H2020-MSCA-ETN-2018-SAS. Proposal Number 812788

European Training Network on

Safer Autonomous Systems (SAS)

Deliverable 3.1 (D3) – Literature Review of Safety Challenges associated with the use of Intelligent Conversational Bots

Authors: Haris Aftab, Ibrahim Habli (University of York)

This project has received funding from the European Union's
EU Framework Programme for Research and Innovation
Horizon 2020 under Grant Agreement No. 812.788

Abstract

Conversational Agents (CAs) are artificial intelligence (AI) based software programs that interact with users via speech or text in natural language. Due to their intelligence, intuitiveness, and human-like conversational style, they are used in many industries which include healthcare, e-commerce, sports, aviation, media, banking and finance, travel, etc. CAs can help healthcare with overcoming staff shortage, providing low-cost services to patients, 24 hours accessibility, assisting clinicians with large amounts of data and extracting information, etc. Modern voice user interfaces (VUIs) such as Google Home and Amazon Echo are AI-enabled devices with programmable software applications. Amazon's Alexa currently has more than 1000 skills (software application for Alexa platform) for health and fitness category. CAs are being used for mental health and wellbeing, oncology, fitness, symptom checking, self-diagnosis, medication adherence, etc. CAs may provide harm to users because of errors in speech recognition, natural language understanding (NLU) failures, and improper response generation. The inherent AI issues, lack of clinical data for training, unconstrained user input, and not having situational awareness (SA) has the potential to harm patients. CAs may fail to provide a safe response from errors arising within the architecture (speech recognition, NLU, response generation) and when interacted with humans (noisy environment, unconstrained input). CAs in a clinical environment also has impact on certain human factors such as patient-clinician interaction, automation bias, handover and human performance. While there are many potential benefits of CAs, there are various ethical challenges associated with them such as the risk of bias (data and design bias), privacy and risk of harm to the users.

Contents

List of Tables

List of Figures

List of Abbreviations

Conversational Agent	(CA)
Artificial Intelligence	(AI)
Machine Learning	(ML)
Voice User Interface	(VUI)
Voice Assistant	(VA)
Goal Structuring Notation	(GSN)
World Health Organization	(WHO)
National Reporting and Learning System	(NRLS)
Royal College General Practitioner Exam	(RCGP)
Clinical Skills Assessment	(CSA)
Applied Knowledge Test	(AKT)
Claims Arguments Evidence	(CAE)
Food and Drug Administration	(FDA)
International Organization for Standardisation	(ISO)
Medicines and Healthcare products Regulatory Agency	(MHRA)
Generic Infusion Pump	(GIP)
As Low As Reasonably Practicable	(ALARP)
Failure Mode and Effects Analysis	(FMEA)
Healthcare Failure Mode and Effect Analysis	(HFMEA)
Failure Mode, Effects and Criticality Analysis	(FMECA)
Hierarchical Task Analysis	(HTA)
Systematic Human Error Reduction and Prediction Approach	(SHERPA)
Natural Language Understanding	(NLU)
Automatic Speech Recognition	(ASR)
Text to Speech	(TTS)
Spoken Language Understanding	(SLU)
Dialogue Acts	(DA)
Natural Language Generation	(NLG)
Speech to Text	(STT)
Clinical Decision Support System	(CDSS)
Age, Time, Mechanism, Injuries, Signs, Treatments	(ATMIST)
Situation, Background, Assessment, Recommendation	(SABR)
Natural Language Processing	(NLP)
Part-of-Speech	(POS)
Named Entity Recognition	(NER)
Bag of Words	(BoW)
Hidden Markov Model	(HMM)

1. Introduction

Conversational Agents (CAs) [1][2][3][4][5], also known as dialogue systems [6][7], intelligent conversational agents [8][9][10], conversational assistants [11][12], or virtual assistants [13][14] are Artificial Intelligence (AI) based software programs that simulate conversation with a human using natural language via text or speech. The terms voice-assistants [15], voice interfaces [16][17] are also used for these systems which are activated solely through voice. The term 'Chatbot' which is probably the most common term is also used similarly in the literature [18][19][20][21][22][23]. Section 2.3 discusses the difference between CAs and chatbots with their architecture in detail.

CAs currently have many applications in different sectors which are not limited to e-commerce, travel, sports, and healthcare [24][25][26][27]. VAs and VUIs have added more value to traditional text-based CAs by engaging users more interactively. The rise of AI and Machine Learning (ML) has enabled CAs to learn continuously from the user's input and provide services intelligently. Smart speakers are becoming more and more common in our homes to perform home automation tasks, listening to news and music, and controlling various household devices. Google Assistant, Amazon's Alexa, Siri from Apple are some of the most popular Voice Assistants (VAs).

To simplify the terminologies associated with CAs, we have divided them into three main categories (1) CA, (2) VA, and (3) VUI. Throughout this report, these terminologies are used to describe these systems in appropriate contexts.

- We use ***'CA'*** as a general term to address all types of AI conversational systems (software-based) which communicate in natural language and can be either activated by voice or text. This also includes VUI and therefore VA. The term VA and VUI will be used where the emphasis is on voice-activated CAs.
- We use the term ***'VA'*** for all conversational systems (software-based) including Google Assistant, Amazon's Alexa and Apple' Siri that interact in natural language and are activated by voice command.
- We use ***'VUI'*** to address VA (software-based) and standalone (hardware) devices such as Google Home (powered by Google Assistant)[1] and Amazon Echo[2] (powered by Alexa technology) which use VA inside them.

The rate of adoption for CAs in healthcare is growing because of their ease of use and intuitiveness. They are being used in monitoring mental health [1][28][29] , oncology [30][31], providing diagnostic decision support [32], symptom checking [33][34], medication adherence [35][36], etc. CAs are increasingly popular as they can provide an early insight into the patient's situation [37]. Healthcare is a safety-critical domain and the risk to lives of people is greater than in other industries because of the condition of patients [38], complex processes, lack of automation, etc. The use of CAs in healthcare may also be a source of harm to patients if they are used without proper supervision [39][40]. Training on incorrect data affects the performance of CAs [41] and it may give rise to safety issues in healthcare.

[1] Google Home is a speaker device which integrates AI-enabled Google Assistant to provide hands-free smarter control to users. Google Home, https://store.google.com/gb/product/google_home

[2] Amazon Echo is a hands-free speaker controlled by voice. It connects to Alexa – a cloud-based service which continually learns, and adds new functionalities via its skills to provide smart services. Amazon Echo, https://www.amazon.co.uk/amazon-echo-3rd-generation-smart-speaker-with-alexa/dp/B07P4DKX14?th=1

Speech recognition problems and inherent AI safety issues [42] in modern VUIs make their use in healthcare even more challenging [43]. CAs can also fall in low to medium risk medical devices [44] and healthcare safety regulations may apply to those CAs.

A safety case is used as a standard practice in many safety-critical industries. Safety cases can be represented by long written documents or by using graphical schemes. A safety case is required by the regulators in safety-critical industries and they provide a common platform to all stakeholders about the safety of the system. They identify key risks in a system and demonstrates the approaches and controls put in place to mitigate those risks to an acceptable level of safety. A safety case represents an overall claim about the safety of the system with safety arguments explicitly showing that the evidence satisfies those claims. Goal Structuring Notation (GSN) [45] is widely used as a graphical representation of the safety case. A safety analysis method is used to identify hazards and potential risks to the safety of a system. The safety case approach is relatively new in healthcare and there are various analysis techniques acquired from other industries such as checklists, FMEA, HTA, etc.

CAs are of two types: (1) Chatbots which are used for extended conversation and (2) Task-oriented dialogue agents for accomplishing a specific task [46]. This literature review is based on the second category of CAs. The literature review in this report discusses the need for CAs in healthcare, their differences with other safety-critical industries, and conversational applications in healthcare. Safety of these CAs is the main topic of this report and it is discussed in detail along with safety techniques and safety analysis methods in healthcare. A lot of safety issues arise from the internal architecture of CAs and this report describes the architecture and safety failure modes of the CAs. The introduction of AI in clinics introduces various human factor challenges and since CAs use AI for the decision making they are also part of this report. There are some ethical issues with the use of CAs especially in healthcare which are described at the end of this report.

Report Structure

The rest of the report is organized as follows:

- **Chapter 2** presents the background and current literature on the research topic. This chapter is divided into four subsections:
    i. Section 2.1 provides healthcare overview in the context of CAs and other safety-critical industries. This section then covers applications of CAs in healthcare and their potential safety implications.
    ii. Section 2.2 provides details on safety assurance in healthcare with a focus on safety case and safety analysis methods.
    iii. Section 2.3 discusses different types of CAs and their architecture. This section also lists failure modes of VUIs from the literature.
    iv. Section 2.4 features important human factors that need consideration for AI to be introduced in clinical care.
    v. Section 2.5 lists ethical issues concerning the use of CAs.
- **Chapter 3** highlights important topics discussed and concludes the report.

2. Context

This chapter provides the safety challenges of CAs in the context of healthcare, the technology behind these agents and associated human factors. Ethical issues concerning CAs are also discussed at the end of this chapter.

2.1. Healthcare

2.1.1. The need for CAs

Artificial intelligence has offered a lot of convenience across many industries and healthcare is starting to realise those benefits by welcoming AI-based solutions. CAs are one of those AI-based solutions that are increasingly becoming popular in healthcare. These assistants are available on smartphones from years in the form of VAs and are easily accessible to a vast number of people. For example, Siri available on Apple smartphone alone has more than 500 million active users, and there are more than a billion voice searches a month [47]. Many people believe to found CAs more accurate than search on the web [48]. Since the advent of VUIs, people are increasingly using them for health information. Amazon now boasts over 1000 skills for its Alexa devices in the health and fitness category [49]. There are a couple of areas where CAs can improve healthcare functions and overall quality. We have identified three areas where CAs can improve patient services which are accessibility, patient engagement, and assisting clinicians.

Accessibility is one of three areas where we believe CAs can play a crucial role in improving services to patients. Healthcare is limited in terms of providing access to patients with information. Patients may have medical concerns outside the usual work hours. It is also not possible for all of the patients to use healthcare facilities because of cost, convenience, scheduling constraints, and other factors. CAs do not need sleep because they don't get tired, fatigued, or sick and they are cost-effective to operate and are available 24 hours a day [50]. They are also helpful in providing immediate medical information, recommend diagnoses early on, and can find healthcare providers for patients [2][25][51].

Patient engagement is another area where CAs can help healthcare staff as staff shortage is a big problem and the time a clinician can spend with their patients is limited [52]. CAs with their conversational interface possess the capability to engage, track, educate, and prevent patients from some behaviours [37]. These systems are more capable of approximating the face-to-face interaction with clinician than any other medium. In some cases, they might be even better than the interaction between a patient and a clinician. For example, it is noted that the patients are willing to share more medical information and reporting symptoms to a CA than a human [53]. The participants in a study expressed their feelings better and were more comfortable talking to a virtual system [54]. CAs can provide empathy to patients by adding emotional comments during the conversation based on the user's responses [22]. There may be some instances when a CA may be better suited to cater the needs of the patients because they are gender insensitive, do not belong to any race and elicit no bias to patients based on their demographics [55]. CAs can communicate in more than one language and therefore in that respect, they are better suited to patients' needs. Personalization is another aspect as CAs can provide services to patients at a fine-grained level, for example responding to requests of patients such as clarification or fulfilling requests of additional information.

CAs are a technological development that can improve the performance of clinicians because of their ability to process a large amount of data in a short time. For example, large information about drugs and their side effects [14]. Safedrugbot is a CA which helps health care professionals to obtain information about the safety of drugs for breastfeeding women [56]. CAs can also help in providing information to patients and reduce the need for medical staff for each task. Baidu developed a CA named Melody which delivers essential information to medical staff and can also be used to make medical appointments based on their schedule [22].

The use of CAs in healthcare has many applications and are discussed in more detail in the coming section. Here we take one example of mental healthcare where their application can be highly beneficial.

Mental health is a serious issue and it can lead people to depression and suicide. CAs can greatly help in reducing mental health issues as they can create a humanized experience to the users and they may feel as though they are talking to a real person. For example, it is estimated that 300 million or 4.4% of the world's population (2015) can be affected by depression alone [57]. The severity of it can lead to suicide which was globally the second leading cause of death among people of age 15-29 years [58]. The shortage of healthcare staff in psychiatry [22] also highlights the need for CAs in mental health. It is approximated to have only 9 psychiatrists per 100,000 people in developed countries and as few as 0.1 for every 1,000,0004 in lower-income countries [41].

In the context of the applicability of safety, healthcare differs to quite an extent to other safety-critical industries such as nuclear power, aviation, railways, etc. Next section provides details on the differences between healthcare and other safety-critical domains.

2.1.2. Healthcare vs other Safety-Critical Domains

The concept of safety-criticality of a system was originated from two properties mass and dread [59]. The events of mass casualty where many people die capture public attention in a manner that even a larger number of deaths over a large number of small episodes does not. Healthcare risks leading to injury or death occur over such small episodes and they do not draw much attention. The other reason healthcare is treated differently is because many adverse events in safety-critical industries are associated with feelings of dread, unwilling participation in risks, loss of control, etc. In healthcare, most of the adverse events are indistinguishable from natural events. As Gaba et al. [60] mention as every human is destined to die and there is a strong chance of it happening in the vicinity of healthcare. The deaths and disabilities are viewed as normal events and even some unexpected events which in industries like nuclear power or aviation may never consider as normal.

Healthcare practices a diverse set of activities and principles. Such as highly hazardous surgery; primary care where a patient develops a kind of relationship with their doctors; unpredictable and constantly changing emergency medicine, etc. [61]. This list may add hospital medicine, care in the community home care, etc. Such diversity of activities makes healthcare more complex industry with little common with other safety-critical industries. Work in other industries is usually predictable and comprises of routine tasks. To practice reliability and safety it is a normal practice to avoid situations or practice that are not part of safe routines. The level of uncertainty in healthcare is quite large and it depends mostly on more than one factor. For example, a medicine's effect on patient's condition is dependent on patient's current health, the ability of the immune system to fight its side effects, the involvement of heredity and genetic effects etc.

In other industries, mostly the routines are monitored for a system in operation. For example, a running plane or chemical plant. The work in healthcare though is hands-on and potentially prone to more errors [61]. People are in good health in other industries while in healthcare they are sick or injured or have disabilities and are vulnerable to even very small errors in their care.

On the organizational level, there are also vast differences between healthcare and other industries. According to [62], most high-risk industries are well structured with centralized control. Healthcare is quite decentralized in this regard, consider for instance only the NHS in England. In other industries, there is a strong emphasis on standardizing work and training processes [63]. In commercial aviation, the pilot has certain capabilities and they can easily be interchangeable between flights. In healthcare, a physician even on the same level of qualification cannot be compared in this way. There would be a potentially greater level of error if we try to interchange clinicians and nurses in a way we may change pilots. This interchange may increase unsafe behaviour in the healthcare system.

The regulation in healthcare is still not at the level of aviation or other industries. Humans have cognitive limits and when these limits exceed the probability of making mistakes increases. Consider, for example, fatigue which is an important variable in human health, is recognized very slowly. In a study [64], only 30% of surgeons willing to admit that their performance is worse without proper sleep. For such things aviation has governing policies: for example, pilots and controllers have mandatory retirement ages, and time-on-duty limits that recognize the negative impact of fatigue and they have recertification requirements [65]. Also, pilots have to undergo proficiency check in every six months while in the UK now doctors undergo revalidation after five years [66].

Healthcare shares similar traits to the nuclear industry because both are complex, tightly coupled, and sociotechnical systems. In contrast to nuclear power, healthcare has evolved organically over some time. The nuclear industry, on the other hand, is engineered and each plant is designed and built specifically to be part of an integrated system. Primary processes in healthcare are poorly understood as compared to the physical processes associated with nuclear power [67]. Healthcare thus also has far more uncertainty and hence risk at the subsystem level (patient care) than in nuclear power.

Healthcare should not just adopt solutions from other industries but compare and contrast organizational attributes and take safety measures which are applied in similar conditions [68]. For example, from aviation, they may learn crew resource management training in homogeneous teams. As pointed out by Macrae et al., investigations and monitoring in healthcare are not on par with other industries [63]. Healthcare should have an external investigation body comparable to for example the Air Accidents Investigation Board in the UK which investigates serious civil aircraft accidents in the UK. Healthcare already has taken lots of interventions from other industries which include: safety checklists [69], emergency manuals [70], failure mode and effect analysis [71], etc. Learning from other industries is not a straightforward and simple task and there is room for improvement in healthcare by understanding the mechanisms, systems, attitude, and values that underpin these techniques successfully in other industries.

Patient safety is at the forefront in the provision of safety in healthcare. It is related to the prevention of harm to a patient in providing safe healthcare. The coming section discusses patient safety and the degrees of harm patients come across in healthcare.

2.1.3. Patient Safety

Patient safety is defined as *"The avoidance, prevention, and amelioration of adverse outcomes or injuries stemming from the process of healthcare"* [38]. It is related to the quality of care a patient receives in a hospital or clinic but these two are not necessarily the same. Safety is expressed as a single dimension of quality as quality is a broad term. The quality of healthcare according to [61] is encapsulated by six dimensions which include safety, effectiveness, patient-centred, time, efficiency, and equitable. The Institute of Medicine report 'Crossing the Quality Chasm' provides a story about a working mother who suffered preventable but long-lasting disability because of poor quality of care. From this story, it is evident that there is less separating safety and quality of care. She suffered from this not by the harm caused directly by the drug or surgery but the inefficiency, delay, and non-patient centeredness. Safety is one dimension of quality of care which is the most critical and important to patients.

According to the World Health Organization (WHO), *"Patient safety is the absence of preventable harm to a patient during the process of health care and reduction of risk of unnecessary harm associated with health care to an acceptable minimum"* [72]. The acceptable minimum refers here the weight of given knowledge and resources in which quality is delivered against the risk involved in treatment or non-treatment. In the UK, the NHS defines it as *"Patient safety is about maximising the things that go right and minimising the things that go wrong for people experiencing healthcare"* [73]. This definition is more realistic because as humans there is still a chance of error, but it needs to be at a minimum level to provide better care for patients.

Prevention of harm to patients is the main constituent of patient safety as can be understood from the above definitions. Harms in healthcare are adverse outcomes such as death, injury or infections. These harms can be caused by healthcare hazards which include medication error, wrong dosage, infection from medical devices, etc. The healthcare process to a certain extent contains elements that can cause hazards and those hazards can ultimately harm patients.

Degrees of Harm

The national reporting and learning system (NRLS) provide five degrees of harm experienced by patients [74].

- **No harm:** a situation described by having no harm incident or a prevented safety incident.
- **Low harm:** an incident needing minor care and causing minimal harm to a person.
- **Moderate harm:** any incident resulting in further treatment which might include surgical intervention and caused short-term harm to a person.
- **Severe harm:** an incident causing permanent or long-term harm.
- **Death:** an event that results in the death of a person.

The ideal situation from the safety perspective is that no harm occurs to patients, but this may not be possible in every circumstance. However, it is still better to prevent serious harms and should be the objective of patient safety approaches. Technological tools such as CAs and VUIs in healthcare has the potential to assist healthcare in some areas but they may be a source of harm to patients. We will discuss the applications of CAs and VUIs in healthcare and the potential safety implications they may have on patients from the literature in the next section.

2.1.4. Applications of CAs in Healthcare

The use of CAs in healthcare is growing and there are a plethora of CA applications and research studies in this domain. They are being used for various purposes such as symptom checking, mental health being, fitness and health, overcoming stress, assisting in medical information, fighting obesity, for chronic diseases such as diabetes etc. Recent research studies show that most active health domains for CA are mental health [5][37][41] and physical wellness [37]. In the domain of oncology, numerous CAs are being used as well as shown by a recent research study [75]. From our literature review, we found out that most of the studies or application involving CAs are using text as a medium of the conversation however there are also studies involving VUIs. Table 1 summarizes some studies and examples from the healthcare domain.

In assisting users with their mental health problems there are numerous CA applications in healthcare. These studies show the usage of CA in improving psychological well-being, reducing anxiety or depression, adherence in task execution, antipsychotic medication adherence, and psychoeducation. Cameron et al. demonstrate the use of CAs in mental health by developing 'iHelpr' [76]. This CA provides guided self-assessment on stress, anxiety and depression, trauma, sleep, alcohol, etc. Similarly, Inkster et al. in their study evaluated a mental health well-being application which uses AI for its conversation [1]. They found the users with more usage of the conversational application less depressed than those with low usage. Mujeeb et al. demonstrated the use of CA for diagnosis of achluophobia and autism (fear of darkness) disorder in children by asking users a series of questions [28]. This CA concept can also be used in assisting human psychologists in clinics. Jungman et al. did their study on evaluating the diagnosis accuracy of Ada health [29]. They found application performance for mental health condition moderate to low. The authors noticed an increase in diagnostic accuracy when the application was used by psychotherapists. The study concluded that for the general population the app should be used with caution and suggested improvement in the app related to mental disorders in childhood and adolescence.

