutral: SPEECH W=78, ns.; MENU W=72, ns. The third and final measure of conversational naturalness asked participants how much agency they had felt having during the conversation. This measure had a median rating of 4 (2) among participants in the SPEECH condition and 4 (1) among those in the MENU condition. These ratings were not significantly different than a neutral rating: SPEECH W=78, ns.; MENU W= 90, ns.

Item	Anchor 1	Anchor 2	Condition	Med (IQR)	Comparison to neutral
How satisfied are you with the agent?	Not at all satisfied	Very satisfied	SPEECH	5 (2)	W=102
			MENU	5 (1)	W=144
How much would you like to continue working with the agent?	Not at all	Very much	SPEECH	4.5 (3.25)	W=78, ns.
			MENU	4.5 (2)	W=96, ns.
How much did you trust the agent?	Not at all	Very much	SPEECH	5 (1.25)	W=102
			MENU	6 (1)	W=126
How much do you like the agent?	Not at all	Very much	SPEECH	5 (2)	W=108
			MENU	6 (1.25)	W=126
How knowledgeable was the agent?	Not at all	Very knowledgeable	SPEECH	6 (1.25)	W=144
			MENU	7 (1)	W=144
How would you characterize your relationship with the agent?	Complete stranger	Close friend	SPEECH	4 (3.25)	W=66, ns.
			MENU	3 (2.25)	W=54, ns.
How similar do you feel that you are to the agent?	Very different	Very similar	SPEECH	5.5 (3)	W=108
			MENU	5.5 (2.25)	W=108
How natural was your conversation with the agent?	Not at all	Very natural	SPEECH	3 (3.25)	W=60, ns.
			MENU	4 (2)	W=66, ns.
I felt like I could express myself during the conversation	Strongly disagree	Strongly agree	SPEECH	4 (3)	W=78, ns.
			MENU	4 (2.25)	W=72, ns.
I felt like I had agency during the conversation	Strongly disagree	Strongly agree	SPEECH	4 (2)	W=78, ns.
			MENU	4 (1)	W=90, ns.

Table 10. Averages, per condition, of several single scale items for rating the virtual counselor. The last column shows the statistic comparing the rating to a neutral score of 4, at $p < .05$.

Trust and Working Alliance

The degree to which participants trusted the agent was measured on a composite scale developed to measure trust between individuals as it relates to self-disclosure [144] (results in Table 11). Each item has 8-points, and thus the composite score of the scale ranges from 1-8, where 1 is the lowest level of trust and 8 is the highest. This measure was normally distributed (Shapiro-Wilks = 0.94, ns.) and therefore means and standard deviations are reported. The mean rating of trust in the agent was 6.67 (0.94). The mean rating among participants in the SPEECH condition was 6.46 (1.05) and 6.88 (0.81) among those in the MENU condition.

Working Alliance was assessed using the Bond subscale of the Working Alliance Inventory (WAI-Bond) [62]. This measure consists of 12 7-point Likert scale items, with anchors ranging from ‘Strongly Disagree’ to ‘Strongly Agree’. The measure was normally distributed (Shapiro-Wilks=0.94, ns.) and therefore means and standard deviation are reported. The WAI-Bond had a mean of 4.6 (1.07) and the mean rating among participants in the SPEECH condition was 4.23 (1.21) and 4.98 (0.79) among those in the MENU condition.

Measure	Scale	Overall Mean (SD)	Condition	Mean (SD) per condition
Interpersonal trust	1-8	6.67 (0.94)	SPEECH	6.46 (1.05)
			MENU	6.88 (0.81)
Working Alliance (Bond)	1-7	4.6 (1.07)	SPEECH	4.23 (1.21)
			MENU	4.98 (0.79)

Table 11. The mean composite trust and working alliance with the virtual counselor per condition.

8.4.1.1 First Hypothesis

My first hypothesis is based on the relationship between being able to express oneself freely in client-centered therapy sessions and constructs of trust, rapport, therapeutic alliance [107].

Participants in the SPEECH condition are expected to have experienced a higher sense of agency, expressiveness, and naturalness compared to those in the MENU condition.

Therefore, participants in the SPEECH condition are hypothesized to have higher level of satisfaction, trust, and working alliance with the agent compared to participants in the MENU condition. To evaluate this hypothesis, I conducted tests using inferential statistics to determine if there is a significant difference between the conditions on these variables.

Conversation Naturalness

The items assessing participants' degree of expressivity, agency, and naturalness of the conversation are single scale ordinal items and the between subject differences will thus be tested using the non-parametric Wilcoxon signed-rank test for independent means. Participants' sense of *expressivity* was not significant between the two conditions, $W=67$, ns. Additionally, the *agency* participants felt after interacting with the agent was not significantly different between the conditions, $W=75.5$, ns. Lastly, the naturalness of the conversation was not found to be significantly different between the two conditions, $W=86.5$, ns.

Agent Ratings

The measures of satisfaction, trust (single item), liking, wanting to continue working with the agent, and how knowledgeable the agent is are a single scale ordinal items and will therefore be assessed using the Wilcoxon signed-rank test for independent means. The measure of satisfaction was not significantly different between the conditions, $W=94.5$, ns. Trust on the single scale item was not significantly different between conditions, $W=95.5$, ns. The item measuring wanting to continue working with the agent was not significantly different between conditions, $W=73.5$, ns.

The measure of liking the agent was not significantly different between conditions, $W=86$, ns. There was not a significant difference between the conditions on how knowledgeable participants found the agent to be, $W=101.5$, ns.

Trust and Working Alliance

The composite measures of trust and WAI-Bond were found to have a normal distribution and differences between conditions are therefore assessed using the t-test for independent means. Composite trust in the agent was not significantly different between the two conditions, $t(20.72)=1.12$, ns. Similarly, differences between conditions on the WAI-Bond scale were not significant, $t(18.97)=1.81$, ns.

8.4.1.2 Second Hypothesis

My second hypothesis is based on the idea that motivation gained by self-expression is related to positive changes in attitudes towards behavior change [108] and that counselor remarks that take the individual's circumstances into account positively impact how clients incorporate the counselor's suggestions [92]. The hybrid dialog management approach allows individuals to talk to the virtual counselor using their own words and the counselor may reflect their sentiments back to them. Therefore, participants who talk to an agent in the SPEECH condition, as opposed to the MENU condition, are hypothesized to have significantly higher readiness, motivation, confidence, and commitment towards drinking less alcohol.

Readiness and Stage of Change

The RCQ-TV measures individuals' readiness to change their drinking behavior [121]. The questionnaire has three constructs used to allocate individuals to one of three stages of change: precontemplation, contemplation, and action [109]. A summary of results on pre-post changes between conditions are reported in Table 12.

