Connecting with *Elizabeth*:
Using artificial intelligence as a data collection aid

Craig F. Kolb, Ask Afrika
Dr Peter Millican, University of Oxford

Abstract

The present paper proposes a new approach to data collection in marketing research. A simple form of artificial intelligence, commonly referred to as 'chatbot' technology, is shown to be a viable way of probing open-ended questions. A straightforward open-ended question was asked of ten subjects, who were then successfully probed by a chatbot named *Elizabeth*. Criteria are put forward for evaluating chatbot interviewing success, and hypotheses are suggested to direct future research on chatbot-based interviewing.

Introducing the 1960s A.I. Rogerian psychotherapist – ELIZA

One of the earliest attempts at creating a 'chatbot', a computer program which could converse with a human in 'natural language', was developed by a computer scientist, Dr. Joseph Weizenbaum. Even though a fairly simple program, it had a massive impact on the psychological community and popular imagination of the 1960s. Designed to mimic a Rogerian psychotherapist, many were impressed by the apparent intelligence of the program, called ELIZA, and some even pondered the possibility of its replacing human psychotherapists. Running on the imposing machines of the day, the human subject communicated with ELIZA via a teletypewriter (with input typed in and responses fed back to the user via a printer). Weizenbaum found that his early subjects were often convinced of the intelligence of the computer, some showing clear signs of emotional attachment (Stephens, 2005).

Weizenbaum (1966) gave the following now classic example of a rather comical dialogue between ELIZA and a young woman (ELIZA's responses are in capitals).

> Men are all alike.
>
> IN WHAT WAY?
>
> They're always bugging us about something or other.
>
> CAN YOU THINK OF A SPECIFIC EXAMPLE?
>
> Well, my boyfriend made me come here.
>
> YOUR BOYFRIEND MADE YOU COME HERE?
>
> He says I'm depressed much of the time.
>
> I AM SORRY TO HEAR YOU ARE DEPRESSED.
>
> It's true. I am unhappy.
>
> DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY?
>
> I need some help, that much seems certain.
>
> WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP?
>
> Perhaps I could learn to get along with my mother.
>
> TELL ME MORE ABOUT YOUR FAMILY.
>
> My mother takes care of me.
>
> WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?
>
> My father.
>
> YOUR FATHER?
>
> You are like my father in some ways.
>
> WHAT RESEMBLANCE DO YOU SEE?
>
> You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE?

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU?

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU?

My father is afraid of everybody.

WHAT ELSE COMES TO YOUR MIND WHEN YOU THINK OF YOUR FATHER?

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE?

Admittedly most conversations with ELIZA are far less convincing than this one, and it should be stressed that the appropriateness of the final sentence, apparently connecting bullies with the boyfriend's compulsion, was lucky coincidence.[1] But the fact that this kind of conversation is achievable at all by such a simple program is remarkable, and it is not surprising that in the early days of computing, ELIZA's performance was widely considered to be extremely impressive. Indeed it seemed to promise a golden future for natural language conversation systems, a promise which has so far remained unfulfilled.

How does it work?

There are two key components to ELIZA's workings, first the process used by the program itself, and secondly an easily missed component – the assumptions made by the human being.

The process

The process used by ELIZA to construct sentences and respond to user input (or non-input as is sometimes the case), is described by Weizenbaum (1966) as:

1. *The identification of the 'most important' keyword occurring in the input message.*

2. *The identification of some minimal context within which the chosen keyword appears; e.g., if the keyword is 'you', is it followed by the word 'are' (in which case an assertion is probably being made).*

3. *The choice of an appropriate transformation rule, and, of course, the making of the transformation itself.*

4. *The provision of a mechanism that will permit ELIZA to respond 'intelligently' when the input text contained no keywords.*

5. *The provision of machinery that facilitates editing, particularly extension, of the script on the script writing level.*

Human assumptions

The second aspect relied on the assumptions made by the human 'patient' about the expected behaviour of a psychotherapist in order to make intelligent responses within the limits of the role.

---

1 A standard chatbot technique for simulating coherent conversation, pioneered by ELIZA, is to spot phrases beginning with 'my' (e.g. 'my boyfriend made me come here', 'my mother takes care of me', 'my father'), and then to feed them back to the user at a later stage, with pronouns adjusted and 'my' replaced by 'Does that have anything to do with the fact that … ?'. In this case it worked well, but it can turn out badly (e.g. 'Does that have anything to do with the fact that your father?').

