Chatbots are one class of intelligent, conversational software agents activated by natural language input (which can be in the form of text, voice, or both). They provide conversational output in response and, if commanded, can sometimes also execute tasks. Although chatbot technologies have existed since the 1960s and have influenced user interface development in games since the early 1980s, chatbots are now easier to train and implement. This is due to plentiful open source code, widely available development platforms, and implementation options via software as a service (SaaS). In addition to enhancing customer experiences and supporting learning, chatbots can also be used to engineer social harm—that is, to spread rumors and misinformation, or attack people for posting their thoughts and opinions online. This paper presents a literature review of quality issues and attributes as they relate to the contemporary issue of chatbot development and implementation. Finally, quality assessment approaches are reviewed, and a quality assessment method based on these attributes and the analytic hierarchy process is proposed and examined.

Evaluating Quality of Chatbots and Intelligent Conversational Agents

NICOLE RADZIWILL AND MORGAN BENTON

INTRODUCTION

"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it." – Weiser (1991)

Whether users realize it or not, the "people" they interact with online are not all people. Customer service chat and commercial social media interactions are increasingly managed by intelligent agents, many of which have been developed with human identities and even personalities (Simonite 2017). Even though the technology itself is not new, reliable linguistic functionality, availability through software as a service (SaaS), and the addition of intelligence through machine learning have increased its popularity. Between 2007 and 2015, chatbots participated in a third to one half of all online interactions (Tsvetkova et al. 2016), and the rate at which new chatbots are being deployed has increased since then.

Social, conversational bots can be used to provide benefits to companies, which use them to reduce time to response, provide enhanced customer service, increase satisfaction, and increase engagement. Unfortunately, some chatbots are specifically designed to be harmful. For example, networks of fake users (called "sybils" on Twitter) have been implemented to artificially inflate "follower" counts to increase social status for users who purchase them, to spread fake news or rumors, and even to intimidate users who express certain political beliefs (Ferrara

et al. 2016). In the 2016 U.S. presidential election, up to a fifth of the comments and responses on Twitter were driven by fully or partially autonomous Twitter accounts (Rainie, Anderson, and Albright 2017)

Due to their flexibility and ease of use (compared to web-based or mobile applications), some have speculated that conversational agents may be a universal user interface and may replace "apps"—a democratization of the versatile command line (Solomon 2017). In addition, chatbots and conversational agents are anticipated to be important interfaces in virtual reality environments (Beilby and Zakos 2014). As a result, it is important to understand the issues and quality attributes associated with developing and implementing high-quality chatbots and conversational agents, and identify a mechanism for quality assurance across these factors.

The rest of this paper is organized as follows. First, the authors present the historical context of chatbots and the background for chatbots and conversational agents, which are two types of dialog systems. Next, they outline the methodology for their literature review along two topics: 1) quality attributes for chatbots and conversational agents; and 2) quality assessment approaches. Finally, the authors synthesize these results and present an approach for evaluating the quality of these technologies in terms of key quality attributes.

BACKGROUND

Chatbot systems originated with programs like ELIZA (Weizenbaum 1966), which were intended to demonstrate natural language conversation with a computer. An early-stated goal of such systems was to pass the Turing test (Turing 1950) in which a human interrogator deems a computer sufficiently "intelligent" to pass as a human. However, primitive systems like ELIZA used key-word matching and minimal context identification, and lacked the ability to keep a conversation going. As shown in Figure 1, through interactions with the program, it was easy to guess that ELIZA was a computer.

Researchers continued to develop demonstration systems with natural language capabilities, but none were capable of passing the Turing test. In the early 1980s, ALICE (Wallace 2003) was created, becoming significant not for its conversational capabilities but because it led to the development of artificial intelligence markup language (AIML). AIML is used to declare pattern-matching rules that link user-submitted words and phrases with topic categories. It is extensible markup language (XML)

based, and supports most chatbot platforms and services in use today.

Chatbots receive natural language input, sometimes interpreted through speech recognition software, and execute one or more related commands to engage in goal-directed behavior (often on behalf of a human user). As intelligent agents, they are usually autonomous, reactive, proactive, and social. The most advanced systems employ machine learning (often Markov chains or deep neural networks) so they may also adapt to new information or new requests.

Chatbots are one category of conversational agents, which are software systems that mimic interactions with real people. They are typically not embodied in the forms of animals, avatars, humans, or humanoid robots (those programs are considered to be "embodied conversational agents"). Conversational agents, a class of dialog systems, have been a subject of research in communications for decades. Interactive voice response systems (for example, "Press or say 1 for English") are also dialog systems, but they are not usually considered conversational agents since they implement decision trees (McTear, Callejas, and Griol 2016). These terms are related to each other in Figure 2.
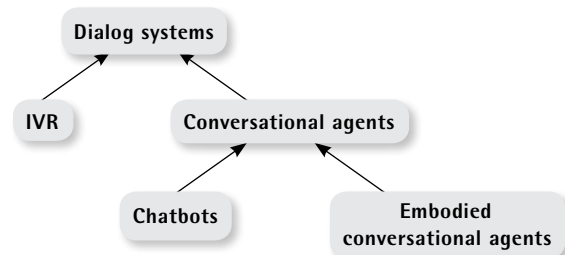
FIGURE 1 A sample dialog with ELIZA (http://www.masswerk.at/elizabot/eliza.html)