Stage of Change	Condition	Count Before	Count After
Precontemplation	SPEECH	11	3
	MENU	4	2
Contemplation	SPEECH	1	3
	MENU	5	5
Action	SPEECH	0	6
	MENU	3	5

Table 12. Counts of participants per stage of change between conditions before and after interacting with the virtual counselor.

The stage of change (SOC) measure in this study is derived from the RCQ-TV by assigning a number from 1-3 to participants, given their stage. Those in precontemplation got an SOC of 1, contemplation an SOC of 2, and action an SOC of 3. The change in SOC was calculated by subtracting the pre SOC from the post SOC. This pre-post change in SOC was not normally distributed (Shapiro-Wilk=0.84, $p < .05$), thus a non-parametric test was used to test for a significant difference between the conditions. The pre-post difference in the SOC measure was significantly higher for participants in the SPEECH condition ($M=1.33$, $SD=1.56$) compared to those in the MENU condition ($M=-0.17$, $SD=1.19$), $W=195$, $p < .05$ (Figure 19).

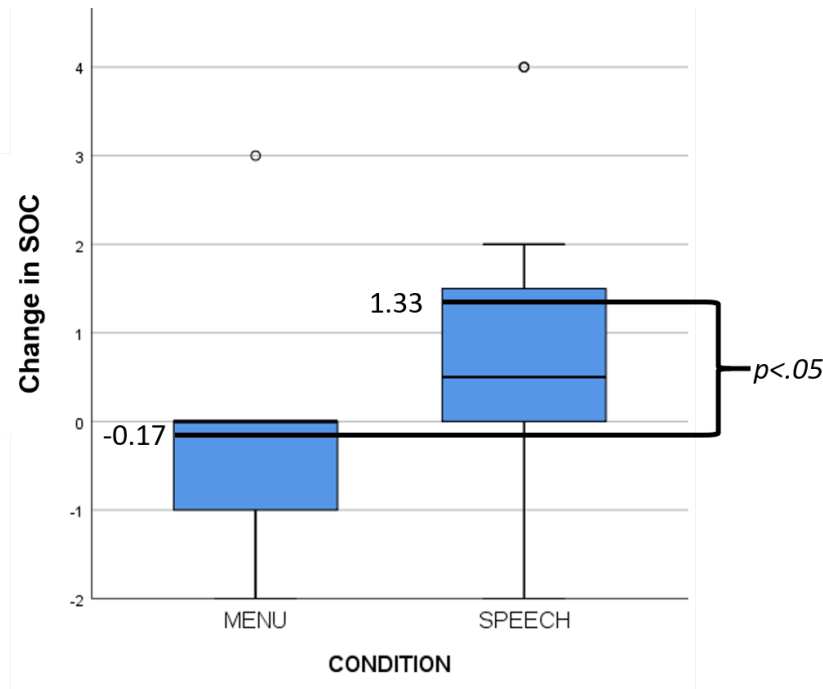


Figure 19. The change in the stage of change (SOC) measure was significantly higher among participants in the SPEECH condition compared to those in the MENU condition.

Change in Drink

The change in drink measure is the number of standard alcoholic drinks participants say they are going to drink in the next week subtracted from the number of standard alcoholic drinks that participants are currently drinking per week. This measure was not normally distributed (Shapiro-Wilks=0.61, $p < .01$), therefore a non-parametric test was used to assess differences between the two conditions. The mean change in drink for participants in the SPEECH condition was 4.08 (9.15) drinks and 5.5 (6.35) drinks for those in the MENU condition. These differences were not significantly different between conditions, $W=86$, ns.

Correlations

There were several significant correlations between outcome variables of interest (Table 13). Expressivity and wanting to continue working with the agent are positively correlated with the pre-post change in the number of drinks participants said they would have ($\sigma=.5$, $p < .01$ and $\sigma=.6$,

$p < .01$) and positively correlated with having a drinking plan for the next week ($\sigma = .55$, $p < .01$ and $\sigma = .45$, $p < .05$). Expressivity and agency are strongly positively correlated with overall satisfaction with the agent ($.8$, $p < .01$ and $.6$, $p < .01$).

	Continue work with agent	Drink change	Having drink plan	Agent satisfaction	Agency
Expressivity	$\sigma = .76^{**}$	$\sigma = .5^{**}$	$\sigma = .55^{**}$	$\sigma = .8^{**}$	$\sigma = .7^{**}$
Continue work with agent		$\sigma = .6^{**}$	$\sigma = .45^*$	$\sigma = .94^{**}$	$\sigma = .53^{**}$
Drink change			$\sigma = .6^{**}$	$\sigma = .71^{**}$	$\sigma = .47^*$
Having drink plan				$\sigma = .46^*$	ns.
Agent satisfaction					$\sigma = .6^{**}$

Table 13. Significant Spearman’s Rank Order correlations between outcome variables of interest. * $p < .05$ and ** $p < .01$.

8.4.2 Qualitative Analysis

Semi-structured interviews were conducted with participants following the session with the virtual counselor to qualify the quantitative measures. The interviews were audio recorded, transcribed, and thematically analyzed. The questions asked participants for their overall impressions of the system, their thoughts on being able to speak or use menu options to talk to the agent (depending on condition) compared to other modes of interacting with computers, and how the agent had made them feel during the session. Chi-squared tests were conducted where possible.

General Impressions

When asked about their overall experience using the system, participants had a generally positive impression across conditions, $X^2(1) = 6$, $p < .05$ (Table 14). The difference between conditions was not significant, $X^2(1) = 0.22$, ns. Participants who indicated having a positive impression felt the system was “a very good kind of system, very informative, the agent” (P8-M) and some found it

“really interesting, I've never experienced anything like that before” (P15-M). The system was also seen as “pretty interactive” (P5-S) and “nice, because people need someone to just hear and talk” (P17-S). Other positive impressions that came through were finding it “positive because it was informative” (P13-M) and liking it “because it was kind of personalized to my answers” (P14-S). The negative impressions seemed related to how the agent presented herself and not feeling “any relationship with the agent, or that they would be able to help me if I had a problem” (P7-M) and that she was “uncanny ... the way she talks and responds is artificial and weird” (P24-S). There were issues of feeling like the agent did not have human qualities “it was a computer program and not a real person” (P22-M) and negative experiences with respect to the flow of conversation, stating that “it was stagnant” (P16-S).

General Impressions	Positive	Negative	Statistical test
All	18	6	$X^2(1)=6, p<.05$
SPEECH	8	4	$X^2(1)=0.22, ns.$
MENU	10	2	

Table 14. Count of participants who mentioned having generally positive and negative impressions of the virtual counselor system, per condition, and tests for differences between groups.