More specifically one expects a psychotherapist to be enigmatic and never to answer a question for you, but to keep nudging you to find the answer to your own problems.

This was touched on by Weizenbaum (1966) while explaining why ELIZA took on the role of a Rogerian psychotherapist:

'This mode of conversation was chosen because the psychiatric interview is one of the few examples of categorized dyadic natural language communication in which one of the participating pair is free to assume the pose of knowing almost nothing of the real world.'

A variation on this pose is perhaps deliberate 'obscurantism' – the difference being that the one party to the conversation really is ignorant of the necessary facts, and attempts to be obscure in order to hide that ignorance. This tends to occur within certain modes of discourse such as business meetings, consulting (prompting Deloitte's to develop software to combat consulting verbosity), and last but not least, politics. In this case one is not free to assume the 'pose of knowing nothing'and so it is replaced by reliance on the human tendency (which is more prevalent in certain cultures) of the hearer not revealing ignorance (even if it is fully justified given the ambiguity of the discourse).

If an unequal power balance exists (lack of power parity), such tolerance for incomprehensible or irrational utterances increases. This is likely to occur even more frequently in certain cultures. For instance, the Japanese are known to employ a deliberately abstract way of speaking to maintain a public facade (tatemae) by using words and phrases that are so abstract as to be virtually meaningless. One is apparently expected to assume that what is being said is not necessarily true (JGuide: Stanford guide to Japan information sources, 2005).

In a different variation – different in that it is not a deliberate deception – marketing research interviewing allows for legitimate ignorance on the part of the interviewer. Marketing research has become widely enough experienced by the general populace so that the respondent can be relied upon to have certain commonsense expectations about how the exchange will work. In particular the interviewer is free to assume the pose of 'this is about you, what you know and your opinions; not about me' . To add to this, other typical assumptions include: 1) the exchange will be fairly focused and will not deviate too much from the topic under discussion; 2) it will be a fixed-initiative dialogue under the control of the interviewer.

While in the context of a general conversation the ideal chatbot would hold masses of commonsense knowledge, in an interview setting this is generally not necessary and in fact may be an obstacle. Therefore ignorance becomes quite tolerable in this context. As a result chatbot technology is not prevented from performing the modest task of probing open ends and conducting short semi-structured interviews.

The problem with open-ended questions

Open-ended questions (also referred to as open ends) are questions where the response options are not pre-determined beforehand, the respondent being free to answer in his or her own words. They are frequently included in quantitative and qualitative investigations, within the field of marketing research and many other academic and applied fields of social science where data must be collected in an unstructured way.

Situations in which open-ended questions are particularly appropriate include those where the range of response options is not well understood, where more comprehensive responses are required, and where avoidance of 'suggested' responses (i.e. leading questions) is desirable and the researcher wants a response in the respondent's own words.

Typically, responses to open-ended questions are very superficial unless probed further. As Ivis *et al.* (1997) found, respondents are more forthcoming when answers are suggested, as would be the case where a range of possible response options are read out. However, suggesting answers is not always desirable if an option might never have occurred to the respondent naturally (during the interview or in the market place), because a suggested answer can then amount, in effect, to a leading question. The other alternative, of recording responses purely by category but not showing or reading these category options to the respondent, can result in information loss.

The examples in Table 1 come from a South African telephonic survey of mobile (cellular) phone users conducted in 2004, who were asked why they had chosen a particular mobile phone network (the responses were not probed).

Table 1: Example of un-probed responses to an open-ended question

Why did you choose A instead of other network providers?
Because I like it.
Because it was cheap.
Because my husband had problems with C and B.
Because of upgrading.

Note: The three cell phone operators are referred to as A, B, and C, as in this paper we are not concerned with actual brands.

As can be seen the responses are far from comprehensive. If, for example, the purpose of the open-ended question was to obtain a list of network operator attributes for use in hard laddering, these responses would not be very useful. Most of the answers probably reflect only the most top-of-mind attributes, while many of them are too vague to be of any use at all.