©2017, ASQ

FIGURE 2 Relationships between classes of software-based dialog systems

©2017, ASQ

The emphasis in this paper is on the text-based conversational agents found online and in some internet of things (IoT) devices, which are sometimes (but not often) embodied. This contrasts with voice-activated conversational agents like Siri, Google Now, Cortana, and Samsung S Voice, which are not considered chatbots. Although the earliest software in this genre appeared in internet relay chat environments in the mid-1990s, bots have evolved to serve multiple purposes, such as content editing and brokering complex transactions (Tsvetkova et al. 2016). These programs serve a range of roles, from personal assistant, to intelligent virtual agent, to companion. There are agents designed to serve as personal university advisors (Ghose and Barua 2013), educational agents developed to help improve learning outcomes (Kerly, Hall, and Bull 2007), and "Art-Bots" (Vassos et al. 2016) to engage museum visitors in participatory installations.

Chatbots are distinct from *bots*, which are compromised computers that often run malicious software and can be linked together as *botnets* to coordinate large-scale denial of service attacks (Thing, Sloman, and Dulay 2007). However, chatbots can be launched from botnets to shape social perceptions. No one who spends time online is immune from the potential harm of chatbots. Even the director of the annual Loebner prize competition in artificial intelligence, an event that pits the most sophisticated chatbots against one another, was fooled into thinking a chatbot on a dating service was interested in him romantically (Epstein 2007; Metz 2017).

Researchers are actively investigating ways to mitigate the harm from chatbots that engage in social engineering. Alarifi, Alsaleh, and Al-Salman (2016), for example, built a corpus of these "sybils" and a browser plug-in to help human users better distinguish between humans and machines. The harm derives from the ease with which these chatbots, especially on social networks like Twitter, pass the Turing test and convince human participants that their criticisms, abuse, and even rape and death threats are real and originate with other humans (McElrath 2017). The potential for social engineering emphasizes the need to critically examine quality attributes for chatbots, in part to protect the well-being of individuals and societies.

Development and implementation of chatbots today is easier, and chatbots themselves are more powerful. Development platforms, some of which implement SaaS approaches (for example, Pandorabots, Chatfuel,

Botsify), serve to split the testing responsibility between service provider (which is responsible for testing inputs, execution of actions, and realistic outputs) and client (who evaluates ease of use and effectiveness of task accomplishment). Because of easy integration with social media and developer productivity tools (for example, Slack, GitHub), chatbots may even be instrumental in improving the work processes in traditional and agile development teams (Storey and Zagalsky 2016).

Machine learning approaches are also increasingly integrated to make these agents more adaptive to different input styles and new tasks. Systems are no longer dependent on deterministic responses from rules-based pattern matching like ELIZA and ALICE. More commonly, systems leverage supervised learning (which requires large training sets), unsupervised learning (like Markov-chain based models), and hybrid intelligence (where humans participate in the training process over time). Supervised learning and hybrid intelligence approaches are more extensive and costly, but they can result in systems that are better at on-the-fly problem solving and take less time to achieve goals (Wilson, Daugherty, and Morini-Bianzino 2017).

METHODOLOGY

The purpose of this research was to examine the academic and industry literature to: 1) provide a comprehensive review of quality attributes for chatbots and conversational agents; and 2) identify appropriate quality assurance approaches. To do this, publications on quality in chatbots and conversational agents were identified through systematic searches of Google Scholar, JSTOR, and EBSCO Host from 1990 to 2017. The search terms that were used were chatbots, conversational agents, embodied conversational agents, quality, quality attributes, quality of chatbots, quality of conversational agents, and quality assurance in various combinations. Papers were drawn from the domains of engineering, technology, psychology, communications, and anthropology.

The selection of articles was based on three criteria. First, scholarly references were emphasized, supplemented only by industry publications from 2016 and 2017. Publications were selected if they contained at least one of the search terms in the title and/or abstract to ensure the relevance of data collection. Only publications were selected that had quality as

a major or influential aspect of the research. Finally, highly technical articles focused on programming and engineering aspects of chatbots, including improving the quality of speech recognition, were excluded.

The initial search generated a sampling frame of 7,340 articles, which was further refined by adding search terms for "evaluation," "assessment," "quality metrics," and "metrics." Most recent articles (from 2016 and 2017) were inspected next, followed by articles between 2013 and 2015, and then from 2007 to 2012. The selection of timeframes was made to limit the number of articles for review to less than 300, and, ultimately, 36 scholarly articles and conference papers were determined to be relevant to the objectives of this paper. They were supplemented with 10 articles from industry or trade magazines. Beyond this, only seven articles were identified for the second portion of the study focusing on quality assurance, and all were used in the analysis.

QUALITY ATTRIBUTES

The authors extracted quality attributes from each of the 32 papers and 10 articles, and grouped them based on similarity. After two or three iterations, they noticed that, in general, they were aligned with the ISO 9241 concept of usability: "The effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments" (Abran et al. 2003). In particular, effectiveness refers to the accuracy and completeness with which specified users achieve their goals, and efficiency refers to how well resources are applied to achieve those goals. Not surprisingly, particularly in a SaaS environment, the burden of demonstrating efficiency and effectiveness falls more on the service provider (since this will rely on common functionality), while the need to ensure customers are satisfied remains with the implementer. Table 1 outlines the quality attributes organized in terms of ISO 9241.