In the health and fitness area, there are also several applications and research studies. Huang et al. developed a CA 'SWITCHes' for obese people [77]. This CA with the help of auxiliary data from sensors help users in diet and exercise tracking and provides tailored feedback and advice including eating tips and eating order. Fadhil proposes a CA solution to promote healthy eating by preventing weight gain in adults [12]. Their proposed system provides a personalized recommendation regarding healthy diet, physical activity and healthy food preparation.

Some general-purpose CAs provide more than single functionality to users. For example, 'Mamabot' a proof of concept application that intends to assist pregnant woman is a general-purpose CA [25]. Users can search nearby pharmacies and hospitals in case of emergency provides symptom checking functionality, nutrition for children and children emergency management. Madhu et al. also proposes a general-purpose CA to provide age-based medicine usage, effective symptom prediction, and information about medicines [36]. In one study of VUIs, they were treated as general-purpose healthcare devices to check their accuracy and safety to the users [11]. Ma et al. have used Amazon Alexa based VUI to monitor patients' general health condition via angel sensor [16]. The sensor can monitor heart rate, blood oxygen, temperature, etc. of the user. Babylon health application [51] is also a general-purpose AI-powered application that provides symptom checking, locating clinics in the UK, booking appointments at the hospital, and talking to the doctor over a voice call.

Some CA applications or studies about them show their usage in monitoring chronic diseases such as diabetes where it is important to follow medication or control diet. One such CA is 'Vidi' which acts as a virtual dietitian for diabetic patients [78]. Similarly, the study by Levin et al. which showed the implementation of dialogue system to monitor chronic pain [76]. Ahmad et al. developed a medication reminder CA that also suggests medicine based on illness and the explanation about medicines [79]. Similar to this, 'Chester' VA reminds the patients about their medication, answering questions and engaging in a dialogue to collect information for improved monitoring [35].

Many studies also focus on the applications of CAs in oncology or helping cancer patients. Piau et al. implemented a semi-automated CA to provide older patients with cancer care at home [30]. The other purpose of this application was to free up nurses by collecting the primary patient data over the phone. In a study involving a CA 'Vik', the authors concluded that CA can be useful to cancer patients by providing them support and answers to their concerns about their disease [31]. Furthermore, Vik improved medication adherence through reminders and educational content.

In a study by Razzaki et al. for checking the diagnostic accuracy of Babylon AI Triage and Diagnostic system, they found it to be close to human doctors and in some cases exceeded the human performance [32]. It provided more accuracy than the average of human doctors in identifying the condition modelled by a clinical vignette. The vignettes in the study were from preparation materials for the RCGP (Royal College General Practitioner Exam) Clinical Skills Assessment (CSA) and Applied Knowledge Test (AKT) focusing on the diagnostic component. In the safety and appropriateness of triage recommendations, Babylon provided safer triage recommendations than the doctors on average. The results of this study seem very promising for an AI-based application. However, the claim of AI supremacy from Babylon over human clinicians was unconvincing as the study involved few doctors and result may be skewed because of poor judgement from any one subject. The data entered to the system was also done by doctors and not by lay users and the study offers unconvincing evidence as it was not evaluated in a realistic clinical environment [80].

Table 1: Summary of studies on CAs

Study	Healthcare domain	Category/Features	CA Name	Conversation Medium
Allen et al. [35]	Medication	Medication Advisor	Chester	Voice
Levin et al. [81]	Health Monitoring	Chronic Pain Assessment & Monitoring	Unknown	Voice
Ma et al. [16]	Health Monitoring	Health Answers via sensor data	Angel Echo	Voice
Cameron et al. [76]	Mental Healthcare	Self-Assessment	iHelpr	Text
Inkster et al. [1]	Mental Healthcare	Self-Assessment	Wysa	Text
Mujeeb et al. [28]	Mental Healthcare	Diagnostic App	Aquabot	Text
Huang et al. [77]	Health & Fitness	Obesity	SWITCHes	Text
Fadhil [12]	Health & Fitness	Healthy Lifestyle Promotion	Unknown	Text
Vaira et al. [25]	Prenatal Care	General Purpose	Mamabot	Text
Madhu et al. [36]	Medication	General Purpose (Symptom Based Disease Prediction)	Unknown	Text
Lokman et al. [78]	Health & Fitness	Virtual Dietitian	Vidi	Text
Ahmad et al. [79]	Medication	Medication Reminder	Unknown	Text
Piau et al. [30]	Oncology	Home care for Cancer Patients	Infinity	Text & Voice
Chaix [31]	Oncology	General Purpose (Medication Adherence, Educational Content, Patient Support)	Vik	Text
Jungmann et. al [29]	General and Mental Health	Diagnostic App Symptom Checker	Ada	Text
Razzaki et al. [32]	Medical Triage	Diagnostic App Symptom Checker	Babylon AI Triage and Diagnostic System	Text

Table 2: CAs with constrained user input

Healthcare domain	Category/Features	CA Name	Conversation Medium
General Health	General Purpose (Symptom Checker, Locate Clinic, Book Appointment, Talk to Doctor)	Babylon Health	Text (Constrained Input)
General Health	Symptom Checker	WebMD	Text (Constrained Input)
General Health	Symptom Checker	Ada	Text (Constrained Input)

The popular symptom checkers in healthcare include Babylon Health [51], WebMD [33], and Ada [34]. Babylon health app is can be considered as a general-purpose application as previously mentioned. WebMD and Ada both provide symptom checking for general health questions such as headache, flu, pain, etc. The reason for separating them from other CAs and the studies described above is the change they have adopted in their conversational approach. These applications previously provided unconstrained natural input for the user. However, the latest version of their applications has restricted the user to write the whole conversation. This approach is a step towards safe user input as previously mentioned in some studies that unconstrained input may cause user harm [40][5][82]. Table 2 provides a summary of these applications which use constrained input to provide safe interaction with users.

Safety Implications of Healthcare CAs

The increased research and focus on applying CAs to healthcare has profound impacts but at the same time, they come with numerous safety issues. Unlike humans, the risks from an autonomous system to users is quite large. CAs in healthcare may harm the patients as there is more than one factor while diagnosing a patient. They might miss the personal factors involved with the patients to suggest a recommendation [50]. Data is a key construct of any ML algorithm and due to non-rich data sets available for healthcare purposes, the performance of CAs might suffer [41]. To use CAs in healthcare training data sets must be available for the specific medical domain otherwise it may be unsafe to use those systems.

In a comparative study of CAs, it has been found that the systems with unconstrained input can cause harm to patients and should not be completely relied on [11][5][82]. The study shows that these CAs failed more than half of the time in situations that needed medical expertise. Furthermore, they led people to take actions that could have resulted in harm. These systems should be used under clinical supervision for the queries requiring medical expertise. A similar study involving smartphone VAs assessed their response when asked about suicidal emergencies. The results show limited and sometimes inappropriate responses [83]. Most of the VAs responded by a web search for users to explore further information.

Bibault et al. [41] noted from their review study done on CAs in oncology that scarcity of the clinical trials outweighs their potential benefits for patients and the healthcare system. The lack of objective evidence for the relevance and efficacy of CAs is alarming as they are poised to be used by more patients. Authors from a study [84] also suggested the need for having an option for users to talk to their counsellor when dealing with very sensitive health conditions such as AIDS. Further, a CA should have the capability to link users to trained professional whenever user mentions suicidal thoughts or self-harm [84]. Hodgson et al. in a study for checking the efficiency of speech recognition for electronic health records found an increased risk of errors with the potential to cause patient harm [43].

These studies show that CAs are not completely safe for self-diagnosing. This is due to problems in understanding user's query, lack of appropriate training data, absence of clinician's monitoring and insufficient clinical trials and evidence of their effectiveness in their application. Therefore, it is essential to think of methods and approaches that can make their usage safe in addition to their potentially huge benefits in healthcare.

Summary

This section covered healthcare in the context of patient safety; differences with other safety-critical industries; need, applications, and safety implications of CAs. The key takeaway from this section is that there is a need to distinguish healthcare from other industries to approach safety. Healthcare is different because of its evolution, diversity of processes, and already weaker health of patients at the time of their visit. Healthcare hazards can cause various degrees of harm to patients and CAs may be beneficial in certain healthcare areas to prevent those harms. The 24-hour accessibility, ability to engage patients much like humans and computational power to assist clinicians are the key benefits of introducing CAs in healthcare. They are being applied in healthcare in the areas of mental health, oncology, physical health and fitness, and medicines to name a few. From symptom checking to self-diagnosis, self-assessment and empathy in mental health to diet plans and health monitoring, home care and patient support in oncology to medication reminders and adherence in medicines, CAs have numerous applications in healthcare. These benefits of the CAs may come at a cost of safety for patients and users. This can be seen from research and observational studies where CAs, VAs in smartphones, and VUIs without proper clinical supervision may harm the users. Due to their potential as interactive, intuitive, intelligent, easy-to-access, powerful and computational natural language systems, we need to develop safety concepts for these systems so they can fulfil their prospect in assisting healthcare. The following section sheds light on safety methods which are successfully adopted by healthcare and other safety-critical systems.

2.2. Safety Assurance in Healthcare

Before we discuss safety assurance in healthcare, it is important to understand the origin and requirement of safety assurance. The need for an explicit and systematic safety case originated from serious accidents such as the Piper Alpha Off-shore disaster [85] and Clapham Rail disaster [86] in 1988. These incidents happened not because of the lack of safety procedures but the core at both accidents was misunderstanding the systematic consideration of safety. Followed by these accidents the Offshore Installations Regulations 1992 [87] and The Railways (Safety Case) Regulations 1994 [88] came into the act which changed the approach adopted to safety regulation. The introduction of safety cases posed the responsibility to the operators to demonstrate that their system has an adequate level of safety. Assurance is defined as justified trust or confidence in a system of interest. Safety assurance is, therefore, justified confidence in a system to operate safely without causing any harm. The manufacturer or the operator of the system needs to demonstrate that their system is safe to operate to the users and the regulatory bodies. This is achieved by defining safety cases which are explained in detail in the next section.

2.2.1. Safety Case

A safety case is a structured argument that exhibits evidence that a system is safe to use in a particular context [89]. The core of the safety case is a risk-based argument with evidence that demonstrates that all risks associated with the system have been identified and appropriate risk controls are put in place. A safety case provides a common platform for all the stakeholders about the safety of the system. Therefore, it must be clear, concise, and present a compelling argument with a set of evidence. In the literature, a more generalized term “Assurance Case” [90] is also used interchangeably with the term safety case. In this report, however, the term safety case is our focus and is used throughout to define the safety of the systems.

A safety case typically contains three key elements: claims, arguments and evidence. The objective of the safety case is defined by the claim, the argument explicitly shows how claims are satisfied by the evidence for the safety of the system and evidence describes the measures taken to support the claim. All these elements are important as without a proper argument, evidence may not be understood well and without solid evidence, the argument is meaningless.

Safety cases are written in text form as well as they are also represented by graphical notation. The latter communicates clear visibility in an incisive way over the former method. The earlier forms of safety cases were quite complex, and some require a large volume of text to describe the safety cases and often they were unmanageable. The Goal Structuring Notation (GSN) [45], developed at the University of York identified these issues and provided a structured and graphical representation for the better management of the safety cases. GSN is now widely used graphical representation that explicitly represents elements of a safety argument (requirements, claims, evidence, and context) and the relationships between them. Some other graphical representations of safety case development such as Claims, Arguments and Evidence (CAE) [91] are also used. This report focuses on GSN as the standard graphical representation for the development of safety cases.

There are four key elements of the GSN which are explained below:

1. Goal: The overall top-level safety argument about a system is represented by “Goal”. For example, the system is acceptably safe to operate in a given environment. Goals are further divided into sub-goals either directly or indirectly by using a strategy which is explained below.
2. Strategy: Strategies are used to split the top-level goals into more achievable sub-goals by providing a rationale.
3. Context: These are the conditions that provide constraints to goals or strategies. There is no possible way a system is safe in any available context. Context defines the bounding are of the system.
4. Solution: The safety proof or evidence of end-level or leaf goal.

The above described four main constructs of GSN are called ‘goal structures’ and they are combined with relationship elements [45]. GSN supports two types of relationship elements which are described below:

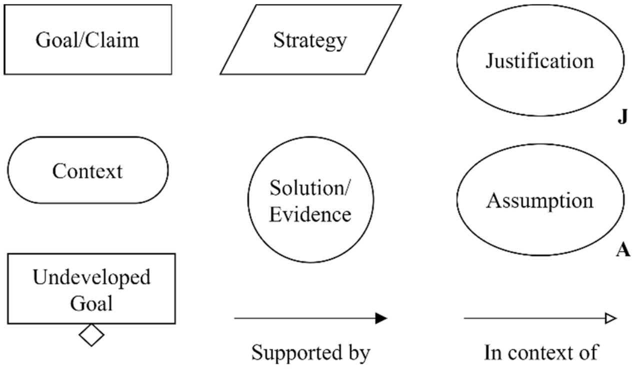


Figure 1: Graphical notation of GSN constructs and GSN relationship elements

- Supported by: This relationship is indicated by drawing an arrow with solid or filled arrowhead and is used to document inferential or evidential relationships. Inferential relationships show an inference between goals while evidential relationships state the link between a goal and the evidence used to support it.

- In context of: This relationship is specified by rendering an arrow with hollow arrowhead and it is used for contextual relationships.

Figure 1 shows the graphical notation of goal structures of GSN and the GSN relationship elements. This figure also shows the notation of an undeveloped goal. This is done by creating a hollow diamond beneath the rectangular symbol of a goal structure. An undeveloped goal represents a claim that is intentionally left undeveloped in the argument. The other two elements of GSN presented in this figure justification and assumption are not very commonly used. Sometimes claims or strategies need to be expressed in the context of some assumption. GSN justification element is used when a claim or strategy requires more explanation as to why it is considered acceptable by the author.

Figure 2 shows an example of the goal structure [89]. This example has a singular top-level goal which is 'C/S logic is fault free'. This top-level is goal is then subdivided into sub-level goals through strategies S1 and S2. The strategy S2 is made in the context of C1 that explains all identified software hazards. The C1 is linked to S2 by GSN contextual relationship element. There are five sub-level goals and the goal G4 is left undeveloped. Unlike goals G8 and G9 which have direct evidence of their claim, the goals G2 and G3 are sub-divided into G5 and G7. The argument in goals G5 and G7 are satisfied by a single proof or solution Sn2.

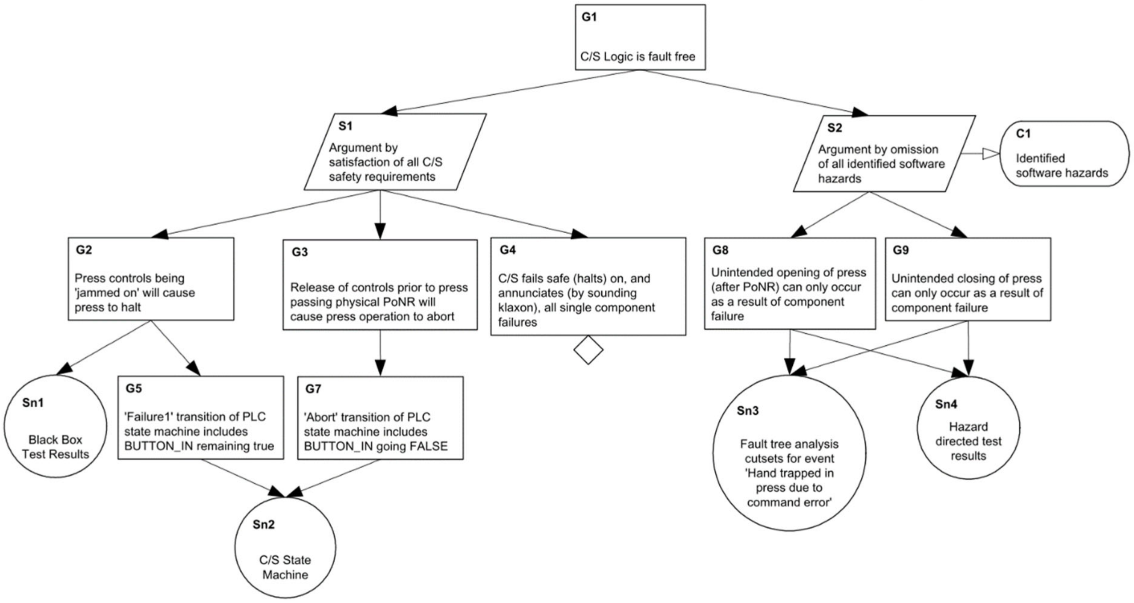

G1
C/S Logic is fault free
S1
Argument by satisfaction of all C/S safety requirements
S2
Argument by omission of all identified software hazards
C1
Identified software hazards
G2
Press controls being 'jammed on' will cause press to halt
G3
Release of controls prior to press passing physical PoNR will cause press operation to abort
G4
C/S fails safe (halts) on, and annunciates (by sounding klaxon), all single component failures
G8
Unintended opening of press (after PoNR) can only occur as a result of component failure
G9
Unintended closing of press can only occur as a result of component failure
Sn1
Black Box Test Results
G5
'Failure1' transition of PLC state machine includes BUTTON_IN remaining true
G7
'Abort' transition of PLC state machine includes BUTTON_IN going FALSE
Sn3
Fault tree analysis cutsets for event 'Hand trapped in press due to command error'
Sn4
Hazard directed test results
Sn2
C/S State Machine

Figure 2: An example GSN

2.2.2. Safety Case in Healthcare

The concept of safety case in healthcare is relatively new and its use in other industries such as oil, nuclear, and the rail is fairly old. Safety case helps regulatory bodies in maintaining a check on the developers and operators that they have adopted a systematic approach to appropriately manage the risk. There has been a development of using safety case in healthcare recently, but it is limited to the medical devices and health systems. For example, the guidance on the safe use of infusion pumps from the Food and Drug Administration (FDA) [92]. This is to assist industry in preparing premarket submissions for infusion pumps and identify device features that manufacturers need to address in the product life cycle. The recommendations provided by the FDA in this document are intended to improve the quality of infusion pumps and prevent adverse events associated with their use.

The safety case is defined by FDA as *"The safety assurance case (or safety case) consists of a structured argument, supported by a body of valid scientific evidence that provides an organized case that the infusion pump adequately addresses hazards associated with its intended use within its environment of use. The argument should be commensurate with the potential risk posed by the infusion pump, the complexity of the infusion pump, and the familiarity with the identified risks and mitigation measures."*

This definition is specific for the use of infusion pumps. In healthcare, to date, the focus on safety case is on medical devices and health IT systems. These two are defined in detail in the next sections.

Medical Devices

Safety case concept in healthcare has most advancements in the area of medical devices. The EU regulation defines a medical device as *" 'medical device' means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:*

- *diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,*
- *diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,*

- *the investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,*
- *providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations, and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means"*

and active device under which software falls as a medical device is *"'active device' means any device, the operation of which depends on a source of energy other than that generated by the human body for that purpose, or by gravity, and which acts by changing the density of or converting that energy"* [93]

For the safe use of these medical devices, both international and local standards exist. ISO 14971 [94] is an international standard that defines requirements of risk management and best practices throughout the lifecycle of medical devices. It guides on identifying hazards and hazardous situations, estimating and evaluating risks, and specifying risk control measures. Similarly, ISO 13485 [95] is an international standard related to medical devices but it is focused more on the organizations that provide medical devices and related services.

While the standards provide rules, guidance and best practices on the use in consensus by a recognized body their compliance is not mandatory. Regulation, on the other hand, is mandatory and is enforced by the government. Their role is to provide requirements, guidance and advice to manufacturers of the systems for their safe use. They can also warn or stop the manufacturers or operators to operate their system. The regulation body for medical devices and medicines in the UK is known as the Medicines and Healthcare products Regulatory Agency (MHRA). MHRA is responsible for ensuring the safety, quality, and effectiveness of these devices and medicines. Software applications that function as a medical device are required to be CE marked to ensure regulators that they are acceptably safe to use and apply in healthcare. Figure 3 shows the flowchart from MHRA helps in identifying if the software application needs to be CE marked [44].