While this is easily remedied by requesting that the interviewer probe the responses, such probing is not possible where the data collection method involves self-completion. This brings us to the point where we introduce a possible solution which can be applied in environments where interviewing is carried out by a computer, such as computer-aided web interviewing (CAWI) or computer-aided personal interviewing (CAPI).

An exploratory test: *Elizabeth*, a chatbot system adapted for interviewing

Today the field of artificial intelligence ranges from games and chatbots designed merely for entertainment, through subject-specific 'expert systems', to major projects aiming to develop highly sophisticated and powerful programs that could potentially replace humans and even outperform them over a broad field. One of the most ambitious attempts to date is the 20-year long, $25 million CYC project which follows a 'bottom up' approach to learning, the ultimate aim being to develop an intelligence with common sense as well as extensive knowledge and reasoning capabilities. CYC's databases are continuously updated by a team of capturers who program in commonsense reasoning (Grossman, 1998; Reingold and Nightingale, 1999). Modelling and exploiting common sense is a surprisingly complex and problematic task, which is partly why it has proved so difficult to create natural language systems that can maintain a plausible conversation except in a very restricted domain. Although back in the 1960s ELIZA seemed to promise so much, there was general disappointment when the Loebner awards began in the early 1990s, providing annual tests of how far chatbots had progressed. Judges lamented that the programs of the day seemed to do little better than the 25 year-old ELIZA (Shieber, 1994).

However our claim in this paper is that a fairly modest extension to the standard chatbot techniques that already existed in ELIZA four decades ago, is in fact quite adequate to the task of solving

certain specific problems in marketing research, notably those concerned with the acquisition of information by open-ended probing. The distinctive nature of these problems, and of the context within which they arise, enables a relatively limited and very cost-effective technology to yield valuable results, while making no pretence to general conversational or all-round commonsense capability. To investigate this possibility the *Elizabeth* chatbot was used, programmed not to deceive (as chatbots standardly are), but instead to emphasise structured questioning, limited but appropriate responses, relevantly triggered probing, and recording of information.

Introducing Elizabeth

In searching for a way to implement a system with qualities broadly similar to those demonstrated by ELIZA but appropriately enhanced, various options were considered. It was eventually decided that *Elizabeth*, a chatbot originally developed by Dr. Peter Millican of Oxford University as an educational tool (Millican, 2003-5), would be most suitable. The main attractions of this system are that it is user-friendly, highly flexible, rigorously specified and comprehensively documented, and can be reprogrammed using a simple syntax which does not require a knowledge of general programming languages. In addition its operations can be inspected in detail, with all data, programmed behaviour, and real-time internal processing revealed in easily understood tables that can be accessed through a straightforward interface, greatly facilitating development and 'debugging' of processes. *Elizabeth* is freely available for non-commercial use, and as such is an ideal platform for marketing research practitioners wanting to experiment with chatbot interviewing systems.

Although *Elizabeth* takes inspiration from ELIZA, it has been developed in such a way that the simple stimulus-response behaviour and limited memory capabilities of the original (e.g. its ability to remember phrases such as 'boyfriend made me come here' when they occur after 'my') have been generalised, abstracted, and supplemented with facilities for programmable iteration, recursion, and script self-modification.[2] These augmented capabilities can still be used very easily for chatbot design of the conventional sort, but they make *Elizabeth* distinctive amongst chatbots in also possessing a powerful algorithmic capability, as illustrated by scripts designed by Dr. Millican (available and documented in the standard downloadable package) for such things as recursive arithmetic, questionnaire structuring, grammatical analysis, and even resolution theorem-proving. He developed the system during 20 years in the School of Computing at Leeds University, as a means of introducing novice students to Artificial Intelligence in an entertaining and straightforward manner, enabling them to start off by designing a playful chatbot (e.g. to converse about their favourite football team or music band) but then to move on, within the same structure and using extensions of the same methods, to explore and implement a range of relatively advanced A.I. techniques.