Impressions on Input Modalities

Participants expressed a variety of views on how they interacted with the virtual counselor before they were asked to reflect on the particular input modality they used (Table 15). Those in the SPEECH condition communicated with the agent verbally and those in the MENU condition pressed a button from a list of menu options at every dialog turn. Participants expressed grievances concerning their mode of interaction with the agent, feeling like one’s “responses are limited to whatever the options are” (P2-M) and thinking that “she expects a certain answer to continue in the open ended questions, and I don't know how to get to those answers” (P4-S). For

some participants, the flow of conversation was interrupted by the input modality because “it got hung up here and there on a couple of occasions” (P5-S) and feeling “weird because I feel like in a counseling session you have the ability say what you want and there were only like dropdown buttons” (P10-M).

	Had negative view on input
All	16
SPEECH	7
MENU	9

Table 15. Some participants expressed negative views about the input modality they used prior to being explicitly asked to talk about that topic.

Participants were explicitly asked to comment on the input modality they used to communicate with the virtual counselor and compare that to other modalities they are familiar using to interact with computers (Table 16). Participants expressed being uncertain about the capabilities of the system and what it expected from them. Some had a feeling that it was deterministic and “had a feeling that no matter what I say she would reach the same end goal” (22-M), while others weren’t “sure what was going to happen when I said something” (P16-S) and were not “sure about her full capabilities with my speaking” (P14-S).

Participants also expressed feeling limited by the input modality stating they might “have a different opinion than what is available as an option” (P6-M) and that they “could only pick the options that were there, it was not flexible enough” (P7-M). One participant felt like they could not divulge what they wanted to say and “didn't really feel like opening up to the maximal potential I could ... just thinking about saying what the computer wanted me to say” (23-S).

Some participants had issues with how the input modality affected the conversation flow and feeling like “at some point you want a two-way conversation ... you just wanted to have a

conversation basically” (P12-M). Others expressed that the agent disrupted the flow of the conversation when she “cut me off before I finished or wait too long after I finished” (P24-S).

When considering other input modalities that they have used to interact with computers, some participants expressed preferences. There were participant that liked speaking to the agent over other modalities, feeling “sometimes when you select the prewritten response or type in, you have less room for honesty or giving the full amount” (P14-S) and a preference for speaking “over writing answers ... it gives you the feeling of talking to some person” (P10-S). Some participants preferred using the menu options, feeling that the clicking “experience is better and engaging, I don't have to wait” (P8-M).

Some had preferences with contingencies, stating that “if given the option to talk, I think that would work ... as long as it understands me” (P13-M) and that “talking if it's done right is always better” (P20-M). Others wanted to express their thoughts their own words and “would rather speak or type something out, so it could be my own words instead of clicking on something else” (P18-M).

	Uncertainty about expectations	Felt limited by the input modality	Issue with conversation flow	Preferred their given modality
All	8	10	4	11
SPEECH	5	2	3	8
MENU	3	8	1	3

Table 16. Counts of participants that were uncertain about the agent’s expectations, felt limited by their respective input modality, had issues with the flow of conversation, and talked about their modality preferences.

How the virtual counselor made participants feel

Participants were asked how the agent had made them feel during the session and follow up questions were asked to discern the possible cause of feelings they might have experienced (Table 17).

Several participants felt positive about the agent being informative and knowledgeable about the topic. They felt she was “very informative, just giving out the facts honestly” (P2-M) and that “it didn't feel like it's just educating me or sit down and listen to this” (P17-S). Some were surprised or shocked by the information the virtual counselor presented about their drinking in a way that made them appreciate the agent: “It was kind of shocking to see the percentile I was in compared to other students” (P6-M). For some it was “a few facts I was surprised by” (P11-M) and one participant felt that “she's scaring me a lot like my mother” (P12-S) in a way that signaled that the agent cared and felt worried. Several participants felt comfortable and at ease, stating that “she made me feel pretty comfortable” (P14-S) and that “she made the topic feel comfortable because it was more like advice” (P3-M). Many participants felt positive that the agent was non-judgmental and took a neutral stance regarding their alcohol use. They felt that she was “just educat[ing] without being ‘you should know’ ... it was a very natural way of conversing” (P17-S) and that the agent’s non-judgmental stance “made me feel comfortable ... she's not trying to judge anyone, just giving information” (P15-M).

Some participants felt like the agent showed human qualities or behavior that impacted them positively and helped the conversation feel natural and non-robotic. Participants stated that “her behavior was great, it made a difference ... looked like it was having an actual conversation” (P10-S), that that virtual counselor’s behavior “just felt natural, wasn't just a robot thing” (P20-S), and that the interaction had been “interesting because it didn't seem robotic” (P1-M).

Others felt the agent had made them feel weird and awkward, and that this experience impacted their perception of the conversation: “as far as a real conversation, it was very

awkward” (P16-S) and “the agent made me feel a little awkward ... it's a weird experience talking with a virtual human” (P7-M).

	Positive – agent informative	Surprise or shock	Comfort or at ease	Not judged	Positive – agent felt human	Weird or awkward	Neutral
All	14 (58%)	6 (25%)	10 (42%)	13 (54%)	6 (25%)	6 (25%)	9 (38%)
SPEECH	6	1	7	8	5	3	2
MENU	8	5	3	5	1	3	7

Table 17. Counts of participants, per conditions, who experienced these feelings during the session with the agent.

8.5 Discussion

This study evaluated a hybrid neural-structured dialog management approach for a virtual counselor to conduct alcohol counseling sessions with college students. Participants were recruited to participate in a single session with the virtual counselor and were assessed on their readiness to change their drinking habits, their satisfaction, trust, and working alliance with the virtual counselor, and conversational naturalness.

This study found significant differences on several of the dimensions of readiness to change drinking behavior before and after interacting with the agent across and within conditions. Similarly, measures of motivation, confidence, and commitment to drinking less were significantly higher after interacting with the agent in either condition. Additionally, at the end of the session with the agent, the number of alcoholic drinks participants said they would have in the next week was significantly lower than the number of drinks they said they had on an average week. This indicates that college students with mild to moderate alcohol use disorder can see an increase in their readiness, confidence, motivation, and commitment to changing their

drinking habits after one session with a virtual alcohol counselor that implements an existing BMI treatment intervention.