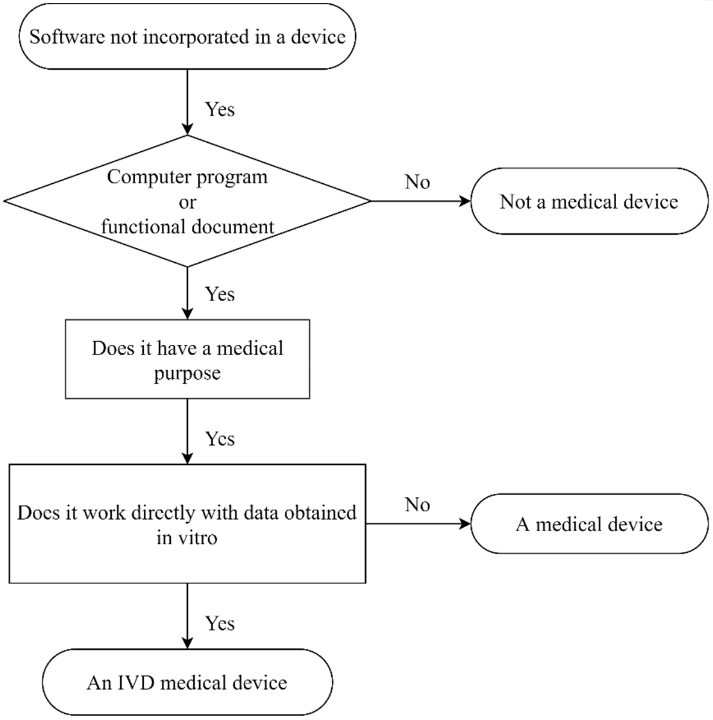

Software not incorporated in a device
Yes
Computer program or functional document
No
Not a medical device
Yes
Does it have a medical purpose
Yes
Does it work directly with data obtained in vitro
No
A medical device
Yes
An IVD medical device

Figure 3: Device determination flow chart

The flow chart in Figure 4 is the continuation of Figure 3 which lays out detail on the definition of a medical device according to MHRA [44]. The MHRA regulation is aligned with international standards for medical devices ISO 14971 and ISO 13485 mentioned earlier. From Figure 4 it is evident that software application that involves diagnosing of disease fall into the category of medical devices. From Chapter 2 on the literature review of CAs, we noticed the use of diagnosis from some applications. Thus, those applications come under the regulation of MHRA. The CAs which fall under the medical device category may become a source of harm to users.

The MHRA document [44] further adds information about symptom checkers which use AI and CA to interact with users. If the software provides a subset of medical conditions that matches user input or indicates the likelihood of a match or provides recommendations for entered conditions, then it is considered as a medical device. On the other hand, the software will not be considered a medical device if it provides only reference information or direct user to suitable care such as GP. Symptom checkers are considered low-risk devices unless they provide a direct diagnosis in which case they are regarded as medium risk devices.

CAs are sophisticated in the use of diagnosis of diseases and a safety case will greatly help both the developers and the regulators for their safe use. Earlier we mentioned FDA guidance [92] on infusion pumps, in this document they also propose GSN as one of many options for developing a safety case for infusion pumps. For a better understanding of safety case for a medical device, below we discuss an example of an infusion pump.

There are two approaches of creating a safety case: (1) satisfying all identified safety requirements, or (2) focusing on safety hazards and showing they have been mitigated to an adequate level. To show that a safety requirement is met, often it is done by exhibiting that hazards have been mitigated or eliminated.

These approaches are not mutually exclusive and can be seen from the safety case of an infusion pump [96]. In this safety case, it has been mentioned as a Generic Infusion Pump (GIP). An infusion pump delivers medicine into a patient's circulatory system with continuous or periodic controlled delivery of medicine. Infusion pumps can be relatively single or quite complex. They all require input programming to control the rate and duration of infusion of medicine. The right amount of medicine and its dosage is critical for the patient and the manufacturers have the responsibility to make sure its safe usage to the regulators.

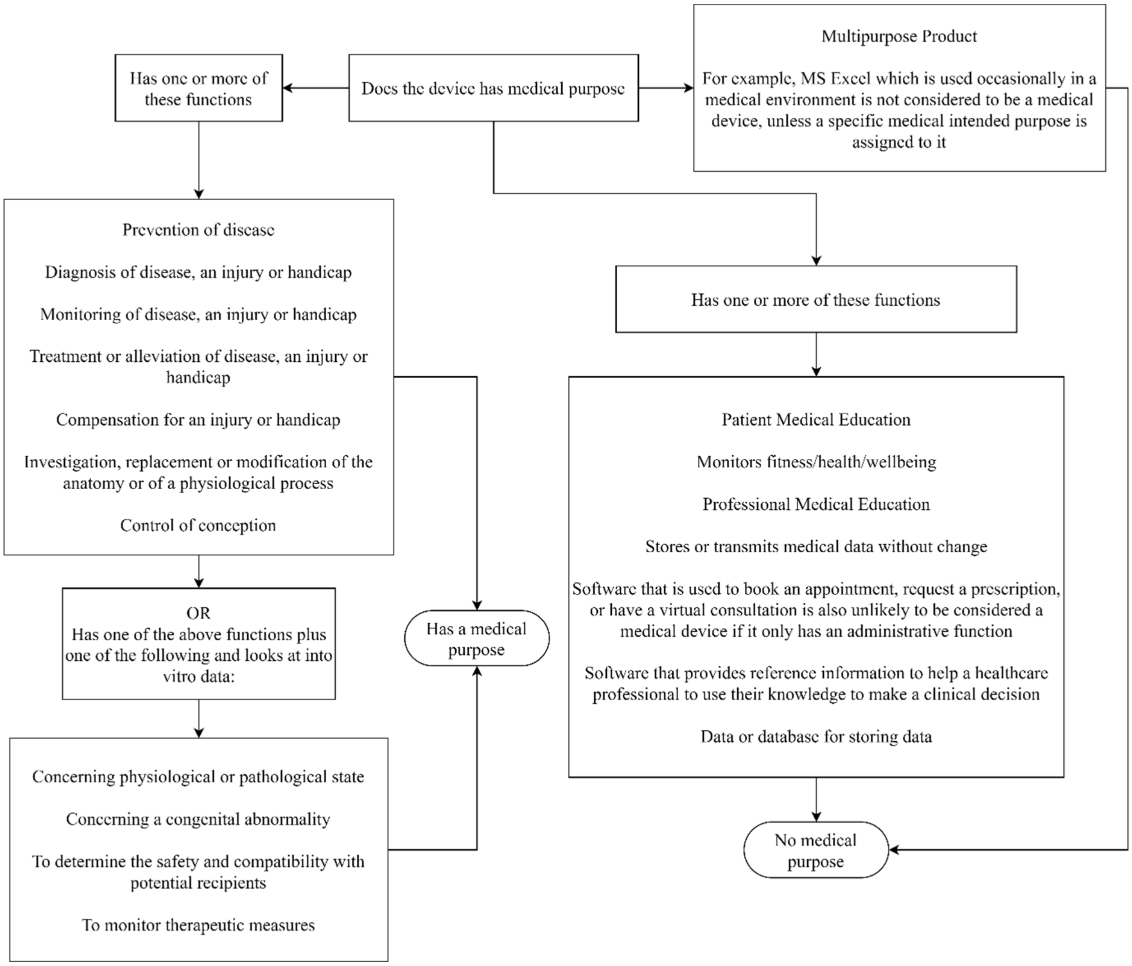

Has one or more of these functions
Does the device has medical purpose
Multipurpose Product
For example, MS Excel which is used occasionally in a medical environment is not considered to be a medical device, unless a specific medical intended purpose is assigned to it
Prevention of disease
Diagnosis of disease, an injury or handicap
Monitoring of disease, an injury or handicap
Treatment or alleviation of disease, an injury or handicap
Compensation for an injury or handicap
Investigation, replacement or modification of the anatomy or of a physiological process
Control of conception
Has one or more of these functions
Patient Medical Education
Monitors fitness/health/wellbeing
Professional Medical Education
Stores or transmits medical data without change
Software that is used to book an appointment, request a prescription, or have a virtual consultation is also unlikely to be considered a medical device if it only has an administrative function
Software that provides reference information to help a healthcare professional to use their knowledge to make a clinical decision
Data or database for storing data
OR
Has one of the above functions plus one of the following and looks at into vitro data:
Has a medical purpose
Concerning physiological or pathological state
Concerning a congenital abnormality
To determine the safety and compatibility with potential recipients
To monitor therapeutic measures
No medical purpose

Figure 4: Medical device determination flow chart

Figure 5 [96] illustrates the top-level claim of the safety case of GIP. The top-level claim or goal "The GIP is safe for use on patients" is subdivided into claims C2 and C30 based on two possible hazards to the patient by its usage. The claim C30 needs further development as shown by the undeveloped goal symbol. The focus of this example is the argument C2 that patient hazards due to unsafe programming are mitigated. Some hazards cannot be mitigated with the safe use of GIP programming and are captured as GSN element assumption A2. The GSN context element Ct6 lists the classes of hazards to patients. The argument C2 further divides into C3 that the GIP is accurately programmed and C22 that the parameters of the drug are safe for the patient. Figure 6 [96] represents the claim C3 and its evidence. The claim C3 is subdivided by 3 claims. The claim C19 refers to the tolerable rate of parameter entry by the person. The evidence to this claim comes from the error log and GIP procedure manual. These leaf of GSN satisfies claim of tolerable entry errors which with other sub-claims satisfies the claim of accurately programmed GIP which in turn with its sub-claims satisfies the main claim of having a safe GIP for use on patients.

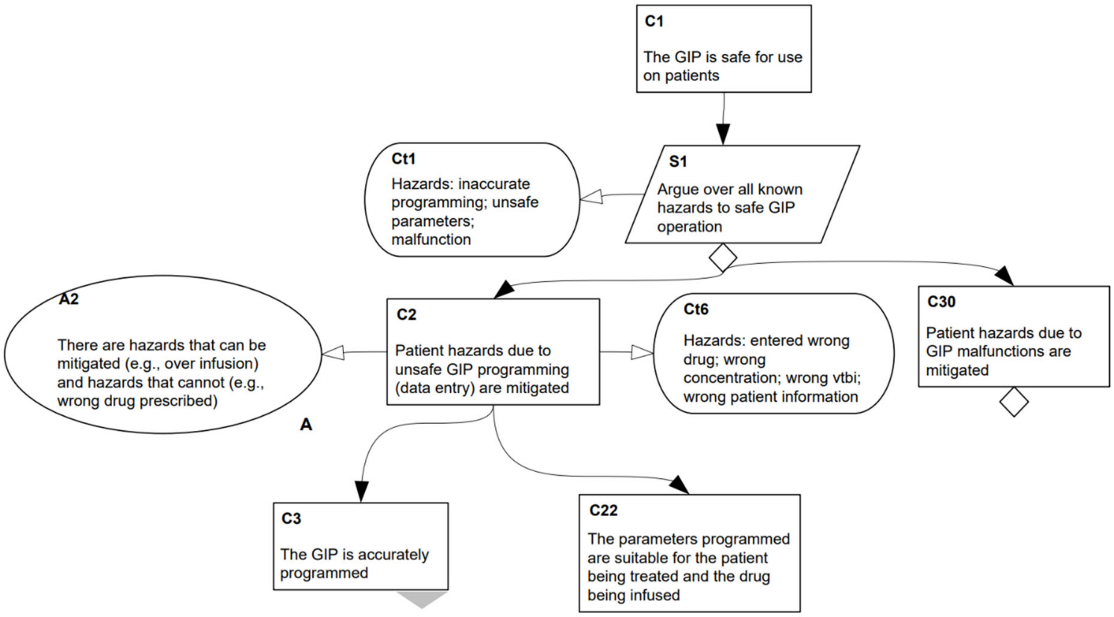

C1
The GIP is safe for use on patients
Ct1
Hazards: inaccurate programming; unsafe parameters; malfunction
S1
Argue over all known hazards to safe GIP operation
A2
There are hazards that can be mitigated (e.g., over infusion) and hazards that cannot (e.g., wrong drug prescribed)
A
C2
Patient hazards due to unsafe GIP programming (data entry) are mitigated
Ct6
Hazards: entered wrong drug; wrong concentration; wrong vtbi; wrong patient information
C30
Patient hazards due to GIP malfunctions are mitigated
C3
The GIP is accurately programmed
C22
The parameters programmed are suitable for the patient being treated and the drug being infused

Figure 5: The GIP assurance case- the GIP is Safe

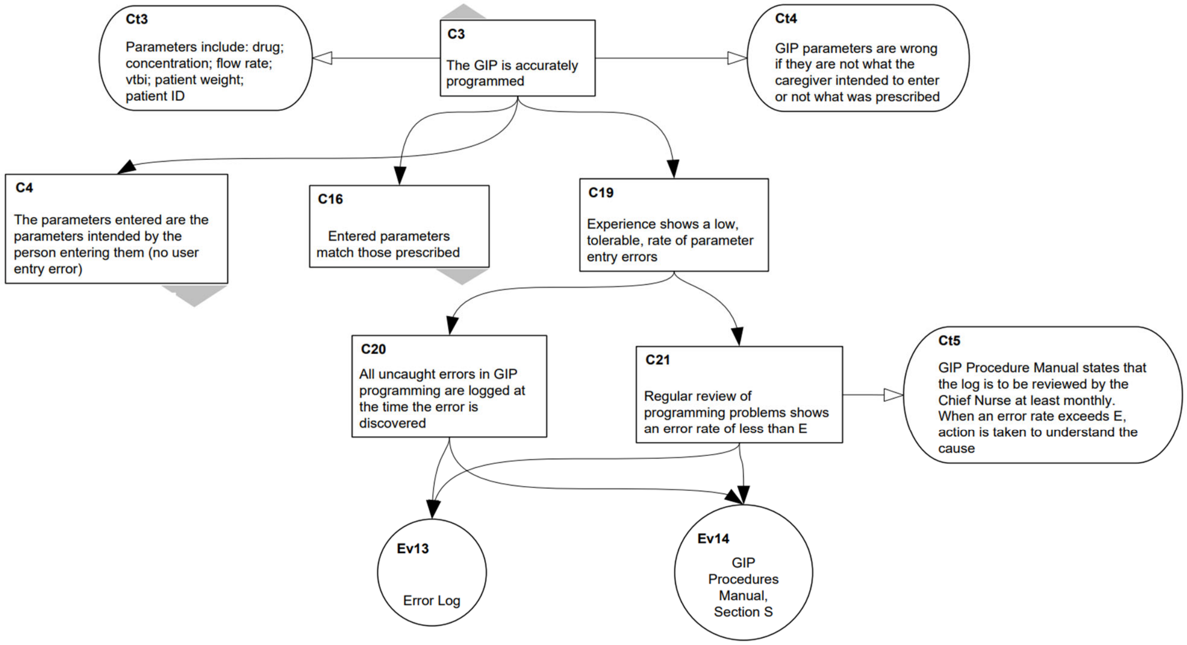

Ct3
Parameters include: drug; concentration; flow rate; vtbi; patient weight; patient ID
C3
The GIP is accurately programmed
Ct4
GIP parameters are wrong if they are not what the caregiver intended to enter or not what was prescribed
C4
The parameters entered are the parameters intended by the person entering them (no user entry error)
C16
Entered parameters match those prescribed
C19
Experience shows a low, tolerable, rate of parameter entry errors
C20
All uncaught errors in GIP programming are logged at the time the error is discovered
C21
Regular review of programming problems shows an error rate of less than E
Ct5
GIP Procedure Manual states that the log is to be reviewed by the Chief Nurse at least monthly. When an error rate exceeds E, action is taken to understand the cause
Ev13
Error Log
Ev14
GIP Procedures Manual, Section S

Figure 6: The GIP assurance case - accurately programmed

Health IT Systems

Health IT systems are defined as *"Product used to provide electronic information for health or social care purposes. The product may be hardware, software or a combination"*. Normally, the health IT system contains software running on computers or mobile devices and medical devices are stand-alone has embedded software. As described in the last section, according to the MHRA device determination flowchart, the software has the potential to be a medical device if it goes beyond dissemination and communication of information. Digital health technology has the potential to mitigate risks and can as well be a source of introducing new clinical risks. In the UK, therefore, the NHS digital has developed national clinical standards DCB0129 [97] and DCB0160 [98] respectively. These two standards are overlapping as the former provides requirements for clinical risk management for manufacturers while the latter sets requirements for the use, deployment, and maintenance of health IT systems. Manufacturing organizations of health IT systems and applications need to do a formal risk assessment and provide evidence of the measures taken to mitigate clinical risks. They are required to produce a clinical risk management plan, hazard log, and clinical safety case report to comply with these standards.

2.2.3. Safety Analysis Methods in Healthcare

Healthcare is often encouraged to use safety techniques from other safety-critical industries. Apart from safety case some hazard and reliability analysis methods are used in healthcare from other industries. In the context of healthcare, a hazard is defined as *"a potential source of harm to a patient"* while the clinical risk is defined as *"combination of the severity of harm to a patient and the likelihood of occurrence of that harm"* by the NHS [98]. To prove the safety of a system is to ensure all potential hazards and risks are mitigated to an acceptable level before the operation of the system. In the UK, this is often stated as low as reasonably practicable (ALARP) principle. In the context of risk management, making sure a risk has been reduced ALARP means weighing the risk against the measures necessary to further reduce it. There are various techniques for assessing and identifying hazards in safety-critical industries that are used in healthcare.

2.2.3.1. Checklists

Checklists are a basic technique of making certain that tasks get done. For example, a shopping list to go to the supermarket. Checklists are based on the argument that there are certain cognitive limitations of the human brain. They are a simple solution to potentially complex problems in many industries. They are used extensively in aviation and are an integral part of their workflow. For example, in aviation, there are three forms of checklists, one for routine operations; one for complex operations; and one for emergencies [66]. Healthcare application of checklists included in the field of surgery and [99] infection control [100].

However, in comparison to checklists in healthcare vs other industries, there are important qualitative differences. As a comparison made by Catchpole et al. showed vast differences between checklists in aviation and healthcare in terms of the size, the number of words used for each checklist, and the type of checklist (which require a confirmation or not) used [101]. This is due to the complex nature of operations in healthcare as Williams et al. argued their effective usage in complex clinical settings [102]. A checklist is a socio-technical intervention and it requires a lot of other ingredients to design and implement in healthcare. For example, to successfully use in healthcare, it requires focused effort, communication and teamwork in addition to simple checks. The communication and teamwork also require training people using it for these skills.

Checklists are a great tool if used with the right design and implementation and in the right place having people with the right skills. Their translation from other industries cannot be applied to healthcare but rather it requires precise design, the right placement, and proper training to the people using it. They are not as simple to apply as they seem and their application in healthcare needs to learn complex factors and changes to culture, design, and teamwork that accompanied them in other industries [69][101].

2.2.3.2. Failure Mode and Effects Analysis (FMEA)

Failure Mode and Effects Analysis (FMEA) is a systematic approach to risk mitigation. FMEA works to identify failure modes that can cause a system to fail and the effects of those failures modes to the overall system. FMEA analysis is used to recognize, prioritize and limit these failure modes. After identifying the failure modes, they are assigned a single Risk Priority Number (RPN) which is calculated by multiplying the probability, severity, and the detectability of failure mode occurring. FMEA technique has different variants such as Healthcare Failure Mode and Effect Analysis (HFMEA), Failure Mode, Effects and Criticality Analysis (FMECA). FMEA is the most popular PHA technique used in healthcare [103]. HFMEA was developed by the Department of Veterans Affairs (VA) National Centre for Patient Safety and it is a five-step process which includes:

(1) Defining the topic
(2) Assembling a multidisciplinary team
(3) Graphically describing the process
(4) Conducting hazard analysis
(5) Determining actions and outcome measures [104].

FMEA is extremely thorough and quite systematic and costlier to implement (in terms of staffing hours required). It identifies examination of all possible failures in a system which also require a lot of time investment which is a serious limitation. FMEA is also resource-intensive and there is a lack of guidance for hazard identification and risk assessment [105]. Additionally, the calculation of RPN is not very relevant as Franklin et al. pointed out the mathematical and logical flaw in calculating RPN and take measure based on the RPN number [106]. There is also a difference in FMEA and HFMEA as the former uses a 10-point scale while the latter uses 4 point scale and HFMEA detectability scores are determined only if there is required a further action for the identified failure. Although it is a powerful technique, the quantitative use of FMEA is not backed up by proper evidence as small variation in one of RPN parameters can have a variable effect on its quantification.