Many of the more sophisticated features of *Elizabeth* were not used in this work, given that interviews are more limited in scope than general conversation, and our intention at this stage was

---

2 *Elizabeth* is designed in such a way that any textual pattern matching, at any point of the processing cycle, can be made to trigger almost any desired modification to the 'script' which determines how future processing takes place (so for example keywords, responses, and other transformations can be added or deleted, memories saved, modified or deleted etc.). The standard processing cycle allows for 'input' transformations to screen the initial input, 'keyword' transformations to select prepared responses depending on keyword identification, and 'output' transformations to enable the text to be further adapted (e.g. to ensure conformity of pronouns etc.). At both the input and output stages, a wide range of match-and-replace operations can be applied, iteratively if desired, enabling complex processing to be achieved very straightforwardly. Any transformation operation can be made to trigger splitting of the text, with the option of sending any of the parts back again through the entire cycle. This enables full recursive behaviour to be implemented, with an additional stage of 'final' transformations available for further testing and processing when the parts have been recombined (which allows further recycling if required). Thus although *Elizabeth* has been designed to be very easy to use from the start, it incorporates the potential for highly sophisticated processing by extensions of the same essentially simple techniques. For further explanation of all this, see the online *Elizabeth* documentation, and in particular the *PowerPoint* presentation bundled with the download package.

to prove a concept rather than to develop a full-fledged self-standing system. The process used (and explained further down) was chosen to make implementation as easy as possible and to lower the potential for errors. It is fairly crude, and requires the syntax to be reprogrammed for each open-ended question posed, but it could easily be streamlined and developed further in order to facilitate more flexible application and scaling-up if the technique proves valuable in practice.

The results: *Elizabeth* in action

Elizabeth was programmed to ask a similar basic question to that asked in the telephonic survey 'Why did you choose your current cell phone network operator?', the aim being to elicit a list of cell phone (mobile) network operator attributes which could be compared to the telephonic survey.

A small convenience sample of 10 employees at Ask Afrika was asked to undergo an interview with an 'automatic interviewer'. The program was sent via e-mail, for installation on the respondent's desktop computer. For illustration three of these interviews are shown below. Note: 1.) These have been edited for spelling errors, but have been left unaltered otherwise; 2.) The network operator company names have been relabelled A, B and C as the emphasis of this research is not a comparison of the companies concerned.

Example 1

> WHY DID YOU CHOOSE YOUR CURRENT CELL PHONE NETWORK OPERATOR?
>
> To save money.
>
> LIST ALL OF THE THINGS THAT WORRY YOU ABOUT YOUR NETWORK PLEASE.
>
> Network coverage and 3g.
>
> WHAT ELSE CAN YOU SAY ABOUT CELL PHONE FEATURES?
>
> I like my new handset, which is very important to me. I also like the extra 'Ahas' – such as emergency services.
>
> WAS IT ONLY ABOUT THE CELL PHONE HANDSET?
>
> No, but they did deliver it to me for free.   I didn't have to go anywhere.
>
> LIST ALL OF THE THINGS YOU SEE AS GOOD ABOUT YOUR NETWORK PLEASE.
>
> Cheap; additional services such as free delivery; nice shows and stuff.
>
> IS SAVING MONEY MORE IMPORTANT THAN OTHER THINGS?
>
> Value for money is important, I don't mind paying A BIT more, but then I want to get more.
>
> WHAT DID FAMILY AND FRIENDS HAVE TO SAY ABOUT YOUR CHOICE OF NETWORK?
>
> They don't really care, but I also have my mom on A now.
>
> HOW DO YOU FEEL ABOUT FAMILY INVOLVEMENT?
>
> I love my family, but we don't really care about each other's cell phone habits.
>
> HOW MUCH OF THE FINAL DECISION ON A PARTICULAR NETWORK WAS REALLY YOURS?
>
> 90%, 10% of it was because of the fact that I work on A projects.
>
> WOULD YOU USE THIS NETWORK AGAIN IN FUTURE, IF YOU HAD TO CHOOSE AGAIN?
>
> Yes.
>
> YES? PLEASE GO INTO A BIT MORE DETAIL.

I like the additional services; and I like the A image.

WHY DO YOU LIKE 'THE ADDITIONAL SERVICES'?

They deliver straight to me!

As can be seen in the transcript above, if this had been left to self-completion, the only network operator attribute which would have been gleaned from this respondent was 'to save money'. The additional attributes yielded by *Elizabeth* included: network coverage, lack of 3G (on this respondent's particular network), emergency services, free handset delivery, and shows.

Example 2

WHY DID YOU CHOOSE YOUR CURRENT CELL PHONE NETWORK OPERATOR?