On several measures related to satisfaction, the virtual counselor was rated higher than neutral in both conditions. The qualitative analysis revealed that a participants expressed positive feelings towards the agent, for example, finding the agent informative and knowledgeable, comforting, and nonjudgmental. Measures of relatedness and conversational naturalness were not significantly different than neutral, suggesting that one session might not be enough for the participants to form a friendship with the agent after and they were ambivalent as to whether the interaction in either condition was natural. The qualitative analysis indicates that participants in both conditions had experiences that may have affected the naturalness of the interaction and flow of conversation. For example, some participants in the MENU condition felt limited by their input modality while some participants in the SPEECH condition were uncertain about the expectations of the system and had issues with conversation flow.

The first hypothesis stated that participants in the SPEECH condition would have a higher level of satisfaction, trust, and working alliance with the agent compared to participants in the MENU condition. The differences between conditions on these measures were not significant and this hypothesis is therefore not supported.

The second hypothesis stated that participants in the SPEECH condition will have significantly higher readiness, motivation, confidence, and commitment towards drinking less alcohol compared to those in the MENU condition. The changes in the number of alcoholic drinks per week, motivation, confidence, and commitment to drink less measured before and after the conversation with the agent were not significantly different between the conditions. The pre-post difference in stage of change was significantly higher among those in the SPEECH

condition compared to those in the MENU condition. This indicates that participants interacting with a virtual counselor using the hybrid structured-neural dialog management approach move further in a positive direction along the trajectory of changing one's drinking behavior than those using the menu-based system. However, when p-values are adjusted for multiple comparisons using the Bonferroni method, the difference is not significant, and the second hypothesis is therefore not supported.

8.5.1 Limitations

This study has several limitations. The sample used in this study may not be representative of all college students in the United States, and therefore these findings may not generalize to the population as a whole. The study was conducted remotely through videoconferencing software and therefore could not adequately control for environmental factors in the same manner that a controlled lab study might. This may decrease the power of the study but potentially make significant findings more generalizable.

The number of non-significant findings could have several explanations. Firstly, the study may have been underpowered and the effect between conditions could be small. A greater number of participants and a controlled laboratory setting could bring out the differences between the conditions on some of these measures. Secondly, the qualitative analysis revealed that some participants that spoke with the agent had at times experienced a disruption in the flow of the conversation and been uncertain about what the system expected them to say. This experience may have reduced participants' sense of expressiveness and agency, which were found to be significantly positively correlated with satisfaction and wanting to continue working with the agent, as well as the change in number of drinks participants said they would have in the future (Table 13).

The counseling session may also not have led to actual behavior change. There was no longitudinal follow-up conducted with the participants to find out if participants reached their goals of drinking less during the week following the session.

8.5.2 Future work

A first step is to improve the speech-based system to alleviate the issues participants raised around disrupted flow of conversation and uncertainty about the agent's expectations. A system that has the capability to handle user interruptions, display appropriate verbal and nonverbal backchannels [93], and implements dialog repair strategies [148] should be integrated into the current virtual counseling architecture. Adding support for a greater number of user input modalities would therefore be an interesting venture, such as gaze and head tracking through video and prosody and intonation from speech, to allow users to communicate conversational functions through nonverbal behaviors.

The efficacy of an improved system should be investigated at multiple sites with larger numbers of participants. It would also be insightful to implement the full 8-week manualized intervention [101] and conduct a randomized controlled trial with follow-ups at 6 and 12 months to get a better understanding the impact that a relational virtual counseling agent can have on longitudinal behavior change.

Chapter 9. Conclusions

In this dissertation, I described the development of a hybrid structured-neural spoken dialog system for conducting counseling in MI-based sessions with college students with mild to moderate Alcohol Use Disorders (AUD). The approach merges a rule-based dialog management approach that maintains control across multiple turns of dialog, with a neural network-based

natural language processing that allows for automated responses to unconstrained client speech in defined discourse contexts.

Effective treatment therapies for substance use disorders are client-centered, in that they allow clients to express themselves in conversation with counselors, which has been found to lead to a more stable form of behavior change [104]. In practice, counseling sessions follow an underlying structure that places boundaries on the kinds of conversations that can be had. At the same time, client-centered counseling requires allowing clients to express themselves freely while coherently carrying on with the conversation. For example, in motivational interviewing (MI) and other client-centered counseling methods, it is important for the counselor to listen to the client's utterances and reflect back to them their intention in a neutral and facilitating manner [89].

Unfortunately, some of the known barriers to engaging in treatment ultimately impede recovery for many patients, for example, poor social support, privacy concerns, time conflicts, lack of treatment availability, and difficulties around admission [102]. Automated counseling may be a scalable solution to address these issues; however, understanding client utterances, responding appropriately, and managing the therapeutic agenda poses significant technical challenges. Modern natural language processing methods, such as deep neural networks, have the potential to meet these challenges and have been used for modeling patient-provider dialog [48, 103] and dialog generation [111]

Models that predict counseling moves were learned automatically from annotated patient-provider counseling session transcripts. The best performing model had two neural networks (LSTMs) and used a sequence modeling approach (CRF) to predict five high-level counseling moves at every dialog turn. The model uses counselor and client utterances at any given dialog

turn to predict a next counseling move for a virtual counseling agent to make. Experts in MI counseling rated two of the model's action types (reflecting and grounding) as being appropriate for MI, to make sense in the context they appeared, and their placement in the context of the dialog to be unlikely to cause harm.

Utterances for a virtual counselor can be automatically generated using a neural network-based language modeling approach by training a transformer [136] on transcripts from patient-provider sessions. The model generates utterances for the virtual counselor, such as the reflections, conditioned on the dialog context. The reflections generated from the language model that was ultimately used in the virtual alcohol counseling system were found to meet the minimum requirements to be considered reflections, to be coherent sentences of English, and were not considered to be harmful. However, their appropriateness and coherence in terms of the dialog context was less clear and further investigations into how to improve discourse coherence should be conducted.

A virtual alcohol use counselor that uses the hybrid dialog system and follows the structure of manualized alcohol use treatment interventions was designed and implemented. An evaluation of the system showed that participants had greater readiness to change their drinking habits and attitudes after one session of BMI with the virtual counselor than they had before. The system was designed to retain control over the flow of the counseling session as a whole and handle open-ended user utterances at particular moments in the session. This design was chosen for safety reasons to ensure that the counselor can move on to a next section of dialog in case the conversation gets derailed during an open-ended section. Samples of the kinds of actions and language produced by the models that control the open-ended parts of the dialog were evaluated by individuals with counseling experience who found them unlikely to cause harm, thus a bigger

risk to the integrity of the counseling session in this case might be misunderstandings and other disruptions to the conversation flow. None of the negative experiences mentioned by the participants in the evaluation study had to do with offensive or inappropriate language from the counselor. However, the current system does not guarantee that the language generation model will never produce an offensive or inappropriate remark. Further studies with an improved system are needed with larger numbers of participants and several health behavior domains to evaluate the capabilities and generalizability of this dialog management approach.