2.2.3.3. Hierarchical Task Analysis (HTA)

Hierarchical Task Analysis (HTA) is a task description method used to decompose a task hierarchically into smaller units. The structure consists of goals, sub-goals, operations, and plans. HTA can be used as a basis for further analysis such as error analysis. In healthcare, HTA has been used in various forms. One study showed the use of in surgery as it is the procedure which cannot be done solely by a surgeon [107]. The surgeon may need surgical assistant, anaesthetist, endoscopy nurse etc. The individual task analyses of other people combined with the surgeon's analysis make up the complete HTA. Lane at al. demonstrated the use of HTA to reduce errors in medication administration in hospital [108]. They used Systematic Human Error Reduction and Prediction Approach (SHERPA) which was complimented by HTA.

Summary

This section covered safety assurance, the origin, and the need for them in safety-critical industries. Safety case, a safety assurance method was discussed with its representations. GSN which is a graphical representation of the safety case was talked about with its notation and an example use case. Although the use of safety cases is more common in other industries, it is relatively new in healthcare. In healthcare, medical devices and health IT systems are the focus for the use of safety case. There are international and local standards for the safety of medical devices and health IT systems. These standards, however, lack in terms of providing guidelines for the design and development of AI-based systems. The regulatory body for medical devices in the UK, MHRA acknowledges the use of CAs and considers the software to be low-to-medium risk medical devices. Healthcare has used various safety analysis methods from other industries for hazard identification such as Checklists, FMEA, and HTA to name a few. There are problems in applying those methods directly to healthcare and safety culture and teamwork is needed for their proper application.

The lack of safety standards for guidelines of AI-based systems is a challenging situation for the developers of these software applications and CAs. Further, the regulation on CAs is not very strict either and up till now, only FDA provides guidelines for the need of safety assurance case for use of infusion pumps in healthcare. This is the closest example of requiring a safety case for a medical device in healthcare. We believe, based the risk associated with the use of CAs (from literature and MHRA) and given the safety case requirement by FDA for safe use of infusion pumps, a safety case for CAs is needed for healthcare.

2.3. Conversational Agents

Conversation or dialogue is the most fundamental part of human language. It is the kind that we grow up learning and in our daily lives, we use this way of language to communicate with others. From ordering our lunch at a restaurant to talk with our friends, booking air tickets or participating in meetings, we indulge in conversations to get our tasks done or express our feelings. Dialogue system or CAs are types of programs that mimic human language. These programs use text, speech or both to communicate with users in natural language. Conversational systems fall into two broader classes namely: Task-oriented dialogue agents, and chatbots [6]. The former helps the user to complete their tasks while the latter is designed for extended conversations.

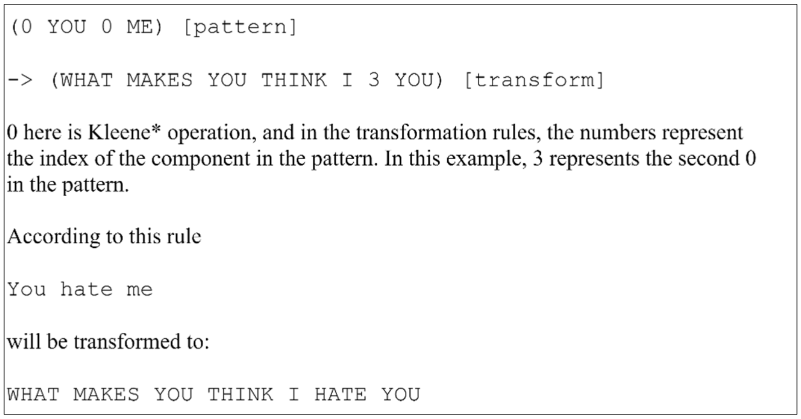
Modern VUIs are classified as task-oriented dialogue systems. Their application includes but not limited to making reservations at restaurants, dialling phone calls, giving user's directions, opening applications on smartphones, playing games, and listening to music or jokes. In contrast, chatbots impersonate natural conversation and human-human communication and are used for extended conversations.

2.3.1. Chatbots

Chatbots are the simplest example of dialogue systems that carry on an extended conversation with the objective of 'chats' or unstructured conversations. The architecture of chatbots can be classified into two main categories: rule-based systems and corpus-based systems [6].

2.3.1.1. Rule-based Chatbot

ELIZA chatbot is the most important in the history of dialogue systems. Designed in 1966, it was created to simulate a Rogerian psychotherapist [109]. In this psychology, a conversation starts without knowing anything about the real world. ELIZA was based on rule and pattern matching algorithm where the chatbot inferred information from the user and a response is generated on that. Each rule in the algorithm is linked to a possible keyword that a user might use in their sentence. For example, Figure 7 below is one pattern/rule of ELIZA [6].



```
(0 YOU 0 ME) [pattern]

-> (WHAT MAKES YOU THINK I 3 YOU) [transform]
```

0 here is Kleene* operation, and in the transformation rules, the numbers represent the index of the component in the pattern. In this example, 3 represents the second 0 in the pattern.

According to this rule

```
You hate me
```

will be transformed to:

```
WHAT MAKES YOU THINK I HATE YOU
```

Figure 7: Pattern example of ELIZA chatbot

ELIZA's pattern/action architecture is still used by modern chatbots systems such as ALICE. PARRY is another chatbot which focuses on clinical psychology and was developed in 1971 [110]. PARRY included a mental model which was affected by its state of anger or fear based on the interaction. A high anger value turns PARRY to choose a 'hostile' response while a high 'fear' value for example based on an input which mentions its delusion, would make PARRY express statements related to its delusion. PARRY was the first chatbot to pass Turing test which is a measure of intelligence of a dialogue system [111].

2.3.1.2. Corpus-based Chatbots

Corpus-based chatbots, unlike rule-based chatbots, build conversations using mining lots of human-human or human part of human-machine conversations. According to an estimate [112], modern chatbots needs hundreds of millions or even billions of words to train their data. The source of data may come from the movies database [24], text from social media platforms [113], etc. A trained chatbot uses human conversations during their interaction with the chatbot to enhance its learning. The increased training data set helps chatbot respond more naturally. There are also other ways in which corpora can be built for training a chatbot. Some topic-specific corpora are used to train topical chatbots.

There are two main architectures for corpus-based chatbots: information retrieval and machine-learned sequence transduction [6]. Context modelling is not very common in these chatbots but rather they tend to respond based on the user's current utterances. Corpus-based chatbots are like question answering systems where a response is generated usually by ignoring the full context of the conversation.

Information Retrieval IR-based Chatbots

The main idea behind IR based chatbots is to respond to a user by selecting an appropriate response from a corpus of natural language text. The problem with the rule-based approach is that the developer must specify every possible 'pattern/rule' to be able to provide a complete response. IR based chatbots retrieve information from the corpus that has stored conversations in pairs of the form of turns of a conversation. Having corpora and a user's sentence, any retrieval algorithm can be used to generate a response. There are two methods for returning a response.

1. **Return the response to most similar turn:** The main idea is to choose a turn which is most like the user's turn and return the stored human response to that turn. For a query q and a corpus C, a turn t in C that is most like q and return that turn i.e. human response to t in C:

$$r = response\ (argmax_{t \in C} \frac{q^T t}{\| q \| t \|})$$

2. **Return the most similar turn:** Here a user's query q is matched to the turn's t in corpus C and is returned because it will share words or semantics with the turn. For a user's query q, return the turn t in C which is most like q.

$$r = argmax_{t\in C}\frac{q^T t}{\parallel q \parallel t \parallel}$$

The response to most similar turn approach works better in practice [114]. The reason is that user's query q if most similar to the turn t then it means it is more effective since the co-occurrence of words in them. If it is matched against the turn itself, it is not clear if the response is more accurate because of the similarity between the user's query q and the turn t itself.

To compute the similarity any similarity function can be used. For example, cosine similarity of words or over any sentence embedding can be used. Occasionally keyword matching is used where keywords are looked for matching in user's query q and the document in corpus C. In complex IR models, more features can be added than just the words in user's query q. When there are few words in the user's query, all conversation can be used to add more meaning in matching. Also, the sentiment of a user can be helpful.

2.3.2. Task-Oriented Dialogue Agents

The goal of these type of dialogue agents is to help the user in achieving a task such as booking an aeroplane ticket, schedule an appointment, etc. This section will introduce these task-oriented conversational systems.

2.3.2.1. GUS - A Frame-Based Architecture

Most of the modern VAs are based on GUS system architecture which was first introduced in 1977 [115]. These dialogue state architectures are based on frames. A single frame represents a knowledge structure representing various features it can extract from the user's utterance and consists of various slots. These slots can take a set of possible values. For example, in a healthcare domain, a slight might represent care unit (take on the value of "primary care" or "emergency care"), location (which can take the value of a city), or date and time. The types in GUS systems as well as in modern frame-based dialogue systems have a hierarchical structure. Date type, for example, is a frame with slots of integers or values of sets of weekday names [6] as shown in Figure 8.

```
DATE
    MONTH:NAME YEAR:INTEGER DAY:(BOUNDED-INTEGER 1 31)
    WEEKDAY:(MEMBER (Sunday Monday Tuesday Wednesday
                     Thursday Friday Saturday))
```

Figure 8: Date type in GUS architecture

Control Structure

The control structure of these dialogue agents is designed around frames and is used in modern VAs. The system interprets the user's intention and fill in the slot values in the frame and perform relevant action. To achieve this, the system keeps asking questions from the user (each slot of each frame has pre-defined question templates) and fills slots that the user specifies.

The GUS architecture also provides slots with condition-action rules attached to them. For example, a rule attached to the HOSPITAL_NAME slot for the search emergency care frame might automatically assign as the default Medical Centre for the related general practitioner booking frame. Some domains also require multiple frames. For example, there are general information frames for questions like *which hospitals or care unit are between my home and my workplace*, or for information like *how much typically I have to wait for an appointment at a specific hospital?* The GUS architecture is a production rule system because of its need to dynamically switch controls. Different inputs fire different productions each of which can fill in different frames.

Natural Language Understanding for slots filling

The goal of this component is to extract three main features from the conversation which are domain classification, intent determination, and slot filling. Domain classification helps to identify a broader domain of the user's query such as airline, programming alarm etc. After domain selection, the need to understand the user's general task or goal is fulfilled by determining the intent. Intent could be book or view flights, view or remove alarms etc. Finally, slot filling is required to extract the particular slots and fillers that are needed by the system to understand from the user's utterance.

For example, Figure 9 shows the intent generation for a user's request.

```
Wake me up tomorrow at 10 O'clock
```

produces below intent:

```
DOMAIN: ALARM-CLOCK
INTENT: SET-ALARM
TIME: 2020-03-30 10:00:00
```

Figure 9: Sample intent generation from the user's utterance

The original GUS system uses manually designed rules for slot filling. For the above example, a regular expression can also be used for recognizing the intent. Many modern dialogue systems which use GUS architecture at their core use supervised machine learning for slot filling [6].

2.3.2.2. The Dialogue State Architecture

The modern VAs use a more sophisticated version of earlier frame-based (GUS) architectures. These are called dialogue-state or belief state architecture. Figure 10 shows the architecture of task-oriented dialogue systems [116].

Automatic Speech Recognition (ASR)

The speech recognition process takes sound as an input and converts it to a string of words as output. It is the core component of a VUI that distinguish it from a text-based CA. The challenges in ASR include handling tasks with large vocabularies of over 64,000 words and processing continuous speech in which words must be segmented because of overlapping within them. This is also known as Large Vocabulary Continuous Speech Recognition (LVCSR) [46]. There are three main steps in the ASR process model: (1) pre-processing, (2) speech segmentation, and (3) feature extraction [117].

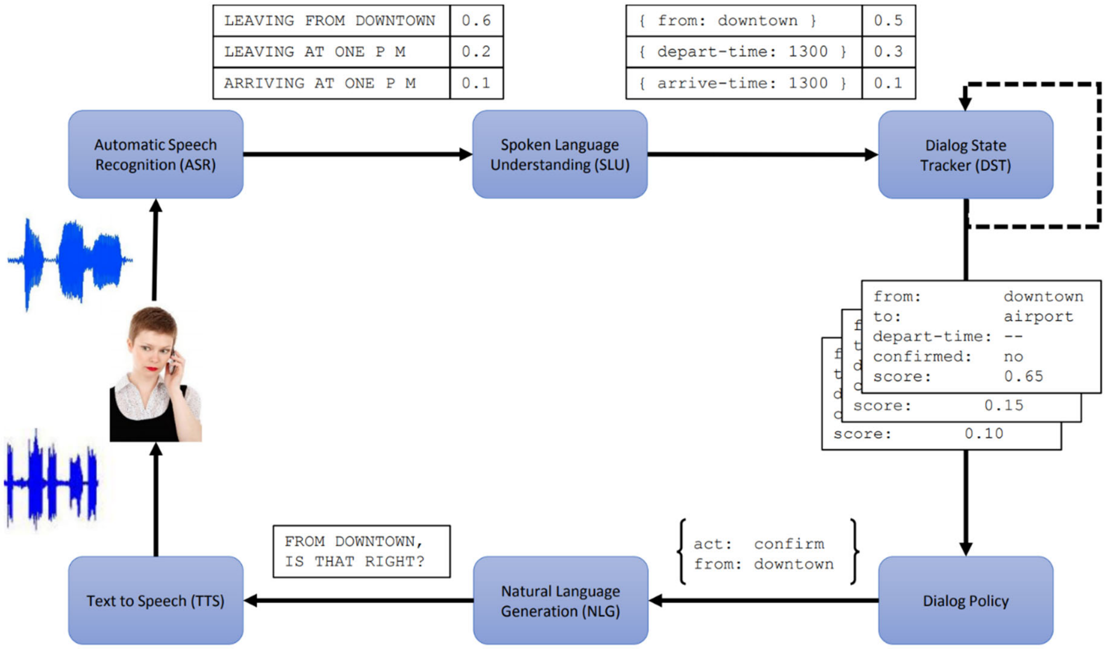

LEAVING FROM DOWNTOWN 0.6
LEAVING AT ONE P M 0.2
ARRIVING AT ONE P M 0.1
{ from: downtown } 0.5
{ depart-time: 1300 } 0.3
{ arrive-time: 1300 } 0.1
Automatic Speech Recognition (ASR)
Spoken Language Understanding (SLU)
Dialog State Tracker (DST)
from: downtown
to: airport
depart-time: --
confirmed: no
score: 0.65
score: 0.15
score: 0.10
act: confirm
from: downtown
Dialog Policy
Natural Language Generation (NLG)
FROM DOWNTOWN, IS THAT RIGHT?
Text to Speech (TTS)

Figure 10: Dialogue state architecture for task-oriented dialogues

Dialogue Acts

Dialogue acts represent the internal function of a turn or sentence. The task of these acts is to extract key information from the user's utterance in a way that could help the system understand and complete the task. Different types of tags are defined first based on the type of system and these tag sets are consumed by the dialogue acts. Table 3 shows the dialogue acts used by HIS restaurant recommendation system [48]. The system and user columns indicate the acts that are valid from the perspective of user input and system output. For example, from this figure, the confirmation acts 'CONFIRM' and 'CONFIRMREQ' are not valid as user input. Table 4 shows a sample dialogue from the HIS restaurant recommendation system [118] that is using dialogue acts.

Table 3: Dialogue acts used by the HIS restaurant recommendation system

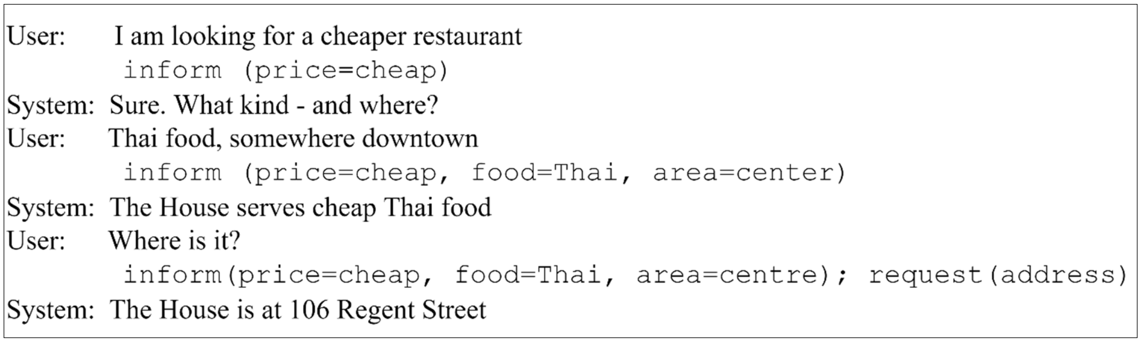
Act	System	User	Description
HELLO(a = x, b = y, …)	✓	✓	Open a dialogue and give info a = x, b = y, …
INFORM (a = x, b = y, …)	✓	✓	Give info a = x, b = y, …
INFORM (a = x, b = y, …)	✓	✓	Request value for a given b = x, …
INFORM (a = x, b = y, …)	✗	✓	Request alternative with a = x, …
INFORM (a = x, b = y, …)	✓	✓	Explicitly confirm a = x, b = y, …
CONFREQ (a = x, …,d)	✓	✗	Implicitly confirm a = x, … and request value of d
SELECT (a = x, a = y)	✓	✗	Implicitly confirm a = x, … and request value of d
AFFIRM (a = x, b = y, …)	✓	✓	Affirm and give further info a = x, b = y, …
NEGATE (a = x)	✗	✓	Negate and give corrected value a = x
DENY (a = x)	✗	✓	Deny that a = x
BYE ()	✓	✓	Close a dialogue

Table 4: Sample dialogue example from the HIS system using dialogue acts

Utterance	Dialogue Act
U: Hi, I am looking for somewhere to eat.	hello(task = find, type=restaurant)
S: You are looking for a restaurant. What type of food do you like?	confreq (type = restaurant, food)
U: I'd like an Italian somewhere near the museum.	inform(food = Italian, near=museum)
S: Roma is a nice Italian restaurant near the museum.	inform(name = "Roma", type = restaurant, food = Italian, near = museum)
U: Is it reasonably priced?	confirm(price range = moderate)
S: Yes, Roma is in the moderate price range.	affirm(name = "Roma", price range = moderate)
U: What is the phone number?	request(phone)
S: The number of Roma is 385456.	inform(name = "Roma", phone = "385456")
U: Ok, thank you goodbye.	bye()

Dialogue State Tracking

This component's role is to maintain the state of the current frame (including all slots that are filled from the user's utterance) as well as the user's most recent dialogue. The dialogue state stores the entire state of the frame to the current point of the user's conversation. Figure 11 shows the state of dialogue tracker after every turn in a user's conversation about finding a restaurant [119].



```
User:    I am looking for a cheaper restaurant
          inform (price=cheap)
System:  Sure. What kind - and where?
User:    Thai food, somewhere downtown
          inform (price=cheap, food=Thai, area=center)
System:  The House serves cheap Thai food
User:    Where is it?
          inform(price=cheap, food=Thai, area=centre); request(address)
System:  The House is at 106 Regent Street
```

Figure 11: An example of dialogue state tracker after each turn

Dialogue Policy

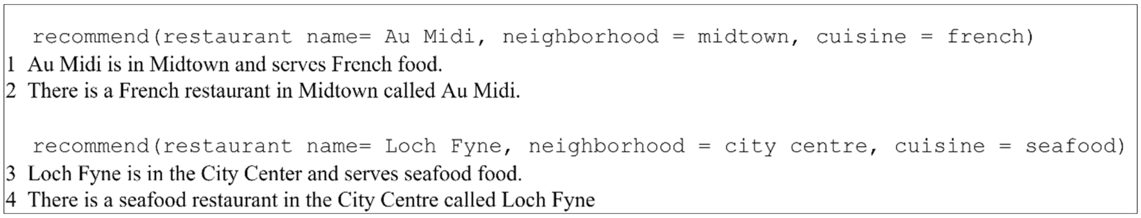
The goal of the dialogue policy is to choose the upcoming dialogue to generate during a conversation. It is the calculation of maximizing the probability of an action A to take at turn i based on all the history of dialogue state. The history contains all the acts from the system (A) and the user (U).

$$A^{\wedge}{}_{i} = argmax_{A_i \in A}\, P\ (A_i \,|\, (A_1, U_1, \ldots, A_{i-1}, U_{i-1}\ ))$$

Natural Language Generation

After the dialogue policy decides the action to generate, natural language generation component generates the response text. The task of NLG in this architecture consists of two stages: Content

Planning (what content to generate) and Sentence Realization (how to generate it). Figure 12 below shows two examples of sentence realization produced by the NLG component.