Best option available at the time.

YOU SEEM POSITIVE ABOUT IT?

Yes.

YES? PLEASE GO INTO A BIT MORE DETAIL.

The price and service of the product was right.

IS SAVING MONEY MORE IMPORTANT THAN OTHER THINGS?

Sometimes.

LIST ALL OF THE THINGS THAT WORRY YOU ABOUT YOUR NETWORK PLEASE.

<void input>

PLEASE THINK CAREFULLY – SURELY THERE IS SOMETHING YOU CAN SAY?

Unavailability of services when you require it, like balance enquiry.

LIST ALL OF THE THINGS YOU SEE AS GOOD ABOUT YOUR NETWORK PLEASE.

Network never down, never experience problems.

ANYTHING ELSE YOU CAN SAY ABOUT THE NETWORK?

No.

NO? PLEASE GIVE A BIT MORE INFORMATION.

I do not have anything more to say about the network, I am happy with it.

WHAT DID FAMILY AND FRIENDS HAVE TO SAY ABOUT YOUR CHOICE OF NETWORK?

Nothing, it is my personal choice.

HOW MUCH OF THE FINAL DECISION ON A PARTICULAR NETWORK WAS REALLY YOURS?

100%.

WOULD YOU USE THIS NETWORK AGAIN IN FUTURE, IF YOU HAD TO CHOOSE AGAIN?

Maybe, depending on product range in comparison with other networks.

In this second transcript, the initial response would have yielded nothing useful if it hadn't been probed further. Notice also that a void response was corrected, and short Yes/No/Nothing answers were not tolerated by the system. Instead of a simple 'yes', this yielded 'price' and 'service' as possible attributes. Balance enquiry reliability, and product range also surfaced as possible network attributes.

Example 3

WHY DID YOU CHOOSE YOUR CURRENT CELL PHONE NETWORK OPERATOR?

I heard that C is good and at the time (in 1998) B was a bit behind, while A didn't exist

SO YOU WERE REALLY HAPPY?

I'm happy with the network, but the service of C is extremely poor! Plus the coverage is not always great

WHAT ABOUT THE SERVICE?

Pathetic!

LIST ALL OF THE THINGS THAT WORRY YOU ABOUT YOUR NETWORK PLEASE.

The coverage was poor at many places where I have been. I was in E-Transvaal and there were many places where I had no coverage

IS IT THAT BAD?

No, but C always brags how brilliant they are, so they should practice what they preach!

LIST ALL OF THE THINGS YOU SEE AS GOOD ABOUT YOUR NETWORK PLEASE.

It's a phone, so it rings when I call and it sends SMSs when I send SMSs and it never gives me problems, so that's nice. It came up with that nice bundle of SMSs for R49 for 150 SMSs, and I like that a lot!

WAS IT ONLY ABOUT THE CELL PHONE HANDSET?

No.

NO? PLEASE EXPAND ON WHAT YOU MEAN.

Well, the cell phone handset is good too. I have a Nokia and they are brilliant phones!

YOU SEEM POSITIVE?

Yes! I feel confident in C, but I am very disappointed with the service at the C shops and with the help line of the phone, i.e. 111 or 110. They are slow, irritating and never solve my issue the first time! But otherwise, I'm happy with C

WHAT DID FAMILY AND FRIENDS HAVE TO SAY ABOUT YOUR CHOICE OF NETWORK?

Not much. However, I convinced my girlfriend to go for B, not C, because of the poor service and so now she has B and is very happy with it. Plus the service is much better!

WHAT WAS WRONG WITH THE SERVICE, PLEASE GIVE DETAIL WHEN REPLYING.

Extremely SLOW, irritating, disrespectful, arrogant and never solved my problem the first time!

HOW MUCH OF THE FINAL DECISION ON A PARTICULAR NETWORK WAS REALLY YOURS?

100% my decision! Although in 1998, I only heard of C, no other network, so I didn't have so much choice!

WOULD YOU USE THIS NETWORK AGAIN IN FUTURE, IF YOU HAD TO CHOOSE AGAIN?

Maybe. I'd maybe give B a try. I am a conservative person, so I tend to stick with what I have.

As can be seen this transcript yielded an abundance of attributes: coverage, service attributes (such as the speed with which issues are handled, and the attitude of service personnel), and brand familiarity.