TABLE 1 Quality attributes of chatbots and conversational agents

EFFICIENCY		
Category	**Quality Attribute**	**Reference**
Performance	• Graceful degradation	• Cohen and Lane (2016)
	• Robustness to manipulation	• Thieltges, Schmidt, and Higelich (2016)
	• Robustness to unexpected input	• Kluwer (2011)
	• Avoid inappropriate utterances and be able to perform damage control	• Morrissey and Kirakowski (2013)
	• Effective function allocation, provides appropriate escalation channels to humans	• Staven (2017)
EFFECTIVENESS		
Category	**Quality Attribute**	**Reference**
Functionality	• Accurate speech synthesis	• Kuligowska (2015)
	• Interprets commands accurately	• Eeuwen (2017)
	• Use appropriate degrees of formality, linguistic register	• Morrissey and Kirakowski (2013)
	• Linguistic accuracy of outputs	• Wallace (2003)
	• Execute requested tasks	• Ramos (2017)
	• Facilitate transactions and follows up with status reports	• Eeuwen (2017)
	• General ease of use	
	• Engage in on-the-fly problem solving	• Solomon (2017)
	• Contains breadth of knowledge, is flexible in interpreting it	• Cohen and Lane (2016)
Humanity	• Passes the Turing test	• Weizenbaum (1966); Wallace (2003)
	• Does not have to pass the Turing test	• Ramos (2017)
	• Transparent to inspection, discloses its chatbot identity	• Bostrom and Yudkowski (2014)
	• Include errors to increase realism	• Coniam (2014)
	• Convincing, satisfying, and natural interaction	• Morrissey and Kirakowski (2013)
	• Able to respond to specific questions	
	• Able to maintain themed discussion	

TABLE 1   Quality attributes of chatbots and conversational agents (continued)

SATISFACTION		
Category	**Quality Attribute**	**Reference**
Affect	• Provide greetings, convey personality	• Morrissey and Kirakowski (2013)
	• Give conversational cues	
	• Provide emotional information through tone, inflection, and expressivity	• Pauletto et al. (2013)
	• Exude warmth and authenticity	• Solomon (2017)
	• Make tasks more fun and interesting	• Eeuwen (2017)
	• Entertain and/or enable participant to enjoy the interaction	• Ramos (2017)
	• Read and respond to moods of human participant	• Meira and Canuto (2015)
Ethics and Behavior	• Respect, inclusion, and preservation of dignity (linked to choice of training set)	• Neff and Nagy (2016)
	• Ethics and cultural knowledge of users	• Applin and Fischer (2015)
	• Protect and respect privacy	• Eeuwen (2017)
	• Nondeception	• Isaac and Bridewell (2014)
	• Sensitivity to safety and social concerns	• Miner et al. (2016)
	• Trustworthiness (linked to perceived quality)	• Herzum et al. (2002)
	• Awareness of trends and social context	• Vetter (2002)
Accessibility	• Responds to social cues or lack thereof	• Morrissey and Kirakowski (2013)
	• Can detect meaning or intent	• Wilson et al. (2017)
	• Meets neurodiverse needs such as extra response time and text interface	• Radziwill and Benton (2017)

©2017, ASQ

There was remarkable consistency across the source material regarding quality attributes; although several of them were repeated across sources, only the primary source (or the source where that attribute was emphasized) appears in Table 1. However, one attribute was in question: whether a chatbot or conversational agent should pass the Turing test. According to most researchers, and following the earliest conversational interfaces, responding and interacting like a human should be the top priority. In fact, this principle guided development for nearly four decades since ELIZA came online.

According to Ramos (2017), who references the Facebook chatbot (named Joy) in addition to a character from a 2014 Disney movie, argues that giving the impression of being human is not a valid quality attribute. "When interacting with a fictional character," he claims, "people are willing to have a suspension of disbelief and enjoy the interaction. In our user testing, I've seen people interact with Joy. When they get a response with an emoji, I've seen the smiles. People know it's not real and don't care that it wouldn't pass a Turing test — especially when Joy saves them money. Just think of the character Baymax from the movie 'Big Hero Six': He doesn't sound or look human, but that doesn't prevent you from empathizing with him. Would Baymax pass a Turing test? No, but that's not what matters."

The position that a chatbot or conversational agent should *not* aim to act human is shared by analysts such as Wilson, Daugherty, and Morini-Bianzino (2017), who believe character development and management is one of the new jobs that will be enabled by the spread of artificial intelligence in automation. Taylor (2016) also argues that passing the Turing test is not a requirement. This incongruity illustrates the important point that the list of quality attributes identified through the literature is suggestive, not prescriptive. Each chatbot implementation will (necessarily) prioritize different quality attributes at different phases of the system's life cycle.

QUALITY ASSESSMENT

The second question raised in this research is how quality assurance can be performed for chatbots and conversational agents. To do this, the authors scanned the literature to find six references where other researchers address the question of quality assurance in this domain. Finally, they synthesized the approaches to

recommend a composite technique, based on the analytic hierarchy process (AHP) by Saaty (1990).

Previous Approaches

The literature review uncovered seven references that focused specifically on quality assurance for chatbots and conversational agents (Vetter 2002; Goh et al. 2007; Košir 2013; Coniam 2014; Kuligowska 2015; Meira and Canuto 2015; Kaleem et al. 2016). Although several other resources were considered, these were the only papers for which quality was the central, unifying theme. Each paper noted the lack of guidance for designing quality into these systems and evaluating the quality of these systems. While Goh called for standard metrics, Košir bemoaned the lack of them.