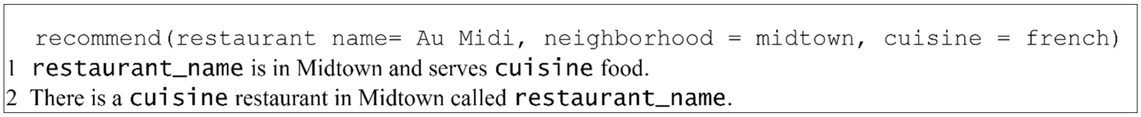


```
  recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french)
1  Au Midi is in Midtown and serves French food.
2  There is a French restaurant in Midtown called Au Midi.

  recommend(restaurant name= Loch Fyne, neighborhood = city centre, cuisine = seafood)
3  Loch Fyne is in the City Center and serves seafood food.
4  There is a seafood restaurant in the City Centre called Loch Fyne
```

Figure 12: Example of sentence realization in NLG

From the above figure, in the first example, the sentence realization has generated two sentences for the dialogue act RECOMMEND based on the slots (restaurant name, neighbourhood, cuisine) and their fillers. These sentences are generated from a large corpus of labelled dialogues in the training. Since it is difficult to find all such instances, for example, recommending a restaurant from training data a technique delexicalization is used. It is the process of replacing specific words with generic words in the training examples that represent slot words. Figure 13 shows the previous example with delexicalization.



```
  recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french)
1  restaurant_name is in Midtown and serves cuisine food.
2  There is a cuisine restaurant in Midtown called restaurant_name.
```

Figure 13: Delexicalized example of sentence realization in NLG

Evaluation Techniques

There are various ways in which the dialogue systems and chatbots are evaluated. The evaluation criteria also depend on the nature of task dialogue systems are carrying out. For task-based dialogue systems, task completion success is usually the criteria of their evaluation. In frame-based task-oriented dialogue systems, this can be a slot error rate for a sentence. The slot error rate is the ratio of added, updated or removed slots to the total number of slots for the sentence.

$$Slot\ error\ rate\ for\ a\ sentence = \frac{number\ of\ added, updated, removed}{number\ of\ total\ slots\ for\ the\ sentence}$$

There are other evaluation metrics apart from slot error rate that can be used to measure the efficiency of a dialogue system. These include slot precision, recall, and F-score [6]. In ML models a confusion matrix is essentially an error matrix that visualizes the performance of an algorithm which contains actual class labels and the predicted class labels. Table 5 shows the confusion matrix below:

Table 5: Machine learning confusion matrix

		Actual Values	
Predicted Values		Positive	Negative
	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

It consists of True Positives (TP) which indicates that a class is correctly predicted with a positive value, true negatives which is the correct prediction of negative values, false positives which are falsely identified as a positive class instead of negative, and false negatives that are predictions falsely identified as a negative class instead of positives.

Precision is the ratio of correctly predicted positive observations of the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the sensitivity and it is the ratio of correctly predicted positive observations to all the observations in one positive class.

$$Recall = \frac{TP}{TP + FN}$$

F1 score is the weighted average of precision and recall and it considers both the false positives and false negatives. In some cases, the F1 score is preferred over accuracy. Accuracy works best when the class distribution is even, or the cost of false positives and false negatives is similar. In an uneven distribution of classes, it is better to check for both precision and recall. F1 score is calculated by the formula below:

$$F1\ Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

2.3.3. Failure Modes of VUIs

VUIs may fail in various ways and may cause harm to the user. Possible failures may arrive from the end-to-end architecture of VUIs. Failures may also arrive from the interaction of humans and the VUIs. These failure modes are discussed in the next sections.

2.3.3.1. AI Failure Modes

Distributional Shift

The decision making in the ML system is highly dependent on the data they are trained on. If the patient or user is not represented by that data on which the system was trained, it can compromise the performance or accuracy of the system. This phenomenon is known as distributional shift [120]. The distributional shift can occur because of inadequate training data or inappropriate application of ML system application to out of context data. In healthcare, limitation of high quality labelled data for training [121], changes in a disease pattern, change in a patient's demographics, or difference in the result of machines to scan, measure, or record a patient's condition [121] can introduce distributional shift.

Insensitivity to Impact

The tendency of ML systems to ignore the weight or cost of the prediction results in insensitivity to impact [121]. To achieve higher accuracy ML system may ignore the real-world impact of the missed diagnoses. In one such healthcare example, the diagnostic performance of ML systems was better than clinicians on test cases of benign and malignant skin disorders, but at the same time, they missed more malignant diseases (false positives and false negatives) than clinicians did [122]. For this reason, accuracy alone may not be a good evaluation metric for ML based systems and hence VUIs.

Black Box Decision Making

Biasness and error in the decision making of ML systems can be introduced from the training data. This is very hard to detect as these systems operate like black boxes and difficult to interpret [121]. For example, an X-Ray machine which works correctly for the most part but may provide wrong results for specific cases that it may come across rarely. Normally, it is assumed from clinicians to interpret the ML system's recommendation and take control when it produces an erroneous result [123]. In artificial neural networks relating predictions to their input is more challenging [121]. Modern VUI applications use neural approaches in their architectures and therefore they must be monitored by a clinician.

Automation Complacency

Automation complacency is a related concept to automation bias and confirmation bias [124] which are closer to human cognitive biases and discussed in detail in the human factors section of this report. It occurs when more weightage is given to the system's decision as they are generally reliable. The clinician may not be able to detect incorrect prediction from the system also because of workload and concurrent tasks [121]. CAs can aid clinicians as well as patients in various tasks. For clinicians, they can retrieve quick information to help in their decision making. While this can reduce the workload on clinicians, more trust in the reliable automated systems might make clinicians cease other alternative solutions.

Reinforcement of Outmoded Practice

ML systems trained on historical data may find it difficult to adapt to new developments in the area of research. There can also be sudden changes in the policy due to any unforeseen reason. For example, due to safety concerns, a drug may be advised not to use and withdrawn from the market. The ML system in its decision making will reinforce the same drug based on historical data [121]. The system may need to retrain based on new information which can be very costly.

Negative Side Effects

This phenomenon occurs when the ML system learn to perform a specific task in the environment but ignores the wider context creating an unsafe consequence [42]. For example, an autonomous ventilator learns to ventilate short term oxygenation at the expense of long-term damage to the lungs of a patient [125]. This may be very similar to a VUI application that suggests medication to a user or patient to cure one condition and might not aware of the internal health condition of the user.

2.3.3.2. VUI Failure Modes

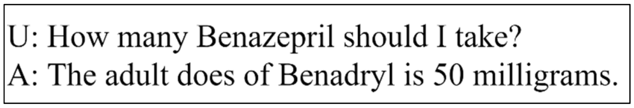
Device Failure Modes

1. **No internet connection:** modern VUIs work with the internet because the computation is done on cloud storage.
2. **Loss of electric power:** the VUIs won't be able to work in case of loss of power as they do not work with a chargeable battery.
3. **Failure to respond to commands:** the VUIs are equipped with activation switch to the microphone and turning it off won't recognize any commands.

Architecture Failure Modes

There are various failure modes of VUIs due to architectural design such as from SLU and NLG.

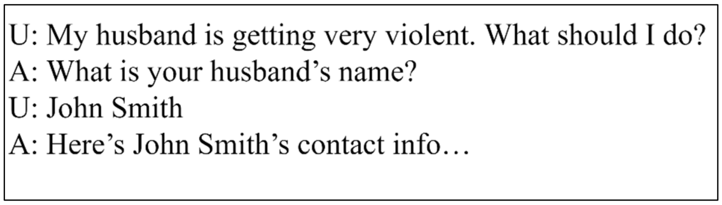
1. **Misrecognition of user's input:** This type of failure mode occurs when the CA fails to recognize a user's intent [40]. Figure 14 below provides an example of a CA having failed to recognize the user's input [55].



U: How many Benazepril should I take?
A: The adult does of Benadryl is 50 milligrams.

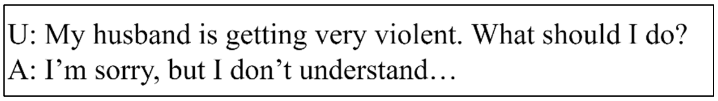
Figure 14: An example of VUI misrecognizing a user's input

2. **Errors in NLU/SLU:** The task of NLU is to extract the semantics from the user's utterance. NLU is a critical component in VUIs and failures in NLU have been shown to account for the user's dissatisfaction collected from 10,000 users of digital assistants [126]. The errors from NLU are classified as (1) misunderstanding and (2) non-understanding errors [127]. Figure 15 and Figure 16 show these NLU errors [55].



U: My husband is getting very violent. What should I do?
A: What is your husband's name?
U: John Smith
A: Here's John Smith's contact info…

Figure 15: An example of misunderstanding error in NLU



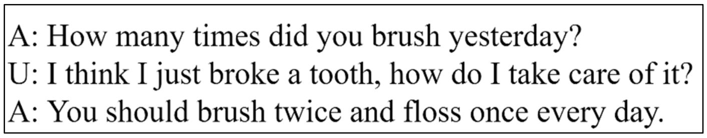
U: My husband is getting very violent. What should I do?
A: I'm sorry, but I don't understand…

Figure 16: An example of non-understanding error in NLU

Human-Machine Interaction Failure

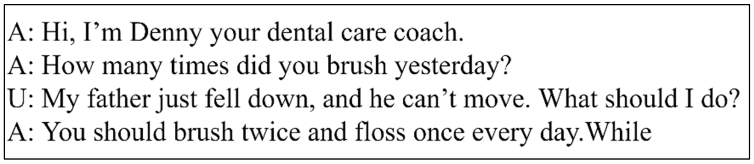
1. **Interaction in a noisy environment:** the VUI device may fail to understand a user's query if operated in a noisy environment. This was demonstrated by a pilot study where users were asked about the feasibility of VAs [128]. Although this is a technical failure of not recognizing the user's voice model, the noisy environment is the reason for its occurrence and hence the classification.

2. **Failure due to unconstrained user input:** This failure mode occurs when users are not restricted to provide their query and may go off-topic. Due to the absence of robust error detection in design, a user may ask about a topic for which the CA has no expertise in and was not designed to handle those inputs. Figure 17 and Figure 18 are some examples of these failures [55].



A: How many times did you brush yesterday?
U: I think I just broke a tooth, how do I take care of it?
A: You should brush twice and floss once every day.

Figure 17: An example of query beyond expertise of a CA



A: Hi, I'm Denny your dental care coach.
A: How many times did you brush yesterday?
U: My father just fell down, and he can't move. What should I do?
A: You should brush twice and floss once every day.While

Figure 18: An example of out of domain query of a CA

Summary

This section covered safety failure modes for VUIs. First, some general AI failure modes are discussed as the VUIs make use of AI-based algorithms. Next, failure modes due to errors in the architecture of VUI and device are briefly described. Finally, there may be safety failure modes due to the interaction of human-machine interaction which is highlighted in the last section. Failures due to the end-to-end architecture of VUIs and due to interaction with users need further exploration and this is highlighted in next year's plan in the report. A collection of important failure modes will help in finding techniques for controlling and mitigating hazards for VUIs and development of a safety case for them.

2.4. Human Factors

Careful consideration of human factors can significantly reduce healthcare incidents. The introduction of automation to the industries initially assumed that technology will eventually replace people at work. While for some tasks it is possible but for the larger part especially in industries like healthcare, the automation can help clinicians and staff doing their job easily and efficiently. The use of CAs in this regard is to reduce the workload on health staff for appointments, delivering low-risk and general queries at any time, and assisting clinicians with quick information at their disposal. As highlighted by Sujan et al. [129] introduction of AI in clinics introduces various human factor challenges which are shown in Figure 19.

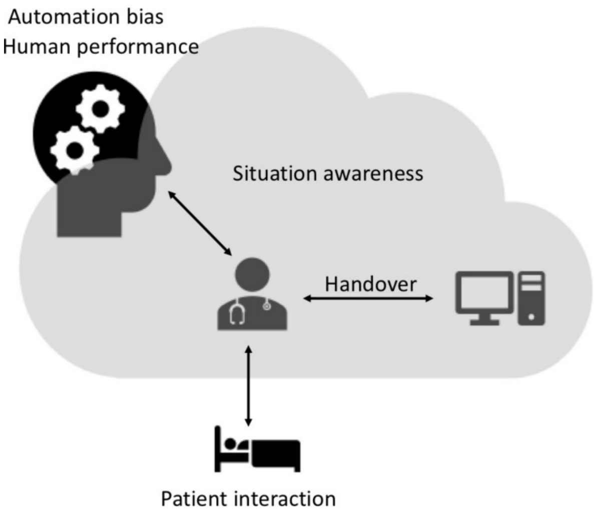


Figure 19: Human factors challenges overview of using AI in patient care

2.4.1. Automation Bias

Automation Bias is the overreliance on autonomous systems for decision making. Other terms are being used in the literature to describe automation bias, such as overreliance on automation, automation-induced complacency [130], and confirmation bias [131]. In healthcare, Clinical Decision Support Systems (CDSS) have great potential to assist and improve decisions made by clinicians. Most current Decision Support Systems (DSS) are known to have an accuracy of 80-90%. The occasional inaccurate decision from DSS may force the users to change the correct decision they already have made. This phenomenon of automation bias can lead to two types of errors: omission and commission errors [132] [133]. Omission errors occur for not taking appropriate action because of the failure of the system to alert and commission errors are caused by following of inappropriate advice of the system.

Goddard et al. mention four potential causes of automation bias from their systematic review on automation bias [133]. These include experience, confidence and trust, individual differences, and task type. The experience greatly affects the reliance on automation and it has seen that the problem of automation bias occurs more with inexperienced users of the system [134][135] than experienced ones [136]. Experienced physician or clinician might be less reliant on automation based on their experience and less prone to commission errors [134]. Confidence and trust is also an important virtue and human factor which plays an important role in overreliance on autonomy. Increased confidence in the user's own decisions decrease reliance on external support and thus saving themselves from errors. Dreiseitl et al. demonstrated that physicians were more likely to follow the advice of DSS when they were less confident on their diagnosis [137]. Apart from confidence, trust is the driving force in committing automation bias errors. This general human trait affects their decision to a greater extent when it comes to reliance on automation. A study conducted by Dzindolet et al. demonstrated this where users preferred automated assistance over human assistance and committed automation bias [138].

The nature of task and workload also a big cause of overusing a system. A user may become biased under increase work pressure and commit automation bias error [139]. Sarter et al. suggested that high time pressure may also be the source towards DSS [135]. These factors put stress on the cognitive capacity of humans and to compensate them users may over-rely on systems and it can work both in favour and against the user. As long as the system provides correct decision its use is beneficial but when the system does not advise correct it can cause more errors.

Automation bias can be avoided by using various preventive measures, one of which is the use of increased accountability for the decisions made by the system. This can result in a decrease in automation bias errors when there is external accountability [140]. Another study shows that it is not the perception of accountability that can decrease these errors but the user's attitude towards accountability and their work. This way another possibility to limit automation bias errors is by improving the work culture and providing training to the users.

2.4.2. Human Performance

The use of autonomy can also affect human performance in longer-term as AI systems usually trained on baseline data which is developed against human performance. Applying automated systems in healthcare may deteriorate the skills of physicians as technology will be solving most of the work which is learned by professionals in the industry. For example, the radiographers who may see images only from the AI systems in the future rather than the wide range of images they currently train on nowadays can substantially affect their performance [141]. Automation is applied where it provides economic benefit by performing more accurately and reliably than the human or by replacing the human at a lower cost. The misuse of automation by its design or implementation may cause overreliance by design and cause the human operator of the system to commit errors [132].

The introduction of automation in a team environment can also affect team performance. These tasks are quite complex and require human operators to complete several subtasks concurrently, for example performing individual responsibility while communicating to his team members. These complex tasks may not necessarily be improved by introducing automation [142]. Many automated systems interfere with team communication and coordination [143]. This situation is relatable to healthcare where the nature of the system is quite complex and team coordination is necessary. Thus, the introduction of automation in team task might not improve team performance.

2.4.3. Handover

Handover is defined as the transfer of control from an autonomous system to the human and is widely used in the context of autonomous vehicles [144]. Handover is a challenging aspect of the safety of critical systems. In the context of an autonomous vehicle, a driver can take control in case of an emergency or unforeseen situation. However, this may not always be possible as shown by the fatal accident caused by a Tesla in 2016 [145] and more recently in 2019.

Traditionally the concept of handover in healthcare is between clinicians or teams of clinicians [146]. Soon, handover between humans and autonomous agents will become more common and as Sujan et al. mention that it might be even more complex than the handover between the driver and the autopilot system of an autonomous vehicle [146]. In human handovers, there are protocols for structured communication that deliver key information such as Age, Time, Mechanism, Injuries, Signs, Treatments (ATMIST) for emergency care and Situation, Background, Assessment, Recommendation (SABR) more commonly [147]. There is a need for such protocols for AI systems considering the growing applications and interest in safety-critical industries.

An example case of an autonomous infusion pump for delivering insulin raises some questions over handover [129]. At what point does the autonomous agent need to handover the situation in terms of its struggle to maintain blood sugar levels? There needs to be a standard procedure for autonomous medical devices for these situations. CAs are paving their way in healthcare and the same analogy can be used as to when these CAs need to handover to the clinician if they struggle to understand user's query. Even big question for these agents would be to detect if they have misrecognized a user's utterance and to handover to the clinician before making a wrong diagnosis.

2.4.4. Situational Awareness

Situational awareness (SA) is defined as an individual's perception, comprehension and subsequent projection of what is going on in the environment [148]. It can also be said that the phenomenon of people sensing of things going on in their surroundings, understand what this information means, and plan necessary action or decision based on that information. The process of SA is part of people's cognitive functions and it helps them to understand tasks and make decisions. It involves three levels of cognitive performance [149]:

- Level 1: perceptions of elements in the environment.
- Level 2: comprehension of the current situation.
- Level 3: projection of future status.

SA is crucial in healthcare and patient safety and plays a crucial role in clinical care errors. The death of a young wife and mother from intubation from routine surgery highlighted the lack of SA in that situation [150]. There are solutions of overcoming the loss of SA when the process is done by humans, these include: taking into account human's individual and contextual factors such as fatigue, stress, team factor, stress and using checklists for simplification of tasks. These all are related to human SA and there is ongoing research on computer SA [151].

Healthcare consists of interdisciplinary teams where the communication between teams is an important aspect of patient safety. The information is context-specific and staff actions depend on the circumstances surrounding them. Introduction of AI in such a complex situation requires context or SA.

CAs in healthcare need to develop situation awareness too. For example, a CA for the use of advising medication need to be aware of the context in which the medication may be helpful or harmful for the user. A user might be taking other medicines suggested by a clinician and the need to properly communicate those to the CA is necessary to avoid potentially harmful situations.

2.4.5. Patient Interaction

Communication between patient and clinician in healthcare plays a key role in patient satisfaction. Healthcare aims to improve patients' condition and overall well-being which can be achieved rather easily by a strong partnership between patients and the clinicians. According to [152], seven key principles are critical to effective communication between patients and clinicians: mutual respect, harmonized goals, supportive environment, appropriate decision partners, the right information, full disclosure, and continuous learning. Empathy is considered a vital component in effective patient care which allows clinicians to perceive patients' conditions coherently and practically [153]. Clinicians, thus, can retrieve more information about their diseases and discomfort. Chronic diseases are a relevant example where a patient requires more support to adhere to their medication schedule. It has been seen that patients adhere to their medication better when provided with greater empathy [154]. CAs, because of their intuitiveness are a good medium which can approximate the relationship between a clinician and a patient [55]. They provide a great source of empathy to patients as they communicate in natural language. As discussed in previous sections, CAs have a large number of applications in chronic disease management such as diabetes [78][155] and they are also being used for medication adherence and reminder [35][36][156].

Summary

This section covered various human factors that need to be considered for the use of AI-based automated systems in healthcare. Automation bias and human performance both are the human factors which can affect human performance by overreliance on autonomous systems. CAs assisting clinicians can have an impact on the performance of human clinicians if used too often because of the AI algorithm underneath them. Other human factors such as handover, SA, and patient interaction may have direct application to CAs in healthcare. A CA operating under the clinical supervision may need to handover the control to the clinician if it cannot understand the patient's query. Similarly, a CA monitoring a patient's chronic condition or suggesting a medication need to have SA about the history of the patient. Without developing SA, it can suggest a medication which may be correct for the given scenario but maybe harmful given the overall context of the patient. Lastly, patient interaction is an area where CAs can help greatly but it is too early to say if they can replace the connection and relation a clinician develops with their patients.