9.1 Future Work

There are many exciting directions one could take to build on and improve this work. There are a multitude of dialog management approaches, as discussed in section 2.3.1, that would be possible to combine and explore for this specific task of alcohol counseling. For example, integrating dialog management systems that have sophisticated methods for handling interruptions and improve feedback during conversations [93], as well as incrementally process user input under uncertainty and initiate dialog repair strategies [148] could improve user interaction capabilities of the current system.

The field of natural language processing moves at a fast pace, with new state-of-the-art methods for tasks, such as language generation and text classification, emerging every year. Future work should continue exploring the use of the latest methods to retrain the models used for the counselor dialog act prediction and language generation and compare with previous approaches. These methods tend to require a large quantity manually annotated data to model these conversations, therefore work that continues to explore the fusion of neural network-based approaches with algorithms that can perform well under conditions of limited data, such as the combination of LSTMs and CRF explored in this dissertation.

Additionally, given fast pace of the field, it is important to explore the safety concerns around having conversational agents advise and counsel patients or users on medical issues [17]. For example, the type of language model used in this study has known flaws [142] that are understudied in the domain of healthcare. The deployment of understudied technologies in real-world healthcare settings raises significant ethical concerns that can only be addressed by further research and development.

Finally, this work could be combined with other digital approaches to health behavior change, such as digital therapeutics [71], and build on previous work using virtual agent counselors in this domain [129] to create a more general domain-agnostic system that interleaves health behavior change intervention materials and features with virtual counseling sessions.

References

- [1] Adiwardana, D. et al. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*. (2020).
- [2] Allen, J. 1995. *Natural language understanding*. Pearson.
- [3] Allen, J.F. et al. 1996. A robust system for natural spoken dialogue. *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (1996), 62–70.
- [4] Allwood, J. et al. 2000. Cooperation, dialogue and ethics. *International Journal of Human-Computer Studies*. 53, 6 (2000), 871–914.
- [5] Ardito, R.B. and Rabellino, D. 2011. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. *Frontiers in psychology*. 2, (Oct. 2011), 270. DOI:<https://doi.org/10.3389/fpsyg.2011.00270>.

- [6] Association, A.P. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
- [7] Bae, S. et al. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 1, 2 (2017), 5.
- [8] Baevski, A. et al. 2019. Cloze-driven Pretraining of Self-attention Networks. (Mar. 2019).
- [9] Balsa, A.I. et al. 2010. The impact of ICT on adolescents' perceptions and consumption of substances. (2010).
- [10] Beck, A.T. 1970. Cognitive therapy: Nature and relation to behavior therapy. *Behavior therapy*. 1, 2 (1970), 184–200.
- [11] Bengio, Y. et al. 2003. A neural probabilistic language model. *Journal of machine learning research*. 3, Feb (2003), 1137–1155.
- [12] Bickmore, T. et al. 2010. Maintaining Engagement in Long-Term Interventions With Relational Agents. *Applied Artificial Intelligence*. 24, 6 (Jul. 2010), 648–666.
DOI:<https://doi.org/10.1080/08839514.2010.492259>.
- [13] Bickmore, T. and Cassell, J. 2001. Relational agents: a model and implementation of building user trust. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2001), 396–403.
- [14] Bickmore, T. and Gruber, A. 2010. Relational agents in clinical psychiatry. *Harvard review of psychiatry*. 18, 2 (2010), 119–130.
- [15] Bickmore, T.W. et al. 2011. A reusable framework for health counseling dialogue systems

- based on a behavioral medicine ontology. *Journal of biomedical informatics*. 44, 2 (2011), 183–197.
- [16] Bickmore, T.W. et al. 2018. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of siri, alexa, and google assistant. *Journal of Medical Internet Research*. 20, 9 (2018).
DOI:<https://doi.org/10.2196/11510>.
- [17] Bickmore, T.W. et al. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *J Med Internet Res*. 20, 9 (2018), e11510.
DOI:<https://doi.org/10.2196/11510>.
- [18] Bickmore, T.W. and Picard, R.W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*. 12, 2 (2005), 293–327. DOI:<https://doi.org/10.1145/1067860.1067867>.
- [19] Bickmore, T.W. and Picard, R.W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*. 12, 2 (2005), 293–327.
- [20] Bien, T.H. et al. 1993. Brief interventions for alcohol problems: a review. *Addiction*. 88, 3 (1993), 315–336.
- [21] Bobrow, D.G. et al. 1977. GUS, a frame-driven dialog system. *Artificial intelligence*. 8, 2 (1977), 155–173.
- [22] Bohnet, B. et al. 2018. Morphosyntactic Tagging with a Meta-BiLSTM Model over

Context Sensitive Token Encodings. (May 2018).

- [23] Bohus, D. and Rudnicky, A.I. 2003. RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda. (2003).
- [24] Bohus, D. and Rudnicky, A.I. 2008. Sorry, I Didn't Catch That! *Recent trends in discourse and dialogue*. Springer. 123–154.
- [25] Bowen, S. et al. 2014. Relative Efficacy of Mindfulness-Based Relapse Prevention, Standard Relapse Prevention, and Treatment as Usual for Substance Use Disorders: A Randomized Clinical Trial. *JAMA Psychiatry*. 71, 5 (May 2014), 547–556.
DOI:<https://doi.org/10.1001/jamapsychiatry.2013.4546>.
- [26] Brooks, D.C. and Pomerantz, J. 2017. ECAR Study of Undergraduate Students and Information Technology, 2017. *EDUCAUSE*. (2017).
- [27] Bunt, H. et al. 2007. An empirically based computational model of grounding in dialogue. *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (2007), 283–290.
- [28] Caldeira, K.M. et al. 2009. College students rarely seek help despite serious substance use problems. *Journal of substance abuse treatment*. 37, 4 (2009), 368–378.
- [29] Campbell, A.N.C. et al. 2014. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *The American journal of psychiatry*. 171, 6 (Jun. 2014), 683–90. DOI:<https://doi.org/10.1176/appi.ajp.2014.13081055>.
- [30] Cassell, J. et al. 2001. BEAT: the Behavior Expression Animation Toolkit. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01* (New York, New York, USA, 2001), 477–486.