Each paper addresses a different aspect of quality assessment, from the effectiveness of responses to individuals' questions to the merits of a customized, goal-oriented approach. The main points from each paper are summarized in Table 2.

The need for a goal-driven approach that incorporates different users' subjective experiences with the chatbots is evident in each of these six results. Furthermore, many useful metrics are suggested, but there is no guidance to determine when each metric should be applied. Although an absolute scale for quality assessment is unlikely given the wide variety of chatbots

TABLE 2 Quality assessment approaches in previous studies

Authors	Emphasis	Conclusions and Recommendations
Vetter (2002)	Constructing test scripts to evaluate utility of conversations	Recommends using PARADISE method (Walker et al. 1997; Sanders and Scholtz 2002) to determine whether conversational constructs meet basic linguistic quality standards. Based on maximizing satisfaction and task success, and minimizing costs. Involves creating a performance function based on confusion matrices (counts of successful and failed communications) for each participant.
Goh et al. (2007)	Effectiveness of question answering	Precision, recall, and F1 could be metrics for how well questions are answered, but they fall short. New measures must take into consideration that utility of responses is subjective, different domains have different knowledge repositories, and information is always growing.
Košir (2013)	Implementation and assessment of one chatbot (thesis)	To overcome the problems outlined by Goh et al. (2007), use an iterative method and track subjective evaluations from multiple evaluators.
Coniam (2014)	Examining linguistic quality of chatbot responses	Literature review established that nearly all chatbots met baseline requirements for linguistic accuracy, "grammatical fit," and "meaning fit," suggesting that underlying frameworks and packages are fundamentally sound.
Kuligowska (2015)	Assessment and comparison of six commercial Polish chatbots	By interviews, identified 10 key quality attributes (visual look, speech synthesis, form of interface, basic knowledge, specialized knowledge, conversation abilities, response to unexpected situations, personality traits, personalization options, ability to comment/provide feedback), and evaluated each one on an ordinal scale (1 … 5) to determine an overall quality assessment (good, very good, excellent).
Meira and Canuto (2015)	Determine quality metrics specifically for embodied emotional agents, where affective characteristics dominate	Authors propose a three-level measurement framework (conceptual level goals, operational level goals, and quantitative level goals) that examine quality of architecture and affective quality. Metrics at the quantitative level include cohesion, coupling, size, and communications or services per module (for architecture), and cooperation, likeability, enjoyment, trust, naturalness, reduction of frustration, believability, and interestingness as metrics.
Kaleem et al. (2016)	Identify and test an assessment approach customized for each conversational agent	The weakness of existing frameworks is that they do not take into account that different conversational systems will have different goals. They adapt the goal-question-metric approach by Fenton and Pfleeger (1998) and suggest pre/post test scores, perception of learning, correct/incorrect responses, and time in system as metrics that could be used for a quality assessment.

and conversational agents (and the fact that some are embodied in avatars or humanoid robots), it should be possible to systematically compare the quality of two or more chatbots. If these metrics are linked to individual elements of the comprehensive categorized list of quality attributes presented in Table 1, then AHP can be used after a prioritization and down selection process to accomplish this comparison. The following section describes this process.

SYNTHESIZED APPROACH: ANALYTIC HIERARCHY PROCESS

AHP is a structured approach for navigating complex decision-making processes that involve both qualitative and quantitative considerations. To use this approach:

1. Create a hierarchy of quality attributes and select appropriate metrics to represent each attribute.

2. Construct pairwise comparisons between the quality attributes for one or more product options.

3. Create comparison matrices, and compute the first principal eigenvector of each one to assess relative and global priority.

4. Combine the priorities and compute inconsistency factors to determine which product option best satisfies the hierarchy of quality attributes. (Software is available to make the last two steps easier; code is provided in the Appendix at the end of the article that illustrates how to do this in practice.)

For chatbots and conversational agents, the product options might be two or more versions of the conversational system, or the "as-is" system and one or more "to-be" systems under development. Consider an example where an organization has made improvements to a chatbot system over the past year, and now it wants to see if the quality has improved. First, the organization should examine the quality attributes in Table 1 to see which ones are the most important, and select metrics (from Table 2, or from experience). They should have multiple users participate in sessions with chatbots, and record the selected metrics. An example is shown in Table 3.

AHP uses pairwise assessments between categories, and within categories. The first step is to set up the attribute hierarchy (see Figure 3 on the next page). Notice that the top level displays the goal, the next level shows the quality attribute categories, and the next level includes the nine quality attributes that were selected in the prioritization process. At the bottom, the two alternatives ("old" and "new") are shown. Because the hierarchical structure is central, this approach aligns well with the goal-oriented approach of Kaleem et al. (2016).