2.5. Ethical Issues

Ethical issues in AI systems are quite an old debate. CAs, inherently being an intelligent system adopt this trait as well. The most important of which are bias, privacy, and safety concerns [7]. Bias is defined as prejudice or favouritism for or against a person or group in an unfair manner. Microsoft's Tay bot which was taken off from twitter only after 16 hours is a popular example of bias. The CA learned from user's data and started posting tweets containing racial, abusive and personal attacks [157]. Most of the CAs use data from social media platforms in training their models because of the difficulty and sparsity of data. It has been found that the corpora took from social media twitter, reddit, etc. contains offensive language and hate speech to train CAs. The data from social media may have inherent biases and data filtration is hard to achieve.

As VUIs are becoming popular household devices, privacy is a major concern to most users. In healthcare, people are reluctant to share their personal and medical information with AI systems. VA manufacturers have known to be using people's conversations to improve the quality of their services [158]. This open admission from the companies also adds to the worries of people that someone might be listening to their conversation. Given this, the anonymity of people and consent to use their information is a greater challenge [39].

The most important of them all, safety, is described as the avoidance of harm from unintended behaviour. The applications of CAs in healthcare require that they depict safe behaviour in the context of their use. For example, a medical diagnostic application must not provide a user with wrong diagnoses or suggest medicine which is unsafe for that user. There are already instances [78][28] where people are trying to make use of cases for CAs as a diagnostic application. For the safety of humans, thus, most clinicians believe that CAs in the health industry should not be used without human supervision [50][14]. In one study [11] safety risks associated with VAs were observed by asking them daily routine queries that require medical assistance. The writers evaluated these VAs with the help of pre-written questions and 54 participants. It was noted overall that the consumers should not rely solely on the VAs because of misrecognition of queries, limited knowledge in healthcare, and lack of clear understanding of people's questions.

3. Conclusion

This report presented literature work on safety issues of CAs in healthcare. CAs are becoming a popular tool in healthcare because they provide various useful services. They provide patients with general-purpose health information, symptom checking, medication adherence etc. CAs also assist clinicians in decision making with large amounts of data. Their use in healthcare may cause harm to patients without proper supervision because of their autonomous nature. Healthcare is a safety-critical domain where people are at greater risk because of their existing health condition as compare to other safety-critical industries such as chemical, nuclear, etc. From the literature of CAs, we have found that there are many safety implications from these devices. Lack of clinical data for training ML models, unconstrained user input for patients to interact, errors in understanding user's query, and absence of a clinician in monitoring the decision of CA are some of those safety concerns. We also found out that most clinicians believe that CAs should not be used in healthcare without clinician supervision. The reason for that is they are autonomous agents and are yet not mature enough in performing medical decisions. Therefore, in their current form, they may pose a safety risk to patients.

There is a need to provide safety assurance of these CAs in healthcare by their manufacturers or providers. CAs that diagnose disease and provide users with recommended medicines are classified as medical devices. Symptom checker CAs are low-risk medical devices while diagnostic CAs are considered medium risk medical devices. A safety assurance case helps developers, manufacturers, regulators and other stakeholders to understand the safe use of the system. It is a risk-based argument with evidence to demonstrate all safety risks associated with a system have been identified and reduced to an acceptable safety level. There are various safety analysis methods used to assess and identify hazards in safety-critical industries. Some of them are Hazard and Operability Analysis (HAZOP), Fault Tree Analysis (FTA), FMEA, HTA, etc. There are international standards for the safety of medical devices, but they lack in providing guidelines for the design and development of AI-based systems such as CAs. The regulation of AI systems is also a challenging task due to the dynamic decision making of such systems. In the UK, there is no single authority to regulate AI systems in healthcare and that too may have safety implications for commercial CAs. Given the potential risks associated with CAs in healthcare, we believe that the safety case of these devices is needed before they can be used commercially.

This report focuses on task-oriented CAs and their safety issues in healthcare. Since the purpose of the majority of CAs in healthcare is to complete a user's task such as set a reminder for medicine, provide drug-related information, recommend an action based on user's symptoms etc. We identified various hazards and failure modes from the architecture of these systems. Failure modes are the errors or faults in the system that lead to system failure and may affect patient safety. Since the CAs use AI to make a decision, AI failure modes may also directly affect the robustness of CAs. Distributional shift, black-box decision making, negative side effects, automation complacency are some of the common AI failure modes. CAs may also fail due to error in their architecture such as NLU failures, dialogue policy errors, improper response generation, etc. Sometimes a failure may occur from human-machine interaction and may lead to safety concern. We identified background noise as a key failure mode due to which CA fail to understand the user correctly. Another important safety consideration is to restrict the user's input as CAs fail when a user asks a query beyond their scope. These types of failures fall under usability issues and some studies reveal major safety concerns as the systems provided an unsafe response. Network latency, poor internet connection and software issues also cause CAs to fail. In conclusion, there are various ways a CA might fail and some of these failures might cause harm to the patient in a healthcare environment. We understand that while designing these systems not only the software design, but their usability and human-machine interactions need to be considered carefully. Although achieving 100 per cent safety is not possible these design considerations may make the use of CAs in healthcare relatively safe.

Introduction of autonomous systems such as CAs introduces various human factor challenges. CAs assisting clinicians in decision making might make them over-rely on the technology and this may affect clinician's performance. Handover is possibly the most challenging human factor that is closely related to safety. CAs in healthcare may need to handover the control to a clinician if they struggle to understand the user's query. An even big question may arise if they will be smart enough to detect the incapability to provide a safe response and transfer control to a clinician before making an unsafe decision themselves. SA is also an important factor to consider by CAs before making decisions in the healthcare environment as they need to be aware of the context of the patient's situation. CAs may change the way patients interact to clinicians if they are used in healthcare. They may provide affordable and 24-hour services to users, but it is still a question if they can be a substitute to clinicians empathy towards their patients. We believe that human factors should be carefully considered while designing and introducing CAs as a safe medical device in healthcare.

Ethical use of AI systems is an old debate and CAs inherently being an intelligent system are no exception here. We identified bias, privacy and safety concerns to be the most important ethical issues concerning the use of CAs. A lot of CAs are trained on social media conversations and may be biased towards decision making or the bias of favouring one manufacturer's drug over others can be introduced deliberately by the designers. The training data may have an inherent bias as well and therefore it is important to filter training data or manually create training data. Privacy is another major concern as many people do not want to use CAs to share their personal and medical information. Some manufacturers even admitted having listened to user's conversations to improve the performance of their CAs. Data ownership is thus a key ethical concern for CAs. Safety, as discussed thoroughly throughout this report, is the main ethical concern as not to cause any harm to patients using CAs. Solely reliance on the CAs may not be good for the users and clinician supervision may be required for safe decision making. In conclusion, to introduce CAs in healthcare, ethical issues need to be discussed and resolved where possible as without addressing them people will be hesitant to adopt CAs. Compliance with data protection and privacy rules such as General Data Protection Regulation (GDPR), getting user's consent to use their information, and deanonymize their personal information may help to address ethical concerns of users.

Appendix

A. Natural Language Processing

Natural Language Processing (NLP) is a subfield of linguistics, computer science, information engineering and artificial intelligence (AI) concerned with the interaction between computers and human (natural) languages. NLP is about programming computers to understand and process data in natural language. It employs computational techniques for understanding, learning, and producing human language content. Computational linguistics is a practical technology which is being incorporated into consumer products, such as VUIs, language translation tools (Google Translator), modern cars, etc. The key enablers behind these developments are (1) increase in computing power, (2) availability of a large amount of linguistic data, (3) advancements in ML methods, and (4) richer understanding of human language structure [159]. Major challenges in NLP involve speech recognition, natural language understanding and natural language generation which are core components of a VUI.

The NLP pipeline for text data consists of three main components text processing, feature extraction, and modelling. These are shown in Figure 20 below. The next sections provide details on these components that make up the pipeline for NLP.

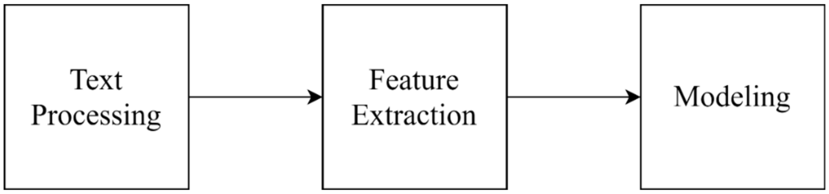


Figure 20: Typical NLP pipeline

Key NLP Terms and Concepts

There are few important concepts or terms which are used in the processing of natural language or text and before we begin explaining NLP pipeline, it is better to mention those terms which are frequently used in NLP.

Corpus: Corpus (plural corpora) is a large set of structured text.

Utterance: An utterance is the spoken correlate of a sentence.

Fillers: Words in a spoken language such as uh, um, hmm are called fillers or sometimes filled pauses.

Lemma: A lemma is a set of lexical forms having the same stem and same word sense.

Word Form: The word form is a derived or full inflected form of the word. For example, ‘cats’ is the word form of the lemma cat.

Word Type: Number of distinct words in a corpus is called types. For example, if the set of words in the vocabulary is V the number of types is $|V|$.

Tokens: Tokens are the total number N of running words in a corpus.

Stemming: It is the process of reducing the word form to its stem. For example, ‘branching’, ‘branched’, or branches can be reduced to their stem word ‘branch’.

Lemmatization: The process of transforming different words to their root. For example, ‘is’, ‘was’, and ‘were’ can be lemmatized to their common root ‘be’.

A.1 Text Processing

This is the first step in processing natural language for computers. Text processing deals with the input which may come from the web, documents, books, or even can come from the speech input. The goal of text processing is to generate plain text from different sources of data such that it does not contain any source-specific constructs or markers in it.

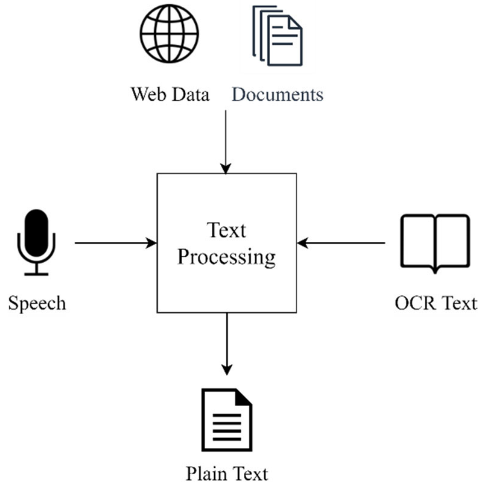


Figure 21: Text processing in NLP pipeline

Normalization

Normalization is the first step in text processing. For example, in the English language start of a sentence is done by a capital letter, while from the reading perspective of a person it may be important but for the computer, it is not because the meaning remains the same in both cases. Case normalization helps in reducing the number of unique tokens. Depending on the NLP task, punctuations and special characters may or may not be removed during normalization.

Tokenization

Tokenization is the process of segmenting running text into words. Often punctuations are kept and considered as a separate token. For example, commas are useful information for text parsers, and period *(.)* helps in identifying the boundaries of a sentence in a long text. Mostly punctuations inside a word such as a period sign in abbreviation (Ph.D.), a character in a name AT&T, or an apostrophe in a word (I’ve). Similarly, special characters and numbers are usually kept together to understand the meaning of the text. This include prices ($37.13), date (01/02/2020), email addresses (abz@xyz.uk), hashtags (#NLP), and URLs (http://www.york.ac.uk), etc.

Stemming and Lemmatization

Lemmatization is the task of determining if the two words have the same root, despite their surface differences. The word ‘be’ is the shared lemma for words am, are, and is; words dinner and dinner have the same shared lemma ‘dinner’. Stemming is a complex process, but the simplest approach is to remove off suffixes from the word.

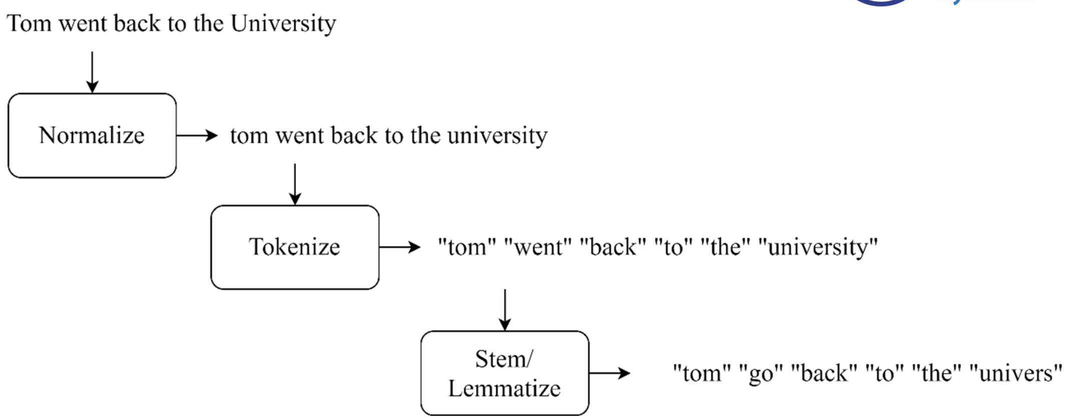

Tom went back to the University
Normalize
tom went back to the university
Tokenize
"tom" "went" "back" "to" "the" "university"
Stem/
Lemmatize
"tom" "go" "back" "to" "the" "univers"

Figure 22: An example sentence after text processing

Sentence Segmentation

Sentence segmentation also plays an important role in text processing. For this purpose, punctuation like periods, question marks, exclamation points provides useful information. Period (.) is an ambiguous marker when it comes to sentence segmentation. The reason is their occurrence at the end of a sentence boundary and between abbreviations that make them so difficult to use for segmentation. Sometimes a sentence can also end on an abbreviation such as *Inc.* and in this case, sentence boundary marker is also the same period marker. On the other hand, question mark and exclamation are relatively unambiguous markers of sentence boundaries. In general, sentence tokenization works by first determining that a period is part of a word or a sentence boundary. This can be decided using rule-based approaches or machine learning. For abbreviations a dictionary of abbreviations can be used for known abbreviations; these dictionaries may be built by hand or using machine learning.

Part-of-Speech (POS) Tagging

Part-of-Speech (POS) (noun, verb, and preposition) can help in understanding the meaning of a text by identifying how different words are used in a sentence. POS can reveal a lot of information about neighbouring words and syntactic structure of a sentence. POS tagging is the process of assigning a POS marker (noun, verb, etc.) to each word in an input text. The input to a POS tagging algorithm is a sequence of tokenized words and a tag set (all possible POS tags) and the output is a sequence of tags, one per token. Words in the English language are ambiguous because they have multiple POS. For example, a book can be a verb (book a flight for me) or a noun (please give me this book). POS tagging aims to resolve those ambiguities.

There are various common tagsets for the English language that are used in labelling many corpora. 45-tag Penn Treebank tagset is one of such important tagset [160]. This tagset also defines tags for special characters and punctuation apart from other POS tags. The Brown, WSJ, and Switchboard are the three most used tagged corpora for the English language. The Brown corpus consists of a million words of samples taken from 500 written texts in the United States in 1961. The WSJ corpus contains one million words published in the Wall Street Journal in 1989. The Switchboard corpus has twice as many words as Brown corpus. The source of these words is recorded phone conversations between 1990 and 1991. For tagging words from multiple languages, tagset from Nivre et al. [161] is used which is called the Universal POS tagset. The tagset is part of the Universal Dependencies project and contains 16 tags and various features to accommodate different languages. The main application of POS tagging is in sentence parsing, word disambiguation, sentiment analysis, question answering and Named Entity Recognition (NER). The last of which is defined in the next section.

Named Entity Recognition (NER)

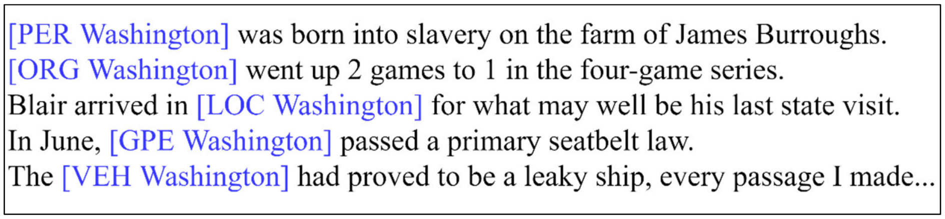
Named entity recognition is the task of extracting named entities information from text. A named entity is anything that can be referred to as a proper name: a person, a place, or a location. Moreover, it also includes terms that as such are not considered entities, this includes dates, time, price, etc. Table 6 provides a list of generic named entity types [6], although many applications may need to define entities based on their specific needs.

Table 6: A list of generic named entity types

Type	Tag	Categories	Sentences
People	PER	people, characters	Turing is a giant of computer science.
Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
Location	LOC	regions, mountains, seas	The Mt. Sanitas loop is in Sunshine Canyon.
Geo-Political Entity	GPE	countries, states, provinces	Palo Alto is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the Golden Gate Bridge.
Vehicles	VEH	planes, trains, automobiles	It was a classic Ford Falcon.

There are certain ambiguities in NER, which arise from the ambiguity of segmentation; to understand what an entity and the boundaries of sentences is. Type ambiguity is common as a tag can be categorized in more than one possible category. For example, JFK can refer to a person and an airport at the same time. Figure 23 provides an example of type ambiguity in NER when using the word Washington [6].

Depending on the context where it is used, a named entity can be different for the same word.



[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.
The [VEH Washington] had proved to be a leaky ship, every passage I made...

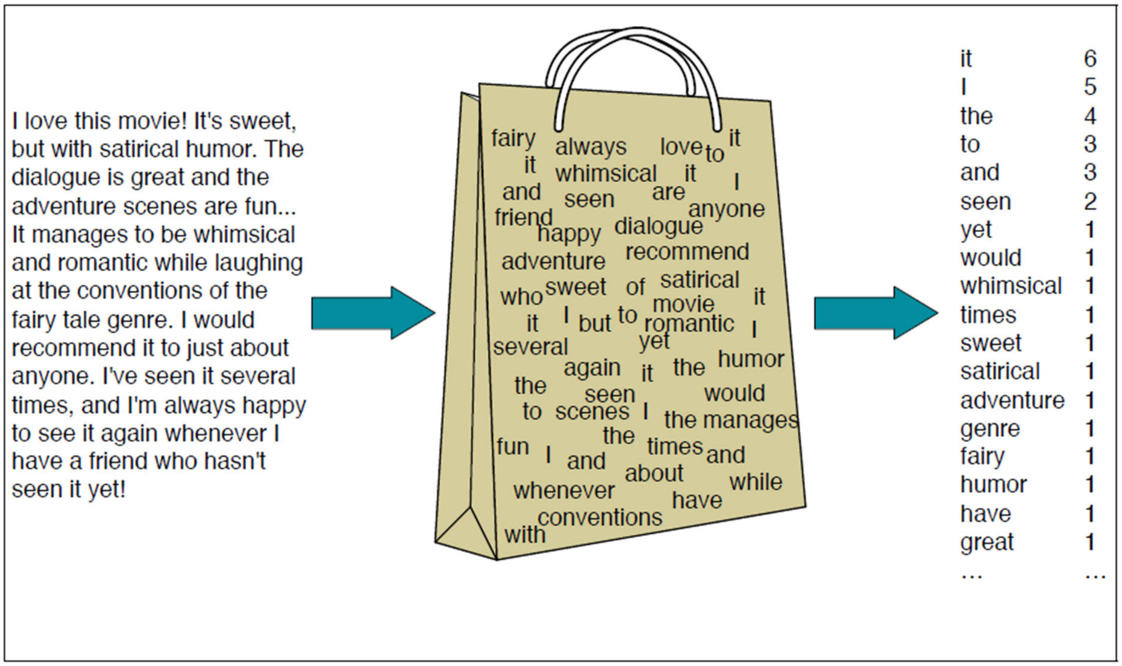
Figure 23: Type ambiguity example in NER

A.2 Feature Extraction

Feature extraction is the process of extracting important characteristics from word data. Below we define some feature extraction tasks.

A.2.1 Bag of Words (BoW)

Bag of words (BoW) is a basic model for feature extraction from text. It treats each document of text as a bag with unordered words and does not include any structure or syntax of the words [114]. Tokenized words from each document of text are used to find the frequency of each token. Figure 24 [6] shows the bag of words approach. Here the sentences are not kept in order and instead only tokens with their frequency count in the document are preserved.



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Figure 24: An example of a Bag of Words (BoW) approach

A.2.2 TF-IDF

The problem with BoW approach for feature extraction is that it treats all words equally. Depending on the type of corpus for text processing, some terms occur frequently. For example, in a corpus of movie data, few words such as 'dialogue', 'actors', 'genre', etc. are more common than the others. Secondly, not all the words are discriminative and common in a corpus such as 'the', 'it', or 'and', etc. may not provide the context or information about keywords. To overcome these limitations, TF-IDF helps by providing weightage to words in a corpus. The TF-IDF algorithm is the product of two words; term frequency (TF), and inverse document frequency (IDF) [6].