- [31] Cassell, J. 2000. *Embodied conversational agents*. MIT press.
- [32] Cassell, J. et al. 2000. Human Conversation as a System Framework: Designing Embodied Conversational Agents. *Embodied conversational agents*. J. Cassell, ed. MIT press.
- [33] Cassell, J. et al. 2001. Non-verbal cues for discourse structure. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*. (2001), 114–123. DOI:<https://doi.org/10.3115/1073012.1073028>.
- [34] Cassell, J. et al. 1999. Turn taking vs. discourse structure: How best to model multimodal conversation. *Machine conversations*. (1999), 143–154.
- [35] Center for Substance Abuse Treatment 1999. *Enhancing Motivation for Change in Substance Abuse Treatment. Treatment Improvement Protocol (TIP) Series, No. 35*.
- [36] Cho, K. et al. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. (2014).
- [37] Clark, H.H. and Brennan, S.E. 1991. Grounding in communication. (1991).
- [38] Clark, H.H. and Schaefer, E.F. 1989. Contributing to discourse. *Cognitive science*. 13, 2 (1989), 259–294.
- [39] DeVault, D. et al. 2004. Natural Language Generation and Discourse Context: Computing Distractor Sets from the Focus Stack. *FLAIRS Conference* (2004), 887–892.
- [40] Devlin, J. et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. (2018).
- [41] Edunov, S. et al. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*. (2018).

- [42] Efird, J. 2011. Blocked randomization with randomly selected block sizes. *International journal of environmental research and public health*. 8, 1 (Jan. 2011), 15–20.
DOI:<https://doi.org/10.3390/ijerph8010015>.
- [43] Firth, J.R. 1957. A synopsis of linguistic theory 1930-55. Reprinted in Palmer FR (ed.), (1968) Selected papers of JR Firth 1952-1959. Longman, London.
- [44] Fowler, L.A. et al. 2016. Mobile technology-based interventions for adult users of alcohol: a systematic review of the literature. *Addictive behaviors*. 62, (2016), 25–34.
- [45] Gamito, P. et al. 2013. Executive functioning in addicts following health mobile cognitive stimulation: Evidence from alcohol and heroin patients. *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare* (2013), 385–388.
- [46] Gaume, J. et al. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of substance abuse treatment*. 37, 2 (2009), 151–159.
- [47] Geisner, I.M. et al. 2004. The relationship among alcohol use, related problems, and symptoms of psychological distress: Gender as a moderator in a college sample. *Addictive Behaviors*. 29, 5 (2004), 843–848.
- [48] Gibson, J. et al. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment*. 111, (2016), 21.
- [49] Glanz, K. et al. 2008. *Health behavior and health education: theory, research, and practice*. John Wiley & Sons.
- [50] Golub, D. and He, X. 2016. Character-Level Question Answering with Attention. (Apr.

2016).

- [51] Grice, H.P. 1975. Logic and conversation. *Speech acts*. Brill. 41–58.
- [52] Grosz, B.J. 1977. *The representation and use of focus in dialogue understanding*. SRI International Menlo Park United States.
- [53] Grosz, B.J. and Sidner, C.L. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*. 12, 3 (1986), 175–204.
- [54] Grosz, B.J. and Sidner, C.L. 1988. *Plans for discourse*. BBN LABS INC CAMBRIDGE MA.
- [55] Gustafson, D.H. et al. 2014. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA Psychiatry*. 71, (2014). DOI:<https://doi.org/10.1001/jamapsychiatry.2013.4642>.
- [56] Gustafson, D.H. et al. 2011. An e-health solution for people with alcohol problems. *Alcohol Research & Health*. 33, 4 (2011), 327.
- [57] Heather, N. et al. 1999. Development of a treatment version of the Readiness to Change Questionnaire. *Addiction Research*. 7, 1 (1999), 63–83.
- [58] Hester, R.K. et al. 2013. Overcoming Addictions, a Web-based application, and SMART Recovery, an online and in-person mutual help group for problem drinkers, part 1: three-month outcomes of a randomized controlled trial. *Journal of Medical Internet Research*. 15, 7 (2013), e134.
- [59] Hill, C.E. and Nakayama, E.Y. 2000. Client-centered therapy: Where has it been and where is it going? A comment on Hathaway (1948). *Journal of Clinical Psychology*. 56, 7

- (2000), 861–875.
- [60] Hochreiter, S. and Schmidhuber, J. 1997. LSTM can solve hard long time lag problems. *Advances in neural information processing systems (1997)*, 473–479.
- [61] Hori, T. et al. 2016. Dialog state tracking with attention-based sequence-to-sequence learning. *2016 IEEE Spoken Language Technology Workshop (SLT) (2016)*, 552–558.
- [62] Horvath, A.O. and Greenberg, L.S. 1989. Development and validation of the Working Alliance Inventory. *Journal of counseling psychology*. 36, 2 (1989), 223.
- [63] Ilievski, V. et al. 2018. Goal-oriented chatbot dialog management bootstrapping with transfer learning. *arXiv preprint arXiv:1802.00500*. (2018).
- [64] Irvin, J.E. et al. 1999. Efficacy of relapse prevention: A meta-analytic review. *Journal of Consulting and Clinical Psychology*. 67, 4 (1999), 563–570.
DOI:<https://doi.org/10.1037/0022-006X.67.4.563>.
- [65] Jacobs, W. et al. 2017. Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. *Cogent Social Sciences*. 3, 1 (2017), 1302785.
- [66] De Jonge, J.M. et al. 2005. The motivational interviewing skill code: Reliability and a critical appraisal. *Behavioural and Cognitive Psychotherapy*. 33, 3 (2005), 285–298.
- [67] K Nair, N. et al. 2015. A systematic review of digital and computer-based alcohol intervention programs in primary care. *Current drug abuse reviews*. 8, 2 (2015), 111–118.
- [68] Kahler, C.W. et al. 2016. Using topic coding to understand the nature of change language in a motivational intervention to reduce alcohol and sex risk behaviors in emergency

- department patients. *Patient education and counseling*. 99, 10 (2016), 1595–1602.
- [69] Kearns, M.J. et al. 1994. *An introduction to computational learning theory*. MIT press.
- [70] Keskar, N.S. et al. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*. (2019).
- [71] Khirasaria, R. et al. 2020. Exploring digital therapeutics: The next paradigm of modern health-care industry. *Perspectives in Clinical Research*. 11, 2 (2020), 54.
- [72] Kingma, D.P. and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. (2014).
- [73] Lafferty, J. et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).
- [74] Laws, M.B. et al. 2018. A sequential analysis of motivational interviewing technical skills and client responses. *Journal of Substance Abuse Treatment*. 92, (2018), 27–34.
DOI:<https://doi.org/https://doi.org/10.1016/j.jsat.2018.06.006>.
- [75] Lee, C.-J. et al. 2010. Recent approaches to dialog management for spoken dialog systems. *Journal of Computing Science and Engineering*. 4, 1 (2010), 1–22.
- [76] Lisetti, C. et al. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*. 4, 4 (2013), 19.
- [77] Lisetti, C. et al. 2015. Now all together: overview of virtual health assistants emulating face-to-face health interview experience. *KI-Künstliche Intelligenz*. 29, 2 (2015), 161–172.
- [78] Lucas, G.M. et al. 2014. It’s only a computer: Virtual humans increase willingness to

disclose. *Computers in Human Behavior*. 37, (2014).