The first step is to make pairwise comparisons between the categories themselves. Typically in AHP, only the numbers 1, 3, 5, 7, and 9 are used. Next, create a matrix where each cell indicates how much more important the category in the row is as compared to the category in the column. Record the reciprocal of the value if the category

TABLE 3 Example data to collect for an AHP chatbot quality assessment

Category	Quality Attribute	Metric	Old	New
Performance	Robustness to unexpected input	% of successes	86–92%	91–93%
	Provides appropriate escalation channels	% of successes	80%	100%
Humanity	Transparent to inspection (known chatbot)	% of users who correctly classify	100%	100%
	Able to maintain themed discussion	0 (low) … 100 (high)	72 (Avg.) 8 (St. Dev.)	85 (Avg.) 12 (St. Dev.)
	Able to respond to specific questions	% of successes	68–82%	80–85%
Affect	Provides greetings, pleasant personality	0 (low) … 100 (high)	89 (Avg.) 3 (St. Dev.)	96 (Avg.) 3 (St. Dev.)
	Entertaining, engaging	0 (low) … 100 (high)	50 (Avg.) 21 (St. Dev.)	66 (Avg.) 4 (St. Dev.)
Accessibility	Can detect meaning and intent	% of successes	85–90%	82–86%
	Responds to social cues appropriately	% of successes	78%	77%

FIGURE 3 Hierarchical model of prioritized quality attributes

©2017, ASQ
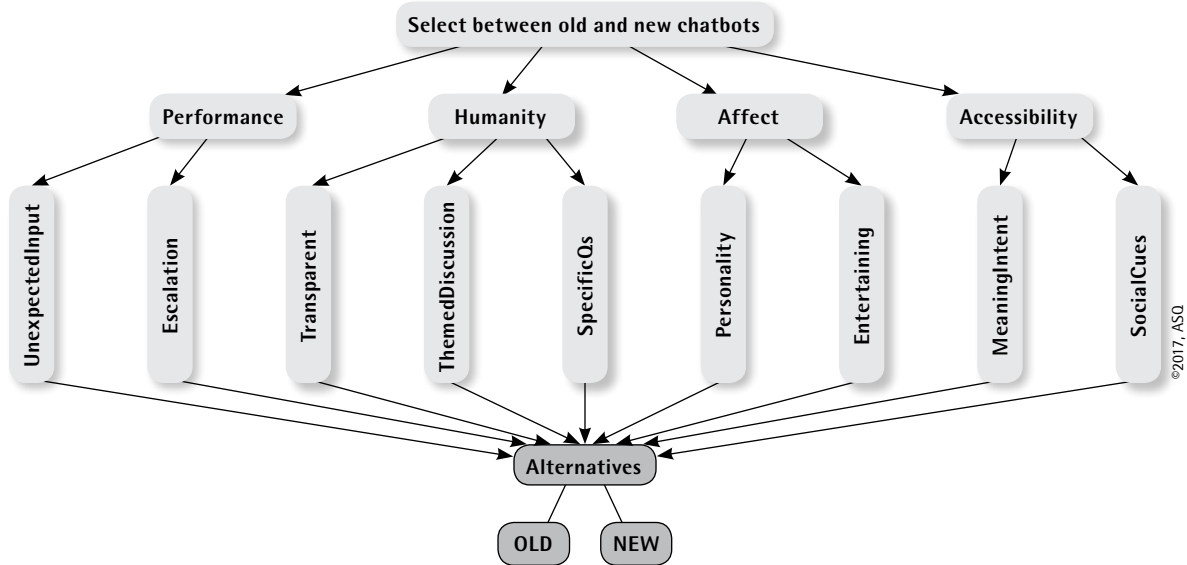
in the column is more important than the category in the row. This means the number 1 will be in the diagonal, since a category cannot be more or less important to itself. For example, in Figure 4, "performance" is much more important than "humanity" or "affect" (7), but it is slightly less important than "accessibility" (1/3). Values of 9 and 1/9 indicate a substantial difference between the importance of the two categories being assessed.

Within each category, create a priority matrix to express the relative merit of each quality attribute as compared to the others. The "humanity" attribute was selected for display in Figure 5 because it has the most attributes for this example. Notice how each measurement has a complementary cell: for example, "transparent" is less important than "themed discussion" (1/5), so it makes sense that themed discussion would be more important than transparent (5, the reciprocal).

There will be four priority matrices total at this level (one for each of the four top-level quality attribute categories).

Finally, use the measured values (in the "old" and "new" columns of Table 3) to compare how the different versions of the chatbot perform in terms of each quality attribute. There will be one comparison for each of the nine quality attributes the authors selected in the original prioritization. Figure 6 shows one of these nine lowest-level priority matrices, showing that the old chatbot is much less effective (1/7) in terms of the escalation attribute—and similarly, the new chatbot is much more effective (7).

When the first principal eigenvector of each of these matrices is determined and combined, a priority metric for each attribute and category is generated. Although the mechanics of the computations are beyond the scope of this article, instructions are provided in the Appendix for

FIGURE 4 Priority matrix for quality categories

	Performance	Humanity	Affect	Accessibility
Performance	1	7	7	1/3
Humanity	1/7	1	1/5	1/7
Affect	1/7	5	1	1/7
Accessibility	1/3	7	7	1

©2017, ASQ

FIGURE 5 Priority matrix for selected "humanity" quality attributes

	Transparent	ThemedDiscussion	SpecificQs
Transparent	1	1/5	1/5
ThemedDiscussion	5	1	1
SpecificQs	5	1	1

©2017, ASQ

FIGURE 6 Priority matrix for selected "escalation" quality attribute

	OLD	NEW
OLD	1	1/7
NEW	7	1

©2017, ASQ

readers to use the R Statistical Software (R Core Team 2016) to perform their own analysis.