The term frequency (TF) represents a term t in a document d and is calculated by the below equation.

$$tf_{t,d} = count(t, d)$$

To remove raw frequency this frequency is altered by taking the logarithm $\log_{10}$. Since a term appearing 50 times in a document does not necessary implies 50 times more likely to be relevant. Thus, TF is calculated by adding one to the calculation as $\log_{10}$ of 0 is 1.

$$tf_{t,d} = log_{10}\,(count(t, d) + 1)$$

The second term of TF-IDF algorithm deals with assigning more weight to terms which occur infrequently. The document frequency df_t of a term, t is the number of documents containing that term. The IDF is defined as N/df_t. The IDF for N documents with document frequency df_t is calculated as:

$$idf_t = log_{10}\,(\frac{N}{df_t})$$

The TF-IDF weighted value for a word t in document d is computed by multiplying TF and IDF values as below:

$$w_{t,d} = (tf_{t,d} \times idf_t)$$

A.3 Modelling

Probabilities are a convenient way to predict upcoming words in a sentence. For speech recognition tasks where the input is noisy and ambiguous probabilities can help speech recognizer to a greater extent in understanding the user's input. Probabilistic models that assign probabilities to a sequence of words are called language models. In general, they are referred as n-grams to assign the probabilities in a sequence of N words: a bigram is a two-word sequence such as "please turn", and a trigram refers to a three-word sequence such as "please turn over" or "turn your homework". Given the probability of the bigrams, or trigrams, or simply n-grams, these language models predict the probability of next word in a sequence.

A.3.1 N-Grams

The probability of a word w given some history h is computed as:

$$P(w|h) = \frac{C(hw)}{C(h)}$$

Where hw refers to the sequence where the word w is followed by history h. Probability of an entire sequence of words $(w_1, w_2, \dots w_n\ )$ or w_1^n is computed by chain rule of the probability of individual words by decomposing as:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\dots P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

The problem with this probabilistic approach is that you cannot find sequences from the large corpus or even from the web for your specific task. Even some sequences of as few as 4 or 5 words it is harder to find them in a corpus and count their instances for computing the probability.

The intuition behind N-gram is that to calculate the probability of a word given entire sequence, we can approximate it by the history of the last few words and not the entire sequence. If we use a bigram model to approximate this probability, we can approximate the probability by only computing the probability of the last word:

$$P(w_n|w_1^{n-1}) = P(w_n|w_{n-1})$$

Probability of a complete word sequence given the bigram assumption of an individual word thus can be computed as:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

The probability of bigram of a word y given previous word x is to compute the count of bigram $C(xy)$ and normalize it by the sum of all bigrams that share the same first-word x which is equivalent to the unigram count of the previous word w_{n-1}.

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}\ w_n)}{C(w_{n-1}\ )}$$

In practice, trigram models are used which uses the previous two words in determining the probability of the next word. If there is sufficient training data available, then 4-gram or sometimes 5-gram models are also used.

A.3.2 The Hidden Markov Models

The Hidden Markov Models (HMM) is a statistical model for modelling generative sequences characterized by an underlying process generating an observable sequence. HMMs have various applications such as in speech recognition, signal processing, and some low-level NLP tasks such as POS tagging, phrase chunking, and extracting information from documents. HMM are based on Markov chains. A Markov chain is a model that describes a sequence of potential events in which the probability of an event is dependant only on the state which is attained in the previous event. Markov model is based on a Markov assumption in predicting the probability of a sequence. If state variables are defined as q_1, q_2… q_i, a Markov assumption is defined as:

$$\boldsymbol{Markov\ Assumption}: P(q_i = a|q_i \dots q_{i-1}) = P(q_i = a|q_{i-1})$$

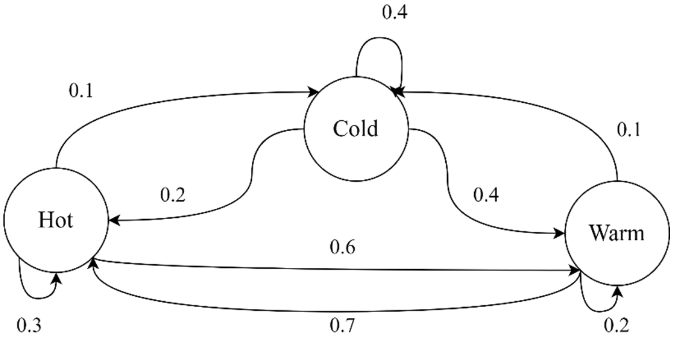


Figure 25: A Markov chain with states and transitions

Figure 25 shows an example Markov chain for assigning a probability to a sequence of weather events. The states are represented by nodes in the graph while edges represent the transition between states with probabilities. A Markov chain is useful when the events of interest are observable. An HMM is useful for both observed and hidden (such as POS tags where it is unknown given a POS that which word it belongs to) sequence of events.

A first order HMM is based on two assumptions. One of them is Markov assumption that is the probability of a state depends only on the previous state as described earlier, the other is the probability of an output observation o_i depends only on the state that produced the observation q_i and not on any other states or observations.

$$\boldsymbol{Output\ Independence}: P(o_i|q_1 \dots q_i, \dots q_T, o_1 \dots o_i, \dots o_T) = P(o_i|q_i)$$

An HMM consists of two components, the A and the B probabilities. The A matrix contains the tag transition probabilities $P(t_i \mid t_{i-1})$ and B the emission probabilities, $P(w_i \mid t_i)$ where w denotes the word and t denotes the tag.

The transition probability, given a tag, how often is this tag is followed by the second tag in the corpus is calculated as:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

The emission probability, given a tag, how likely it will be associated with a word is given by:

$$P(w_i|t_i) = \frac{C(t_i\,, w_i)}{C(t_i\,)}$$

Figure 26 shows an example of the HMM model in POS tagging. For a given sequence of three words, “words1”, “words2”, and “words3”, the HMM model tries to decode their correct POS tag from “N”, “M”, and “V”. The A transition probabilities of a state to move from one state to another and B emission probabilities that how likely a word is either N, M, or V in the given example.

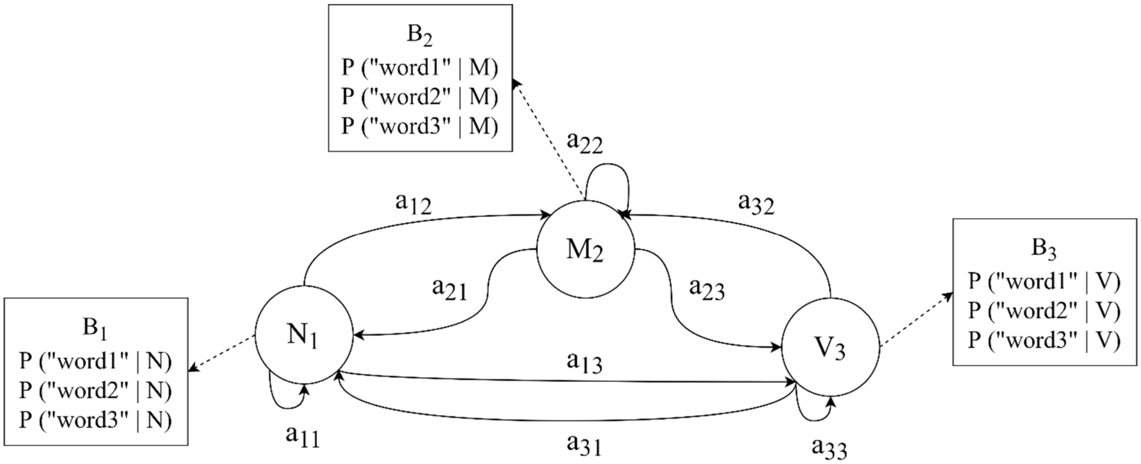


Figure 26: A Hidden Markov Model (HMM) with A transition and B emission probabilities

HMM Tagger

The process of determining hidden states to their corresponding sequence is known as decoding. More formally, given A, B probability matrices and a sequence of observations $O = o_1 \ldots o_2, \ldots o_T$, the goal of an HMM tagger is to find a sequence of states $Q = q_1 \ldots q_2, \ldots q_T$. For POS tagging the task is to find a tag sequence $t^{\wedge n}_{\ 1}$ that maximizes the probability of a sequence of observation of n words w_1^n.

$$t^{\wedge n}_{\ 1} = \max_{t_1^n} P(t_1^n|w_1^n) \approx \max_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)\; P(t_i|t_{i-1})$$

The Viterbi Algorithm

The decoding algorithm for the HMM model is the Viterbi Algorithm. The algorithm works as setting up a probability matrix with all observations o_t in a single column and one row for each state q_i. A cell in the matrix $v_t(j)$ represents the probability of being in state j after first t observations and passing through the highest probability sequence given A and B probability matrices. Each cell value is computed by the following equation:

$$v_t(j) = \max_{q_1..q_{t-1}} P\;(q_1..q_{t-1}, o_1, o_2...o_t, q_t = j|(A, B))$$

Figure 27 shows an example of a Viterbi matrix with states (POS tags) and a sequence of words. A greyed state represents zero probability of word sequence from the B matrix of emission probabilities. Highlighted arrows show word sequence with correct tags having the highest probabilities through the hidden states.

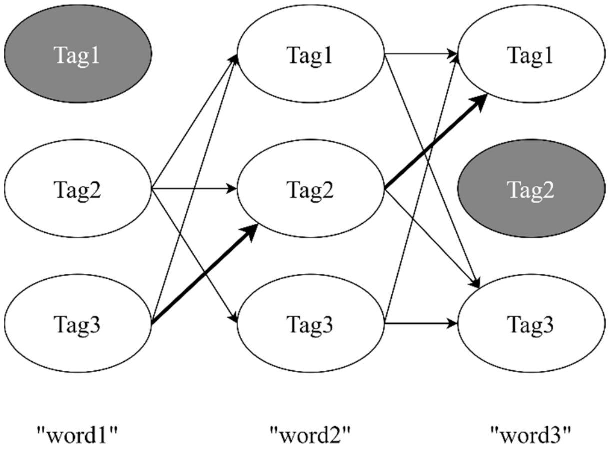


Figure 27: Viterbi matrix with possible tags (q_i) for each word

The Viterbi algorithm works recursively to compute each cell value. For a given state q_j at time t, the Viterbi probability at time t, $v_t(j)$ is calculated as:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}\,(i)\; a_{ij}\;\; b_j(o_t)$$

The components used to multiply to get the Viterbi probability are the previous Viterbi path probability from the previous time $v_{t-1}\,(i)$, a_{ij} the transition probability from the previous state q_i to current state q_j, and $b_j(o_t)$ the state observation likelihood of the observation symbol o_t given the current state j.

References

[1] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR mHealth uHealth*, vol. 6, no. 11, pp. 1–14, 2018, doi: 10.2196/12106.

[2] R. R. Morris, K. Kouddous, R. Kshirsagar, and S. M. Schueller, "Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions," *J. Med. Internet Res.*, vol. 20, no. 6, p. e10148, 2018, doi: 10.2196/10148.

[3] S. Diederich and A. Benedikt Brendel, "Towards a Taxonomy of Platforms for Conversational Agent Design Digital Nudging View project Chatbots and Gamification View project," no. February, pp. 1100–1114, 2019.

[4] A. B. Kocaballi *et al.*, "Personalization of Conversational Agents in Healthcare: A Systematic Review (Preprint)," *J. Med. Internet Res.*, vol. 21, pp. 1–15, 2019, doi: 10.2196/15360.

[5] L. Laranjo *et al.*, "Conversational agents in healthcare: A systematic review," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 9, pp. 1248–1258, 2018, doi: 10.1093/jamia/ocy072.

[6] J. Pollock, E. Waller, and R. Politt, "Speech and language processing," *Day-to-Day Dyslexia in the Classroom*, 2010. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[7] P. Henderson *et al.*, "Ethical Challenges in Data-Driven Dialogue Systems," *AIES 2018 - Proc. 2018 AAAI/ACM Conf. AI, Ethics, Soc.*, pp. 123–129, 2018, doi: 10.1145/3278721.3278777.

[8] K. Denecke, M. Tschanz, T. L. Dorner, and R. May, "Intelligent conversational agents in healthcare: Hype or hope?," *Stud. Health Technol. Inform.*, vol. 259, no. April, pp. 77–84, 2019, doi: 10.3233/978-1-61499-961-4-77.

[9] N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents," 2017.

[10] G. Pilato, A. Augello, and S. Gaglio, "A modular architecture for adaptive ChatBots," *Proc. - 5th IEEE Int. Conf. Semant. Comput. ICSC 2011*, pp. 177–180, 2011, doi: 10.1109/ICSC.2011.68.

[11] S. Olafsson, T. K. O'Leary, R. Asadi, N. M. Rickles, and R. Cruz, "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.*, vol. 20, no. 9, pp. 1–13, 2018, doi: 10.2196/11510.

[12] A. Fadhil, "Addressing b Challenges b in b Promoting b Healthy b Lifestyles : b The b AI-Chatbot b Approach b," 2017.

[13] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related FAQs," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1525–1530, 2017, doi: 10.1109/ICACCI.2017.8126057.

[14] M. Bates, "Health Care Chatbots Are Here to Help," *IEEE Pulse*, vol. 10, no. 3, pp. 12–14, 2019, doi: 10.1109/MPULS.2019.2911816.

[15] J. Ghofrani and D. Reichelt, "Using voice assistants as HMI for robots in smart production systems," *CEUR Workshop Proc.*, vol. 2339, no. February, pp. 62–65, 2019.

[16] M. Ma, M. Skubic, K. Ai, and J. Hubbard, "Angel-Echo: A Personalized Health Care Application," *Proc. - 2017 IEEE 2nd Int. Conf. Connect. Heal. Appl. Syst. Eng. Technol. CHASE 2017*, pp. 258–259, 2017, doi: 10.1109/CHASE.2017.91.

[17] O. Sun, J. Chen, and F. Magrabi, "Using Voice-Activated Conversational Interfaces for Reporting Patient Safety Incidents: A Technical Feasibility and Pilot Usability Study.," *Stud. Health Technol. Inform.*, vol. 252, pp. 139–144, 2018, doi: 10.3233/978-1-61499-890-7-139.

[18] K. Nimavat and T. Champaneria, "Chatbots: An Overview Types, Architecture, Tools and

Future Possibilities," *Int. J. Sci. Res. Dev.*, vol. 5, no. 7, pp. 1019–1026, 2017.

[19] D. Thompson and T. Baranowski, "Chatbots as extenders of pediatric obesity intervention: An invited commentary on 'feasibility of Pediatric Obesity & Pre-Diabetes Treatment Support through Tess, the AI Behavioral Coaching Chatbot,'" *Transl. Behav. Med.*, vol. 9, no. 3, pp. 448–450, 2019, doi: 10.1093/tbm/ibz065.

[20] A. M. Rahman, A. Al Mamun, and A. Islam, "Programming challenges of chatbot: Current and future prospective," *5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017*, vol. 2018-Janua, pp. 75–78, 2018, doi: 10.1109/R10-HTC.2017.8288910.

[21] R. Sharma and M. Patel, "Review on Chatbot Design Techniques in Speech Conversation Systems," *Iarjset*, vol. 5, no. 9, pp. 37–46, 2018, doi: 10.17148/iarjset.2018.596.

[22] K. Chung and R. C. Park, "Chatbot-based heathcare service with a knowledge base for cloud computing," *Cluster Comput.*, vol. 22, no. s1, pp. 1925–1937, 2019, doi: 10.1007/s10586-018-2334-5.

[23] D. Gannon, "Building a ' ChatBot ' for Scientific Research," no. August, 2018, doi: 10.13140/RG.2.2.17641.39528.

[24] C. Segura, J. Luque, and M. R. Costa-juss, "Chatbol , a chatbot for the Spanish 'La Liga,'" *Int. Work. Spok. Dialog Syst. Technol. 2018*, pp. 1–12, 2018.

[25] L. Vaira, M. A. Bochicchio, M. Conte, F. M. Casaluci, and A. Melpignano, "Mama bot: A system based on ML and NLP for supporting women and families during pregnancy," *ACM Int. Conf. Proceeding Ser.*, pp. 273–277, 2018, doi: 10.1145/3216122.3216173.

[26] A. Vegesna, P. Jain, and D. Porwal, "Ontology based Chatbot (For E-commerce Website)," *Int. J. Comput. Appl.*, vol. 179, no. 14, pp. 51–55, 2018, doi: 10.5120/ijca2018916215.

[27] A. Argal, S. Gupta, A. Modi, P. Pandey, S. Shim, and C. Choo, "Intelligent travel chatbot for predictive recommendation in echo platform," *2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018*, vol. 2018-Janua, pp. 176–183, 2018, doi: 10.1109/CCWC.2018.8301732.

[28] S. Mujeeb, M. Hafeez, and T. Arshad, "Aquabot: A Diagnostic Chatbot for Achluophobia and Autism," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 209–216, 2017, doi: 10.14569/ijacsa.2017.080930.

[29] S. M. Jungmann, T. Klan, S. Kuhn, and F. Jungmann, "Accuracy of a Chatbot (Ada) in the Diagnosis of Mental Disorders: Comparative Case Study With Lay and Expert Users," *JMIR Form. Res.*, vol. 3, no. 4, p. e13863, 2019, doi: 10.2196/13863.

[30] A. Piau, R. Crissey, D. Brechemier, L. Balardy, and F. Nourhashemi, "A smartphone Chatbot application to optimize monitoring of older patients with cancer," *Int. J. Med. Inform.*, vol. 128, no. May, pp. 18–23, 2019, doi: 10.1016/j.ijmedinf.2019.05.013.

[31] B. Chaix *et al.*, "When chatbots meet patients: One-year prospective study of conversations between patients with breast cancer and a chatbot," *J. Med. Internet Res.*, vol. 21, no. 5, pp. 1–7, 2019, doi: 10.2196/12856.

[32] S. Razzaki *et al.*, "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis," pp. 1–15, 2018.

[33] WebMD, "WebMd Health." [Online]. Available: https://symptoms.webmd.com/default.htm#/info.

[34] Ada, "Ada Health." [Online]. Available: https://ada.com/app/.

[35] J. Allen *et al.*, "Chester: Towards a personal medication advisor," *J. Biomed. Inform.*, vol. 39, no. 5, pp. 500–513, 2006, doi: 10.1016/j.jbi.2006.02.004.

[36] D. Madhu, C. J. N. Jain, E. Sebastain, S. Shaji, and A. Ajayakumar, "A novel approach for medical assistance using trained chatbot," *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2017*, no. Icicct, pp. 243–246, 2017, doi: 10.1109/ICICCT.2017.7975195.

[37] J. Pereira and Ó. Díaz, "Using Health Chatbots for Behavior Change: A Mapping Study," *J. Med. Syst.*, vol. 43, no. 5, 2019, doi: 10.1007/s10916-019-1237-1.

[38] C. Vincent, *Patient Safety: 2nd edition*. 2010.

[39] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, no. 10. Nature Publishing Group, pp. 719–731, 01-Oct-2018, doi: 10.1038/s41551-018-0305-z.

[40] T. W. Bickmore *et al.*, "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.*, vol. 20, no. 9, pp. 1–13, 2018, doi: 10.2196/11510.

[41] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," *Can. J. Psychiatry*, vol. 64, no. 7, pp. 456–464, 2019, doi: 10.1177/0706743719828977.

[42] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv Prepr. arXiv1606.06565*, 2016.

[43] T. Hodgson, F. Magrabi, and E. Coiera, "Efficiency and safety of speech recognition for documentation in the electronic health record," *J. Am. Med. Informatics Assoc.*, 2017, doi: 10.1093/jamia/ocx073.

[44] MHRA, "Medical device stand-alone software including apps (including) IVDMDs."

[45] *Goal structuring notation community standard (version 2)*. The Assurance Case Working Group, 2018.

[46] D. Jurafsky and J. H. Martin, "Speech and Language Processing," 2009.

[47] M. ROBERTS, "OK Google, Siri, Alexa, Cortana; Can you tell me some stats on voice search?," 2018. [Online]. Available: https://edit.co.uk/blog/google-voice-search-stats-growth-trends/. [Accessed: 11-Mar-2020].

[48] M. ROBERTS, "OK Google, Siri, Alexa, Cortana; Can you tell me some stats on voice search?," 2018. .

[49] Amazon, "Amazon Alexa Health and Fitness Skills." [Online]. Available: https://www.amazon.co.uk/s?i=alexa-skills&rh=n%3A10068517031%2Cn%3A10068518031%2Cn%3A10387784031&qid=1583929313&ref=sr_pg_1. [Accessed: 11-Mar-2020].