Further examples are available from the corresponding author on request.

Criteria for evaluating chatbot interviewing performance

General criteria

In an attempt to set a general criterion by which intelligence could be identified, mathematician and computing pioneer Alan Turing (1950) proposed a simple test. If a computer was able, through online conversation, to fool a human into believing that it was human too, then it was for all practical purposes intelligent.[3] In one of the first large-scale implementations of the test in 1991, Hugh Loebner introduced the somewhat controversial Loebner prize, a $100,000 dollar prize offered to anyone who could develop an A.I. that could pass the test in appropriate conditions.

While certain Loebner competition entrants have managed to exhibit superficial (though sometimes quite extensive) knowledge in a limited area, as remarked earlier they have universally failed to maintain convincing conversations of any significant extent, due in part to a lack of commonsense knowledge. Ignorance of basic facts, such as that a kitchen generally houses a sink or that a house is bigger than an insect, could reveal to a human judge that the respondent was a machine, as could the related inability to assess or respect conversational relevance. Other common failings include inability to cope with idiom, non-standard syntax, conversational shorthand, or anaphoric reference (including references back to earlier remarks or picking up of previous conversational threads). As a result chatbots all fall down badly when attempting to hold general conversations (Mullins, 2005), and Hugh Loebner's prize seems very unlikely to be claimed in the foreseeable future.

All this implies a risk that chatbots will provoke hostility, thus undermining their usefulness for marketing research. Examination of various failed chatbot transcripts reveals that the human subject started to become hostile under the following conditions: When responses were irrelevant to the topic;[4] when an incorrect grammatical transformation occurred (for instance second person instead of third); when incorrect semantics attached to a word because of failure to understand the context; and when the chatbot repeated the same statement over and over.

Criteria in a marketing research setting

Fortunately, the demanding criteria of the Turing Test are not entirely relevant to most types of dialogue carried out in a marketing research setting. As mentioned above, respondents can be made to have a very different set of expectations in a research context, so there is typically no requirement to appear intelligent in general conversation, nor would meeting this criterion equate to good performance. The demands are in fact far less complex.

This echoes the practical dialogue hypothesis (a practical dialogue being a dialogue aiming to accomplish a concrete task) suggested by Allen *et al.* (2001), which states: "The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence". In particular, the marketing research interview falls into the 'information seeking' subclass of practical dialogues, which are particularly undemanding in respect of general competence.

Typically, research interviews can be characterised as fixed-initiative dialogues, where one participant controls the conversation. In contrast mixed-initiative dialogues allow for conversation flow to switch between the two participants, each being able to initiate new lines of conversation

---

3 The Turing Test has spawned a huge literature. For some relatively recent examples, see the papers in Millican and Clark (1996) and Shieber (2004).

4 While a statement irrelevant to a topic could be seen as initiating a new topic in a mixed-initiative dialogue, it can also be seen as evasive and revealing a lack of knowledge, just as it does at times in human conversation.

(Allen *et al.*, 2001). Fixed-initiative dialogue is generally used in marketing research (consciously or not) because it facilitates more efficient transfer of information, which is obviously important to a profit-maximizing marketing research business. Fixed-initiative dialogue fortuitously reduces the demands on a chatbot system, so that it only has to pose questions and field responses to them, not having to deal with new lines of thought and questioning introduced by a respondent.

In determining a set of criteria to evaluate a chatbot interview, Grice's maxims (1975) may be useful since they are built on the premise that conversation is purposive behaviour, and the maxims characterise meaningful conversation (Saygin and Cicekli, 2002). This does not mean that all conversations will conform to them, or even that they should. Grice's maxims include:

- Relevance – Be relevant.
- Quantity – Do not make your contribution less or more informative than is required.
- Quality – Try to make your contribution one that is true: Do not say what you believe to be false; do not say that for which you lack adequate evidence.
- Manner – Be perspicuous: Avoid obscurity of expression; avoid ambiguity; be brief; be orderly.