When the computations are complete, the results will help an organization choose which alternative best satisfies its quality attributes, given that some attributes are more important and others are less important. Figure 7 shows the results for this example problem. Unfortunately, the authors' efforts have been in vain: the old chatbot is weighted much more heavily (66.2 percent compared to 33.8 percent) given the priorities the authors set between their categories and attributes.

The consistency scores should also be reasonable. Ideally, these values should be below 10 percent, but in practice numbers below 20 percent are often acceptable. High values in this column indicate some discrepancy in the individual assessments (for example, prioritizing all elements on one level of a hierarchy low with respect to each other). If there is a problem, inspect all assessments to make sure there are no data input errors.

FIGURE 7 AHP results for the example problem

	Weight	OLD	NEW	Consistency
Select between old and new chatbots	100.0%	66.2%	33.8%	❶18.4%
Accessibility	54.5%	39.1%	15.3%	0.0%
MeaningIntent	47.7%	35.7%	11.9%	0.0%
SocialCues	6.8%	3.4%	3.4%	0.0%
Performance	32.1%	24.6%	7.5%	0.0%
UnexpectedInput	28.1%	21.1%	7.0%	0.0%
Escalation	4.0%	3.5%	0.5%	0.0%
Affect	9.4%	1.6%	7.8%	0.0%
Entertaining	7.8%	1.3%	6.5%	0.0%
Personality	1.6%	0.3%	1.3%	0.0%
Humanity	4.1%	1.0%	3.1%	0.0%
SpecificQs	1.9%	0.3%	1.5%	0.0%
ThemedDiscussion	1.9%	0.5%	1.4%	0.0%
Transparent	0.4%	0.2%	0.2%	0.0%

CONCLUSIONS

This paper provided a review of the academic literature since 1990 and industry articles since 2015 to: 1) gather and articulate quality attributes for chatbots and conversational agents; and 2) discover and synthesize quality assurance approaches to recommend strategies moving forward. There are many ways for practitioners to apply the material in this article:

- The quality attributes (see Table 1) can be used as a checklist for a chatbot implementation team to make sure they have addressed key issues.

- Two or more conversational systems can be compared by selecting the most significant quality attributes.

- Systems can be compared at two points in time to see if quality has improved, which is particularly useful for adaptive systems that learn as they are exposed to additional participants and topics.

The example showed how this goal-oriented approach might be used to evaluate the quality of two different chatbot implementations. Because the method relies on pairwise comparisons, any metric (including those emphasized by the authors in Table 2) can be associated with each quality attribute, and the results will still be valid. Furthermore, this technique can be easily adapted to evaluate different implementations over time, which is essential since most conversational agents learn from experience with users. These factors make the AHP approach particularly robust for assessing the quality of chatbots and conversational agents, resolving the majority of issues identified by previous researchers.

REFERENCES

Abran, A., A. Khelifi, W. Suryn, and A. Seffah. 2003. Consolidating the ISO usability models. In *Proceedings of 11th international software quality management conference*, April, 23-25.

Alarifi, A., M. Alsaleh, and A. Al-Salman. 2016. Twitter Turing test: Identifying social machines. *Information Sciences* 372:332-346.

Applin, S. A., and M. D. Fischer. 2015. New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In *Proceedings of the 2015 IEEE International Symposium on Technology and Society* (ISTAS). New York, NY: IEEE.

Beilby, L. J., and J. Zakos. 2014. U.S. Patent No. 8,719,200. Washington, DC: U.S. Patent and Trademark Office.

Cohen, D., and I. Lane. 2016. An oral exam for measuring a dialog system's capabilities. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 835-841. Palo Alto, CA: AAAI Press.

Coniam, D. 2014. The linguistic accuracy of chatbots: Usability from an ESL perspective. *Text & Talk* 34, no. 5:545-567.

Eeuwen, M. 2017. Mobile conversational commerce: Messenger chatbots as the next interface between businesses and consumers. Master's thesis, University of Twente. Available at: http://essay.utwente.nl/71706/1/van%20Eeuwen_MA_BMS.pdf.

Epstein, R. 2007. From Russia, with love. *Scientific American Mind* 18, no. 5:16-17.

Fenton, N. E., and S. L. Pfleeger. 1998. *Software metrics: A rigorous and practical approach.* Brooks.

Ferrara, E., O. Varol, C. Davis, F. Menczer, and A. Flammini. 2016. The rise of social bots. *Communications of the ACM* 59, no. 7:96-104.

Ghose, S., and J. J. Barua. 2013. Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor. In *Proceedings of the 2013 International Conference on Informatics, Electronics & Vision* (ICIEV), 1-5. New York, NY: IEEE.

Goh, O. S., C. Ardil, W. Wong, and C. C. Fung. 2007. A black-box approach for response quality evaluation of conversational agent systems. *International Journal of Computational Intelligence* 3, no. 3:195-203.

Hertzum, M., H. H. Andersen, V. Andersen, and C. B. Hansen. 2002. Trust in information sources: Seeking information from people, documents, and virtual agents. *Interacting with computers* 14, no. 5:575-599.

Isaac, A. M., and W. Bridewell. 2014. Mindreading deception in dialog. *Cognitive Systems Research* 28:12-19.

Kaleem, M., O. Alobadi, J. O'Shea, and K. Crockett. 2016. Framework for the formulation of metrics for conversational agent evaluation. In RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme, May 28, 2016.

Kerly, A., P. Hall, and S. Bull. 2007. Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems* 20, no. 2:177-185.