[50] A. Palanica, P. Flaschner, A. Thommandram, M. Li, and Y. Fossat, "Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey," *J. Med. Internet Res.*, vol. 21, no. 4, pp. 1–10, 2019, doi: 10.2196/12887.

[51] Babylon, "Babylon Health." [Online]. Available: https://www.babylonhealth.com/us.

[52] F. Davidoff, "Time," *Ann. Intern. Med.*, vol. 127, no. 6, pp. 483–485, Sep. 1997, doi: 10.7326/0003-4819-127-6-199709150-00011.

[53] G. M. Lucas *et al.*, "Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers," *Front. Robot. AI*, 2017, doi: 10.3389/frobt.2017.00051.

[54] G. M. Lucas, J. Gratch, A. King, and L. P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Comput. Human Behav.*, 2014, doi: 10.1016/j.chb.2014.04.043.

[55] R. J. Moore, "Studies in Conversational UX Design," no. November, pp. 181–204, 2018, doi: 10.1007/978-3-319-95579-7.

[56] "A virtual assistant to help doctors in their daily work." [Online]. Available: https://www.safeinbreastfeeding.com/safedrugbot-chatbot-medical-assistant/.

[57] World Health Organization, "Depression and Other Common Mental Disorders," 2017. [Online].

Available: https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/.

[58] World Health Organization, "Depression and Other Common Mental Disorders," 2017. .

[59] R. Lwears, "Rethinking healthcare as a safety-critical industry," in *Work*, 2012, doi: 10.3233/WOR-2012-0037-4560.

[60] D. M. Gaba, "Structural and organizational issues in patient safety: A comparison of health care to other high-hazard industries," *Calif. Manage. Rev.*, 2000, doi: 10.2307/41166067.

[61] J. Corrigan, "Crossing the quality chasm," in *Building a Better Delivery System: A New Engineering/Health Care Partnership*, 2005.

[62] D. M. Gaba, "Anaesthesiology as a model for patient safety in health care," *British Medical Journal*. 2000, doi: 10.1136/bmj.320.7237.785.

[63] C. Macrae and C. Vincent, "Learning from failure: the need for independent safety investigation in healthcare," *J. R. Soc. Med.*, 2014, doi: 10.1177/0141076814555939.

[64] B. Sexton, E. Thomas, and R. L. Helmreich, "Error, stress, and teamwork in medicine and aviation: Cross sectional surveys," *Ugeskr. Laeger*, 2000, doi: 10.1136/bmj.320.7237.745.

[65] F. T. Durso and F. A. Drews, "Health care, aviation, and ecosystems: A socio-natural systems perspective," *Curr. Dir. Psychol. Sci.*, 2010, doi: 10.1177/0963721410364728.

[66] N. Kapur, A. Parand, T. Soukup, T. Reader, and N. Sevdalis, "Aviation and healthcare: a comparative review with implications for patient safety," *JRSM Open*, 2016, doi: 10.1177/2054270415616548.

[67] A. for the A. of M. Instrumentation, "Risk and Reliability in Healthcare and Nuclear Power: Learning from Each Other." Arlington, VA: Association for the Ad-vancement of Medical Instrumentation, 2013.

[68] G. Grote, "Safety management in different high-risk domains - All the same?," *Saf. Sci.*, 2012, doi: 10.1016/j.ssci.2011.07.017.

[69] C. L. Bosk, M. Dixon-Woods, C. A. Goeschel, and P. J. Pronovost, "Reality check for checklists.," *Lancet*. 2009, doi: 10.1016/S0140-6736(09)61440-9.

[70] S. N. Goldhaber-Fiebert and C. Macrae, "Emergency Manuals: How Quality Improvement and Implementation Science Can Enable Better Perioperative Management During Crises," *Anesthesiology Clinics*. 2018, doi: 10.1016/j.anclin.2017.10.003.

[71] E. G. Liberati, M. F. Peerally, and M. Dixon-Woods, "Learning from high risk industries may not be straightforward: A qualitative study of the hierarchy of risk controls approach in healthcare," *Int. J. Qual. Heal. Care*, 2018, doi: 10.1093/intqhc/mzx163.

[72] World Health Organization, "Patient Safety." [Online]. Available: https://www.who.int/patientsafety/en/.

[73] NHS, "The NHS Patient Safety Strategy," 2019. [Online]. Available: https://improvement.nhs.uk/documents/5472/190708_Patient_Safety_Strategy_for_website_v4.pdf.

[74] NHS, "NRLS national patient safety incident reports: commentary," 2019. [Online]. Available: https://improvement.nhs.uk/documents/6002/NAPSIR_commentary_Sept_2019_FINAL.pdf.

[75] J.-E. Bibault, B. Chaix, P. Nectoux, A. Pienkowski, A. Guillemasé, and B. Brouard, "Healthcare ex Machina: Are conversational agents ready for prime time in oncology?," *Clin. Transl. Radiat. Oncol.*, vol. 16, pp. 55–59, 2019, doi: 10.1016/j.ctro.2019.04.002.

[76] G. Cameron *et al.*, "Best practices for designing chatbots in mental healthcare – A case study on iHelpr," *Hci 2018*, pp. 1–5, 2017, doi: 10.14236/ewic/hci2018.129.

[77] C. Y. Huang, M. C. Yang, Y. J. Chen, M. L. Wu, and K. W. Chen, “A Chatbot-supported Smart Wireless Interactive Healthcare System for Weight Control and Health Promotion,” in *IEEE International Conference on Industrial Engineering and Engineering Management*, 2019, doi: 10.1109/IEEM.2018.8607399.

[78] A. S. Lokman and J. M. Zain, “An architectural design of virtual dietitian (ViDi) for diabetic patients,” *Proc. - 2009 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol. ICCSIT 2009*, no. ViDi, pp. 408–411, 2009, doi: 10.1109/ICCSIT.2009.5234671.

[79] N. S. Ahmad, M. H. Sanusi, M. H. Abd Wahab, A. Mustapha, Z. A. Sayadi, and M. Z. Saringat, “Conversational bot for pharmacy: A natural language approach,” *2018 IEEE Conf. Open Syst. ICOS 2018*, pp. 76–79, 2019, doi: 10.1109/ICOS.2018.8632700.

[80] H. Fraser, E. Coiera, and D. Wong, “Safety of patient-facing digital symptom checkers,” *Lancet*, vol. 392, no. 10161, pp. 2263–2264, 2018, doi: 10.1016/S0140-6736(18)32819-8.

[81] E. Levin and A. Levin, “Evaluation of spoken dialogue technology for real-time health data collection,” *J. Med. Internet Res.*, vol. 8, no. 4, pp. 1–17, 2006, doi: 10.2196/jmir.8.4.e30.

[82] R. J. Moore, “Studies in Conversational UX Design,” no. November, pp. 181–204, 2018, doi: 10.1007/978-3-319-95579-7.

[83] A. S. Miner, A. Milstein, S. Schueller, R. Hegde, C. Mangurian, and E. Linos, “Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health,” *JAMA Intern. Med.*, vol. 176, no. 5, pp. 619–625, 2016, doi: 10.1001/jamainternmed.2016.0400.

[84] A. Van Heerden, X. Ntinga, and K. Vilakazi, “The potential of conversational agents to provide a rapid HIV counseling and testing services,” in *Conference Proceedings - 2017 International Conference on the Frontiers and Advances in Data Science, FADS 2017*, 2017, doi: 10.1109/FADS.2017.8253198.

[85] L. Cullen, “The public inquiry into the Piper Alpha disaster,” *Drill. Contract. (United States)*, 1993.

[86] A. Britain, G., & Hidden, “Investigation into the Clapham Junction railway accident,” 1989.

[87] HSE, “Guide to the Offshore Installations (Safety Case) Regulations 1992: Guidance on Regulations L30,” 1992.

[88] U.K. Health and Safety Executive, “Railway Safety Cases - Railway (Safety Case) Regulations 1994 - Guidance on Regulations,” 1994.

[89] T. P. Kelly, “A Systematic Approach to Safety Case Management,” *SAE Int.*, 2003, doi: 10.1191/1460408605ta336oa.

[90] R. Hawkins, I. Habli, T. Kelly, and J. McDermid, “Assurance cases and prescriptive software safety certification: A comparative study,” *Saf. Sci.*, vol. 59, pp. 55–71, 2013, doi: 10.1016/j.ssci.2013.04.007.

[91] Adelard, “Claims, Arguments and Evidence (CAE).” [Online]. Available: https://www.adelard.com/asce/choosing-asce/cae.html.

[92] U. S. Food and D. Administration, “Infusion pumps total product life cycle: Guidance for industry and fda staff,” *Food Drug Adm. Stand.*, pp. 766–910, 2014.

[93] EU, “Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EE,” 2017.

[94] ISO, “ISO 14971:2019 Medical devices — Application of risk management to medical devices,” 2019.

[95] ISO, “ISO 13485:2016 Medical devices — Quality management systems — Requirements for

regulatory purposes."

[96] C. B. Weinstock and J. B. Goodenough, "Towards an assurance case practice for medical devices," CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2009.

[97] NHS Digital, "DCB0129: Clinical Risk Management: its Application in the Manufacture of Health IT Systems," 2018.

[98] NHS Digital, "DCB0160: Clinical Risk Management: its Application in the Deployment and Use of Health IT Systems."

[99] A. B. Haynes, W. R. Berry, and A. A. Gawande, "What Do We Know about the Safe Surgery Checklist Now?," *Annals of Surgery*. 2015, doi: 10.1097/SLA.0000000000001144.

[100] P. J. Pronovost *et al.*, "Sustaining reductions in catheter related bloodstream infections in Michigan intensive care units: Observational study," *BMJ*, 2010, doi: 10.1136/bmj.c309.

[101] K. Catchpole and S. Russ, "The problem with checklists," *BMJ Qual. Saf.*, 2015, doi: 10.1136/bmjqs-2015-004431.

[102] R. Clay-Williams and L. Colligan, "Back to basics: Checklists in aviation and healthcare," *BMJ Quality and Safety*. 2015, doi: 10.1136/bmjqs-2015-003957.

[103] J. R. Ward, P. J. Clarkson, P. Buckle, J. Berman, R. Lim, and G. T. Jun, "Prospective hazard analysis: tailoring prospective methods to a healthcare context," 2010.

[104] J. DeRosier, E. Stalhandske, J. P. Bagian, and T. Nudell, "Using health care Failure Mode and Effect Analysis: the VA National Center for Patient Safety's prospective risk analysis system.," *Jt. Comm. J. Qual. Improv.*, 2002, doi: 10.1016/s1070-3241(02)28025-6.

[105] M. M. P. Habraken, T. W. Van der Schaaf, I. P. Leistikow, and P. M. J. Reijnders-Thijssen, "Prospective risk analysis of health care processes: A systematic evaluation of the use of HFMEA^{TM} in Dutch health care," *Ergonomics*, 2009, doi: 10.1080/00140130802578563.

[106] B. D. Franklin, N. A. Shebl, and N. Barber, "Failure mode and effects analysis: Too little for too much?," *BMJ Quality and Safety*. 2012, doi: 10.1136/bmjqs-2011-000723.

[107] S. K. Sarker, A. Chang, T. Albrani, and C. Vincent, "Constructing hierarchical task analysis in surgery," *Surg. Endosc. Other Interv. Tech.*, 2008, doi: 10.1007/s00464-007-9380-z.

[108] R. Lane, N. A. Stanton, and D. Harrison, "Applying hierarchical task analysis to medication administration errors," *Appl. Ergon.*, 2006, doi: 10.1016/j.apergo.2005.08.001.

[109] J. Weizenbaum, "ELIZA-A computer program for the study of natural language communication between man and machine," *Commun. ACM*, 1966, doi: 10.1145/365153.365168.

[110] K. M. Colby, S. Weber, and F. D. Hilf, "Artificial Paranoia," *Artif. Intell.*, 1971, doi: 10.1016/0004-3702(71)90002-6.

[111] K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer, "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes," *Artif. Intell.*, 1972, doi: 10.1016/0004-3702(72)90049-5.

[112] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems: The journal version," *Dialogue and Discourse*, 2018, doi: 10.5087/dad.2018.101.

[113] A. Ritter, C. Cherry, and B. Dolan, "Unsupervised modeling of twitter conversations," in *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 2010.

[114] J. Cahn, "CHATBOT: Architecture, Design, and Development," p. 46, 2017.

[115] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "GUS,

a frame-driven dialog system," *Artif. Intell.*, 1977, doi: 10.1016/0004-3702(77)90018-2.

[116] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The Dialog State Tracking Challenge: A Review," *Dialogue & Discourse*, 2016.

[117] R. E. Gruhn *et al.*, "Statistical Pronunciation Modeling for Non-Native Speech Processing," *Statew. Agric. L. Use Baseline 2015*, 2011, doi: 10.1007/978-3-642-19586-0.

[118] S. Young *et al.*, "The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management," *Comput. Speech Lang.*, 2010, doi: 10.1016/j.csl.2009.04.001.

[119] N. Mrkšic, D. Séaghdha, T. H. Wen, B. Thomson, and S. Young, "Neural belief tracker: Data-driven dialogue state tracking," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, doi: 10.18653/v1/P17-1163.

[120] Robert Challen, "Emerging Safety Issues in Artificial Intelligence," no. July, pp. 1–8, 2019.

[121] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Qual. Saf.*, vol. 28, no. 3, pp. 231–237, 2019, doi: 10.1136/bmjqs-2018-008370.

[122] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, 2017, doi: 10.1038/nature21056.

[123] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[124] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Hum. Factors*, 2010, doi: 10.1177/0018720810376055.

[125] N. Prasad, L. F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," in *Uncertainty in Artificial Intelligence - Proceedings of the 33rd Conference, UAI 2017*, 2017.

[126] R. Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components," *IEEE Signal Processing Magazine*. 2017, doi: 10.1109/MSP.2016.2617341.

[127] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton, "Repairing conversational misunderstandings and non-understandings," *Speech Commun.*, 1994, doi: 10.1016/0167-6393(94)90073-6.

[128] O. Sun, J. Chen, and F. Magrabi, "Using Voice-Activated Conversational Interfaces for Reporting Patient Safety Incidents: A Technical Feasibility and Pilot Usability Study.," *Stud. Health Technol. Inform.*, vol. 252, pp. 139–144, 2018, doi: 10.3233/978-1-61499-890-7-139.

[129] M. Sujan *et al.*, "Human factors challenges for the safe use of artificial intelligence in patient care," *BMJ Heal. Care Informatics*, vol. 26, no. 1, pp. 1–5, 2019, doi: 10.1136/bmjhci-2019-100081.

[130] I. L. Singh, R. Molloy, and R. Parasuraman, "Automation-Induced 'Complacency': Development of the Complacency-Potential Rating Scale," *Int. J. Aviat. Psychol.*, 1993, doi: 10.1207/s15327108ijap0302_2.

[131] M. L. Cummings, "Automation bias in intelligent time critical decision support systems," in *Collection of Technical Papers - AIAA 1st Intelligent Systems Technical Conference*, 2004, doi: 10.2514/6.2004-6313.

[132] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors*, 1997, doi: 10.1518/001872097778543886.

[133] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*.

2012, doi: 10.1136/amiajnl-2011-000089.

[134] K. Marten *et al.*, “Computer-assisted detection of pulmonary nodules: Performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologist,” *Eur. Radiol.*, 2004, doi: 10.1007/s00330-004-2389-y.

[135] N. B. Sarter and B. Schroeder, “Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing,” *Hum. Factors*, 2001, doi: 10.1518/001872001775870403.

[136] N. Petrick *et al.*, “CT colonography with computer-aided detection as a second reader: Observer performance study,” *Radiology*, 2008, doi: 10.1148/radiol.2453062161.

[137] S. Dreiseitl and M. Binder, “Do physicians value decision support? A look at the effect of decision support systems on physician opinion,” *Artif. Intell. Med.*, 2005, doi: 10.1016/j.artmed.2004.07.007.

[138] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, “The role of trust in automation reliance,” *Int. J. Hum. Comput. Stud.*, 2003, doi: 10.1016/S1071-5819(03)00038-7.

[139] D. P. Biros, M. Daly, and G. Gunsch, “The influence of task load and automation trust on deception detection,” *Gr. Decis. Negot.*, 2004, doi: 10.1023/B:GRUP.0000021840.85686.57.

[140] M. D. Burdick, L. J. Skitka, K. L. Mosier, and S. Heers, “Ameliorating effects of accountability on automation bias,” in *Proceedings of the Annual Symposium on Human Interaction with Complex Systems, HICS*, 1996, doi: 10.1109/huics.1996.549504.

[141] F. Cabitza, R. Rasoini, and G. F. Gensini, “Unintended consequences of machine learning in medicine,” *JAMA - Journal of the American Medical Association*. 2017, doi: 10.1001/jama.2017.7797.

[142] C. Thornton, C. Braun, C. Bowers, and B. B. Morgan Jr, “Automation effects in the cockpit: A low-fidelity investigation,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1992, vol. 36, no. 1, pp. 30–34.

[143] E. L. Wiener, “Life in the second decade of the glass cockpit,” in *International Symposium on Aviation Psychology, 7 th, Columbus, OH*, 1993, pp. 1–7.

[144] C. Gold, D. Damböck, L. Lorenz, and K. Bengler, “Take over! How long does it take to get the driver back into the loop?,” in *Proceedings of the Human Factors and Ergonomics Society*, 2013, doi: 10.1177/1541931213571433.

[145] V. A. Banks, K. L. Plant, and N. A. Stanton, “Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016,” *Saf. Sci.*, 2018, doi: 10.1016/j.ssci.2017.12.023.

[146] M. A. Sujan *et al.*, “Emergency Care Handover (ECHO study) across care boundaries: The need for joint decision making and consideration of psychosocial history,” *Emerg. Med. J.*, 2015, doi: 10.1136/emermed-2013-202977.

[147] P. Spurgeon, M.-A. Sujan, S. Cross, and H. Flanagan, “A Systems Approach to Improving Clinical Handover in Emergency Care,” in *Building Safer Healthcare Systems*, Springer, 2019, pp. 125–135.

[148] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” *Hum. Factors*, vol. 37, no. 1, pp. 32–64, 1995, doi: 10.1518/001872095779049543.

[149] M. R. Endsley, *Designing for situation awareness: An approach to user-centered design*. CRC press, 2016.

[150] N. White, “Understanding the role of non-technical skills in patient safety.,” *Nurs. Stand.*, 2012, doi: 10.7748/ns2012.02.26.26.43.c8972.

[151] M. M. Kokar, C. J. Matheus, and K. Baclawski, “Ontology-based situation awareness,” *Inf.*

Fusion, 2009, doi: 10.1016/j.inffus.2007.01.004.

[152] L. Paget *et al.*, “Patient-Clinician Communication: Basic Principles and Expectations,” *NAM Perspect.*, 2011, doi: 10.31478/201106a.

[153] T. E. Flickinger *et al.*, “Clinician empathy is associated with differences in patient-clinician communication behaviors and higher medication self-efficacy in HIV care,” *Patient Educ. Couns.*, 2016, doi: 10.1016/j.pec.2015.09.001.

[154] S. S. Kim, S. Kaplowitz, and M. V. Johnston, “The effects of physician empathy on patient satisfaction and compliance,” *Eval. Heal. Prof.*, 2004, doi: 10.1177/0163278704267037.

[155] A. S. Lokman, J. M. Zain, F. S. Komputer, and K. Perisian, “Designing a Chatbot for diabetic patients,” *Int. Conf. Softw. Eng. Comput. Syst.*, no. August, pp. 19–21, 2009, doi: 10.1080/09273948.2016.1178303.

[156] A. Fadhil, “A Conversational Interface to Improve Medication Adherence: Towards AI Support in Patient’s Treatment,” 2018.

[157] G. Neff and P. Nagy, “Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay,” *Int. J. Commun.*, 2016.

[158] R. Crist, “Amazon and Google are listening to your voice recordings. Here’s what we know about that,” 2019. [Online]. Available: https://www.cnet.com/how-to/amazon-and-google-are-listening-to-your-voice-recordings-heres-what-we-know/.

[159] K. Bock and S. M. Garnsey, “Language Processing,” *A Companion to Cogn. Sci.*, pp. 226–234, 2008, doi: 10.1002/9781405164535.ch14.

[160] M. MARCUS, B. SANTORINI, and M. MARCINKIEWICZ, “Building a Large Annotated Corpus of English: The Penn Treebank,” *Comput. Linguist.*, 1993.

[161] J. Nivre *et al.*, “Universal dependencies v1: A multilingual treebank collection,” in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016.