DOI:<https://doi.org/10.1016/j.chb.2014.04.043>.

- [79] Magill, M. et al. 2019. A meta-analysis of cognitive-behavioral therapy for alcohol or other drug use disorders: Treatment efficacy by contrast condition. *Journal of consulting and clinical psychology*. 87, 12 (2019), 1093.
- [80] Markland, D. et al. 2005. Motivational interviewing and self-determination theory. *Journal of social and clinical psychology*. 24, 6 (2005), 811–831.
- [81] Marlatt, G. and Gordon, J. 1985. *Relapse prevention: maintenance strategies in the treatment of addictive disorders*. Guilford Press.
- [82] Marlatt, G.A. and Donovan, D.M. 2005. *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors*. Guilford Press.
- [83] Mayfield, E. et al. 2013. Recognizing rare social phenomena in conversation: Empowerment detection in support group chatrooms. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2013), 104–113.
- [84] McLellan, A.T. et al. 2000. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. *Jama*. 284, 13 (2000), 1689–1695.
- [85] McTear, M. et al. 2016. *The Conversational Interface: Talking to Smart Devices*: Springer International Publishing. *Doi: <https://doi.org/10.1007/978-3-319-32967-3>*. (2016).
- [86] Meier, P.S. et al. 2005. The role of the therapeutic alliance in the treatment of substance misuse: a critical review of the literature. *Addiction*. 100, 3 (2005), 304–316.

- [87] Meilman, P.W. et al. 1997. Average weekly alcohol consumption: Drinking percentiles for American college students. *Journal of American College Health*. 45, 5 (1997), 201–204.
- [88] Mikolov, T. et al. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. (2013).
- [89] Miller, W.R. and Rollnick, S. 2012. *Motivational interviewing: Helping people change*. Guilford press.
- [90] Moyers, T.B. et al. 2003. The motivational interviewing treatment integrity (MITI) code: Version 2.0. Retrieved from *Verfügbar unter: www.casaa.unm.edu [01.03. 2005]*. (2003).
- [91] Newman, M.G. et al. 2011. A review of technology-assisted self-help and minimal contact therapies for drug and alcohol abuse and smoking addiction: is human contact necessary for therapeutic efficacy? *Clinical psychology review*. 31, 1 (2011), 178–186.
- [92] Noar, S.M. et al. 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological bulletin*. 133, 4 (2007), 673.
- [93] Nooraei, B. et al. 2014. A Real-Time Architecture for Embodied Conversational Agents: Beyond Turn-Taking. *ACHI 2014, The Seventh International Conference on Advances in Computer-Human Interactions*. (2014).
- [94] Olafsson, S. et al. 2019. Coerced change-talk with conversational agents promotes confidence in behavior change. *ACM International Conference Proceeding Series* (2019).
- [95] Olafsson, S. et al. 2020. Motivating health behavior change with humorous virtual agents. *Proceedings of the 20th ACM international conference on intelligent virtual agents* (2020), 1–8.

- [96] Olafsson, S. et al. 2020. Towards a Computational Framework for Automating Substance Use Counseling with Virtual Agents. *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (2020), 966–974.
- [97] Olafsson, S. et al. 2020. Towards a Computational Framework for Automating Substance Use Counseling with Virtual Agents. *Autonomous Agents and Multi-Agent Systems (AAMAS)* (2020).
- [98] Olafsson, S. et al. 2018. Virtual Counselor for Patients in Medication-Assisted Treatment for Opioid Use. *GREATS2018 at the international conference on Intelligent Virtual Agents* (Sydney, NSW, 2018).
- [99] Oliveira, J. et al. 2015. Cognitive stimulation of alcoholics through VR-based Instrumental Activities of Daily Living. *Proceedings of the 3rd 2015 Workshop on ICTs for improving Patients Rehabilitation Research Techniques* (2015), 14–17.
- [100] Paszke, A. et al. 2017. Automatic differentiation in pytorch. (2017).
- [101] Pedrelli, P. et al. 2013. Combined MI+ CBT for depressive symptoms and binge drinking among young adults: Two case Studies. *Journal of cognitive psychotherapy*. 27, 3 (2013), 235–257.
- [102] Pedrelli, P. et al. 2020. Evaluating the combination of a Brief Motivational Intervention plus Cognitive Behavioral Therapy for Depression and heavy episodic drinking in college students. *Psychology of Addictive Behaviors*. 34, 2 (2020), 308.
- [103] Pérez-Rosas, V. et al. 2017. Predicting Counselor Behaviors in Motivational Interviewing Encounters. *Proceedings of the 15th Conference of the European Chapter of the*

Association for Computational Linguistics: Volume 1, Long Papers (2017), 1128–1137.

- [104] Peters, M.E. et al. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*. (2018).
- [105] Peters, M.E. et al. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*. (2017).
- [106] Petty, R.E. and Cacioppo, J.T. 1996. *Attitudes and persuasion: Classic and contemporary approaches*. Westview Press.
- [107] Pinto, R.Z. et al. 2012. Patient-centred communication is associated with positive therapeutic alliance: a systematic review. *Journal of physiotherapy*. 58, 2 (2012), 77–87.
- [108] Pollak, K.I. et al. 2011. Physician empathy and listening: associations with patient satisfaction and autonomy. *The Journal of the American Board of Family Medicine*. 24, 6 (2011), 665–672.
- [109] Prochaska, J.O. and Velicer, W.F. 1997. The Transtheoretical Model of Health Behavior Change. *American Journal of Health Promotion*. 12, 1 (Sep. 1997), 38–48.
DOI:<https://doi.org/10.4278/0890-1171-12.1.38>.
- [110] Quality, C. for B.H.S. and 2018. 2017 National Survey on Drug Use and Health: Detailed tables. (2018).
- [111] Radford, A. et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*. 1, 8 (2019), 9.
- [112] Rapp, R.C. et al. 2006. Treatment barriers identified by substance abusers assessed at a centralized intake unit. *Journal of substance abuse treatment*. 30, 3 (2006), 227–235.