Košir, D. 2013. Implementacija in testiranje klepetalnika. Doctoral diss., Univerza v Ljubljani.

Kuligowska, K. 2015. Commercial chatbot: Performance evaluation, usability metrics, and quality standards of embodied conversational agents. *Professionals Center for Business Research* 2 (February).

McElrath, L. 2017. Watching the hearings, I learned my "Bernie bro" harassers may have been Russian bots. Shareblue. Available at: http://shareblue.com/watching-the-hearings-i-learned-my-bernie-bro-harassers-may-have-been-russian-bots/.

McTear, M., Z. Callejas, and D. Griol. 2016. Conversational interfaces: Past and present. *The Conversational Interface.* New York, NY: Springer International Publishing.

Meira, M. O., and A. M. P. Canuto. 2015. Evaluation of emotional agents' architectures: An approach based on quality metrics and the influence of emotions on users. In *Proceedings of the World Congress on Engineering* 1.

Metz, R. 2017. Three weeks with a Chatbot and I've made a new friend. *MIT Technology Review.* Available at: https://www.technologyreview.com/s/603936/three-weeks-with-a-chatbot-and-ive-made-a-new-friend.

Miner, A. S., A. Milstein, S. Schueller, R. Hegde, C. Mangurian, and E. Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine* 176, no. 5:619-625.

Morrissey, K., and J. Kirakowski. 2013. Realness in chatbots: Establishing quantifiable criteria. In *International Conference on Human-Computer Interaction*, 87-96. Berlin, Germany: Springer Berlin Heidelberg.

Neff, G., and P. Nagy. 2016. Automation, algorithms, and politics| Talking to Bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10, no. 17.

Pauletto, S., B. Balentine, C. Pidcock, K. Jones, L. Bottaci, M. Aretoulaki, and J. Balentine. 2013. Exploring expressivity and emotion with artificial voice and speech technologies. *Logopedics Phoniatrics Vocology* 38, no. 3:115-125.

R Core Team. 2016. *R: A language and environment for statistical computing.* Vienna, Austria. R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Radziwill, N. M., and M. C. Benton. 2017. *Neurodiversity secrets for innovation and design.* Austin, TX: SXSW Interactive.

Rainie, L., J. Anderson, and J. Albright. 2017. The future of free speech, trolls, anonymity, and fake news online. *Pew Research.* Available at: http://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/.

Ramos, R. 2017. Screw the Turing test – Chatbots don't need to act human. *VentureBeat.* Available at: https://venturebeat.com/2017/02/03/screw-the-turing-test-chatbots-dont-need-to-act-human/.

Saaty, T. L. 1990. *Decision making for leaders: The analytic hierarchy process for decisions in a complex world.* Pittsburgh, PA: RWS Publications.

Sanders, G. A., and J. Scholtz. 2001. Measurement and evaluation of embodied conversational agents. In *Embodied Conversational Agents*, 346-373. Cambridge, MA: MIT Press.

Simonite, T. 2017. Customer service chatbots are about to become frighteningly realistic. *MIT Technology Review.* Available at: https://www.technologyreview.com/s/603895/customer-service-chatbots-are-about-to-become-frighteningly-realistic/.

Solomon, M. 2017. If Chatbots win, customers lose, says Zappos customer service expert. *Forbes.* Available at: https://www.forbes.com/sites/micahsolomon/2017/03/23/customers-lose-if-chatbots-win-says-zappos-customer-service-expert.

Staven, T. 2017. What makes a good bot (or not)? *Unit4 Newsletter.* Available at: *http://www.unit4.com/blog/2017/03/what-makes-a-good-bot-or-not.*

Storey, M. A., and A. Zagalsky. 2016. Disrupting developer productivity one bot at a time. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 928-931. New York, NY: ACM.

Taylor, S. J. 2016. Very human lessons from three brands that use chatbots to talk to customers. *Fast Company.* Available at: https://www.fastcompany.com/3064845/human-lessons-from-brands-using-chatbots.

Thieltges, A., F. Schmidt, and S. Hegelich. 2016. The devil's triangle: Ethical considerations on developing bot detection methods. In *2016 AAAI Spring Symposium Series.*

Thing, V. L., M. Sloman, and N. Dulay. 2007. A survey of bots used for distributed denial of service attacks. In *IFIP International Information Security Conference*, 229-240. New York, NY: Springer US.

Tsvetkova, M., R. García-Gavilanes, L. Floridi, and T. Yasseri. 2016. *Even Good Bots Fight.*

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59, no. 236:433-460.

Vassos, S., E. Malliaraki, F. D. Falco, J. Di Maggio, M. Massimetti, M. G. Nocentini, and A. Testa. 2016. Art-Bots: Toward chat-based conversational

experiences in museums. In *Proceedings of the Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling* (ICIDS), Los Angeles, CA, November 15–18, 433-437. New York, NY: Springer International Publishing.

Vetter, M. 2002. Quality aspects of bots. In *Software Quality and Software Testing in Internet Times*, 165-184. Berlin, Germany: Springer Berlin Heidelberg.

Walker, M. A., D. J. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 271-280. Stroudsburg, PA: Association for Computational Linguistics.

Wallace, R. 2003. *The elements of AIML style.* ALICE AI Foundation. Avilable at: http://www.alicebot.org/style.pdf.

Weiser, M. 1991. The computer for the twenty-first century. *Scientific American* (September): 94–110.

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9, no. 1:36-45.

Wilson, H. J., P. R. Daugherty, and N. Morini-Bianzino. 2017. Will AI create as many jobs as it eliminates? *MIT Sloan Management Review.* Available at: http://sloanreview.mit.edu/article/will-ai-create-as-many-jobs-as-it-eliminates/.

BIOGRAPHIES

Nicole Radziwill is an associate professor in the Department of Integrated Science and Technology at James Madison University in Harrisonburg, VA. She is an ASQ Fellow and is a Certified Six Sigma Black Belt (CSSBB) and Certified Manager of Quality/Organizational Excellence (CMQ/OE). She has a doctorate in quality systems from Indiana State University, and her research focuses on the quality and innovation in cyber-human production systems. She is one of ASQ's Influential Voices and blogs at http://qualityandinnovation.com.

Morgan Benton is cofounder of the Burning Mind Project, and an associate professor in the Department of Integrated Science and Technology at James Madison University in Harrisonburg, VA. He holds a doctorate and master's degree in information systems from the New Jersey Institute of Technology, and a bachelor's degree in leadership studies/sociology/physics from the University of Richmond. His research focuses on innovation in learning, higher education, and transformation.

APPENDIX: Analytic Hierarchy Process (AHP)

The AHP package in the R Statistical Software (R Core Team 2016) can be used to process pairwise comparisons data in the manner shown in this paper. If a user does not have access to an R installation, he or she can also do the analysis on the web at https://ipub.com/apps/ahp/ as shown here. There are three steps in the process:
1. Prepare a YAML file containing all pairwise comparisons
2. Generate and validate the hierarchical model
3. Perform the analysis, checking consistency scores and overall weightings of alternatives.

Step 1: YAML

YAML (YAML Ain't Markup Language) is a data serialization mechanism that is used in some programming environments. Like FORTRAN, it is sensitive to the column in which each line of text begins. Complete instructions for how to generate a YAML file for AHP can be found at https://cran.r-project.org/web/packages/ahp/vignettes/file-format.html. The YAML file for the example in this article is shown in Figure 8 on the next page.

Step 2: Generate and Validate the Hierarchical Model

Next, go to https://ipub.com/apps/ahp/. Click on the "Model" tab, and copy and paste the entire contents of the YAML file in step 1 into the text area. Be sure to start in column 1 of line 1, and be sure not to leave any extra characters or blank lines at the bottom. To check if the YAML file is valid, click "Visualize." If a picture of the decision hierarchy appears, the YAML file meets the requirements and one may proceed.

Step 3: Perform the Analysis

Process the AHP data by clicking "Analyze." Be sure the values in the consistency column are no greater than 20 percent. The alternative that best meets one's goals is the one with the largest weight in the first row.

FIGURE 8 YAML file

```
Version: 2.0
#########################
# Alternatives Section
#
Alternatives: &alternatives
# Here, we list all the alternatives, together
with their attributes.
# We can use these attributes later in the file
when defining
# preferenceFunctions. The attributes can be
quantitative or
# qualitative.
OLD:
NEW:
#
# End of Alternatives Section
####################################
####################################
# Goal Section
#
Goal:
# The goal spans a tree of criteria and the
    alternatives
  name: Select Between Old and New Chatbots
  description: >
    Model quality assessment as a decision
    process.
  author: unknown
  preferences:
    # preferences are typically defined
    pairwise
    # 1 means: A is equal to B
    # 9 means: A is highly preferable to B
    # 1/9 means: B is highly preferable to A
    pairwise:
     - [Performance, Humanity, 7]
     - [Performance, Affect, 7]
     - [Performance, Accessibility, 1/3]
     - [Humanity, Affect, 1/5]
     - [Humanity, Accessibility, 1/7]
     - [Affect, Accessibility, 1/7]
  children:
    Performance:
      preferences:
        pairwise:
         - [UnexpectedInput, Escalation, 7]
      children:
        UnexpectedInput:
          preferences:
            pairwise:
             - [OLD, NEW, 3]
          children: *alternatives
        Escalation:
          preferences:
            pairwise:
             - [OLD, NEW, 7]
          children: *alternatives

    Humanity:
      preferences:
        pairwise:
         - [Transparent, ThemedDiscussion,
           1/5]
         - [Transparent, SpecificQs, 1/5]
         - [ThemedDiscussion, SpecificQs, 1]
      children:
        Transparent:
          preferences:
            pairwise:
             - [OLD, NEW, 1]
          children: *alternatives
        ThemedDiscussion:
          preferences:
            pairwise:
             - [OLD, NEW, 1/3]
          children: *alternatives
        SpecificQs:
          preferences:
            pairwise:
             - [OLD, NEW, 1/5]
          children: *alternatives
    Affect:
      preferences:
        pairwise:
         - [Personality, Entertaining, 1/5]
      children:
        Personality:
          preferences:
            pairwise:
             - [OLD, NEW, 1/5]
          children: *alternatives
        Entertaining:
          preferences:
            pairwise:
             - [OLD, NEW, 1/5]
          children: *alternatives
    Accessibility:
      preferences:
        pairwise:
         - [MeaningIntent, SocialCues, 7]
      children:
        MeaningIntent:
          preferences:
            pairwise:
             - [OLD, NEW, 3]
          children: *alternatives
        SocialCues:
          preferences:
            pairwise:
             - [OLD, NEW, 1]
          children: *alternatives
#
# End of Goal Section
####################################
```