- [113] Rehurek, R. and Sojka, P. 2010. Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (2010).
- [114] Resnicow, K. and McMaster, F. 2012. Motivational Interviewing: moving from why to how with autonomy support. *International Journal of Behavioral Nutrition and Physical Activity*. 9, 1 (2012), 19.
- [115] Rich, C. and Sidner, C. 2012. Using collaborative discourse theory to partially automate dialogue tree authoring. *Intelligent Virtual Agents* (2012), 327–340.
- [116] Rich, C. and Sidner, C.L. 2010. Collaborative Discourse, Engagement and Always-On Relational Agents. *AAAI Fall Symposium: Dialog with Robots* (2010).
- [117] Rich, C. and Sidner, C.L. 1998. COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction*. 8, 3 (1998), 315–350.
- [118] Rickel, J. et al. 2002. Collaborative discourse theory as a foundation for tutorial dialogue. *International Conference on Intelligent Tutoring Systems* (2002), 542–551.
- [119] Rieger, B.B. 1991. *On distributed representation in word semantics*. International Computer Science Institute Berkeley, CA.
- [120] Riper, H. et al. 2009. Curbing problem drinking with personalized-feedback interventions: a meta-analysis. *American journal of preventive medicine*. 36, 3 (2009), 247–255.
- [121] Rollnick, S. et al. 1992. Development of a short ‘readiness to change’ questionnaire for use in brief, opportunistic interventions among excessive drinkers. *British journal of addiction*. 87, 5 (1992), 743–754.

- [122] Roque, A. and Traum, D.R. 2009. Improving a Virtual Human Using a Model of Degrees of Grounding. *IJCAI* (2009), 1537–1542.
- [123] Roy, N. et al. 2000. Spoken dialogue management using probabilistic reasoning. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (2000), 93–100.
- [124] Rubak, S. et al. 2005. Motivational interviewing: a systematic review and meta-analysis. *Br J Gen Pract.* 55, 513 (2005), 305–312.
- [125] Sacks, H. et al. 1978. A simplest systematics for the organization of turn taking for conversation. *Studies in the organization of conversational interaction*. Elsevier. 7–55.
- [126] Saunders, J.B. et al. 1993. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction.* 88, 6 (1993), 791–804.
- [127] Schatzmann, J. et al. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* (2007), 149–152.
- [128] Schmidhuber, J. 2015. Deep learning in neural networks: An overview. *Neural networks.* 61, (2015), 85–117.
- [129] Schulman, D. et al. 2011. An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change Using Motivational Interviewing. *AAAI Spring Symposium: AI and Health.* (2011), 1–4.

- [130] Skantze, G. 2003. Exploring human error handling strategies: Implications for spoken dialogue systems. *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems* (2003).
- [131] Substance Abuse and Mental Health Services Administration 2018. Key Substance Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health. *Substance Abuse and Mental Health Services Administration*. Retrieved from <https://www.samhsa.gov/data>. (2018).
- [132] Thombs, D.L. et al. 2007. Outcomes of a technology-based social norms intervention to deter alcohol use in freshman residence halls. *Journal of American College Health*. 55, 6 (2007), 325–332.
- [133] Traum, D. 1996. Conversational agency: The TRAINS-93 dialogue manager. In Susann LuperFoy, Anton Nijhholt, and Gert Veldhuijzen van Zanten, editors, *Proceedings of Twente Workshop on Language Technology, TWLT-II* (1996).
- [134] Traum, D.R. and Allen, J.F. 1994. Discourse obligations in dialogue processing. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (1994), 1–8.
- [135] Vader, A.M. et al. 2010. The language of motivational interviewing and feedback: counselor language, client language, and client drinking outcomes. *Psychology of Addictive Behaviors*. 24, 2 (2010), 190.
- [136] Vaswani, A. et al. 2017. Attention is all you need. *Advances in neural information processing systems* (2017), 5998–6008.

- [137] Volkow, N.D. et al. 2016. Neurobiologic Advances from the Brain Disease Model of Addiction. *The New England journal of medicine*. 374, 4 (2016), 363–71.
DOI:<https://doi.org/10.1056/NEJMra1511480>.
- [138] Walker, M.A. 1996. Limited attention and discourse structure. *Computational Linguistics*. 22, 2 (1996), 255–264.
- [139] Wallace, B.C. et al. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013), 1765–1775.
- [140] Wallace, B.C. et al. 2014. Automatically Annotating Topics in Transcripts of Patient-Provider Interactions via Machine Learning. *Medical Decision Making*. 34, 4 (May 2014), 503–512. DOI:<https://doi.org/10.1177/0272989X13514777>.
- [141] Wallace, B.C. et al. 2014. Identifying Differences in Physician Communication Styles with a Log-Linear Transition Component Model. *AAAI* (2014), 1314–1320.
- [142] Wallace, E. et al. 2019. Universal adversarial triggers for nlp. *arXiv preprint arXiv:1908.07125*. (2019).
- [143] Wang, K.-C. et al. 2013. A phone-based support system to assist alcohol recovery. *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)* (New York, New York, USA, 2013), 529–534.
- [144] Wheelless, L.R. and Grotz, J. 1977. The Measurement of Trust and Its Relationship to Self-Disclosure. *Human Communication Research*. 3, 3 (Mar. 1977), 250–257.
DOI:<https://doi.org/10.1111/j.1468-2958.1977.tb00523.x>.

- [145] Wolf, T. et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*. (2019).
- [146] Wu, W.-L. et al. 2010. Spoken language understanding using weakly supervised learning. *Computer speech & language*. 24, 2 (2010), 358–382.
- [147] Xiong, W. et al. 2017. The Microsoft 2017 Conversational Speech Recognition System. (Aug. 2017).
- [148] Yaghoubzadeh, R. and Kopp, S. 2016. flexdiam—Flexible dialogue management for incremental interaction with virtual agents (demo paper). *International Conference on Intelligent Virtual Agents* (2016), 505–508.
- [149] Yasavur, U. et al. 2014. Intelligent virtual agents and spoken dialog systems come together to deliver brief health interventions. *Journal on Multimodal User Interfaces, in press*. 1, 1 (2014), 19.
- [150] You, C.-W. et al. 2018. Using Mobile Phones to Facilitate Alcohol Dependent Patients to Improve Family Communication. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (2018), 311–314.
- [151] Young, S. et al. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*. 101, 5 (2013), 1160–1179.
- [152] Young, S. et al. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*. 24, 2 (2010), 150–174.

[153] Zhou, S. et al. 2017. A Relational Agent for Alcohol Misuse Screening and Intervention in Primary Care. *CHI'17 Workshop on Interactive Systems in Healthcare* (2017).