To evaluate *Elizabeth*'s performance in this study, four criteria were eventually decided on. Three of these criteria were suggested by van der Zouwen (2001) to evaluate a human interviewer, and somewhat resemble Grice's maxims. These include: 1) 'relevance of interviewer questions', relevant questions being operationally defined in this paper as questions specified by the researcher which have not previously been answered; 2) avoidance of 'suggestion' (which would contribute to achieving the 'quality' maxim); 3) 'relevance of respondent answers', operationally defined as being those answers which answer the question in part or in full. An answer which lived up to the 'quality' and 'manner' maxims would therefore contribute to relevance, and this obviously precludes: deliberate falsehoods, poor responses which could be improved on if more cognitive effort were exerted, speculation which was not requested of the respondent, and truths which do not answer the question (i.e. going off topic). An interviewer can ensure relevance of respondents' answers, to some extent, by exercising closed loop control. Van der Zouwen and Smit (2005) describe closed loop control as that which can be exerted by the interviewer as events unfold based on a feedback loop, while open loop control refers to the control typically exerted by researchers in setting the initial parameters – such as questionnaire wording – but without the ability to respond if deviations occur.

Based on a modification of Grice's quantity maximum, a last criterion was proposed specifically for this study: 4) 'maximisation of the volume of attributes elicited' (i.e. cell phone network operator attributes).

As judged by three of these criteria, *Elizabeth* performed well. First, on the criterion 'relevance of interviewer questions', in the rare cases where irrelevant statements were made, they were irrelevant only because they had already been answered and were thus redundant, not because the open loop control of the researcher had been broken. In contrast human interviewers were found in a test by van der Zouwen and Smit (2005) to violate open loop control on average 47% of the time across eight test questions – either deviating from the question wording, failing to ask at the right time, or omitting parts of the question wording or response options. *Elizabeth*'s memory feature ensured that the irrelevant statements were not repeated, the set questions serving the purpose of driving the interview forward.

On the second criterion, of 'suggestion', *Elizabeth* would no doubt outperform any human interviewer. Van der Zouwen and Smit (2005) found that suggestion (labelled 'hinting') ranged from 15% to 39% across 4 test questions.

On the third criterion, respondents' answers are generally relevant, however this is purely due to respondent cooperation as *Elizabeth* does not have the knowledge necessary to evaluate the quality of responses, except at the most rudimentary level (non-acceptance of void and crude Yes/No/Nothing responses). Typically, interviewers would be expected to exercise closed loop control and engage in repair behaviour wherever respondents gave irrelevant answers – the options being to repair, repair inadequately or neglect to repair (van der Zouwen and Smit, 2005).

Lastly, the fourth criterion, 'maximisation of the volume of attributes elicited', was successfully met relative to self-completion.

While the chatbot performed well in terms of being able to meet the four criteria, and outperforms self-completion on these criteria (though it is a mute comparison on criteria one and four), some human interviewers should be able to outperform the chatbot. In addition, it is recognised that criteria one, three and four may present problems when the following independent variables are altered: 1) respondent characteristics; 2) the interview length; 3) the context presented to the respondent (e.g. conversation vs. research interview, masquerading as a human vs. chatbot); and 4) chatbot behaviour. In order to better understand which variables impact on success – in terms of meeting criteria, one, three and four – hypotheses for testing are suggested in the 'future research directions' section of the paper.

# How did the present application of *Elizabeth* work?

The *Elizabeth* system was programmed to ask the question 'Why did you choose your current cell phone network operator?', then to follow up on the first response with probes – either based on keywords in the respondent's answers, or alternatively using a set question (as would be the case in semi-structured interviewing).

Keywords were at times bundled into positive and negative emotional categories, with a facility to identify when a positive keyword was being negated by words such as 'not'. Liking for a particular thing was also probed with a question incorporating the phrase used by the respondent. In addition, *Elizabeth* monitored for simple Yes/No answers and void answers, following up with prompts to expand on the response given when these occurred.

In designing a quantitative questionnaire, the designer attempts to take into account all the possible responses that a respondent could give. Similarly when designing a semi-structured questionnaire, a limited set of direct questions (topics) are included to ensure basic areas of interest are covered adequately, and that the interview does not wonder into irrelevant areas, and that repetition does not occur. The present chatbot overcomes this to some extent, by remembering which questions have been asked, and which keywords or concepts have already been probed.

Diagram 1 illustrates the flow of logic in this implementation.

Diagram 1: The simple steps followed by *Elizabeth* in probing open ends
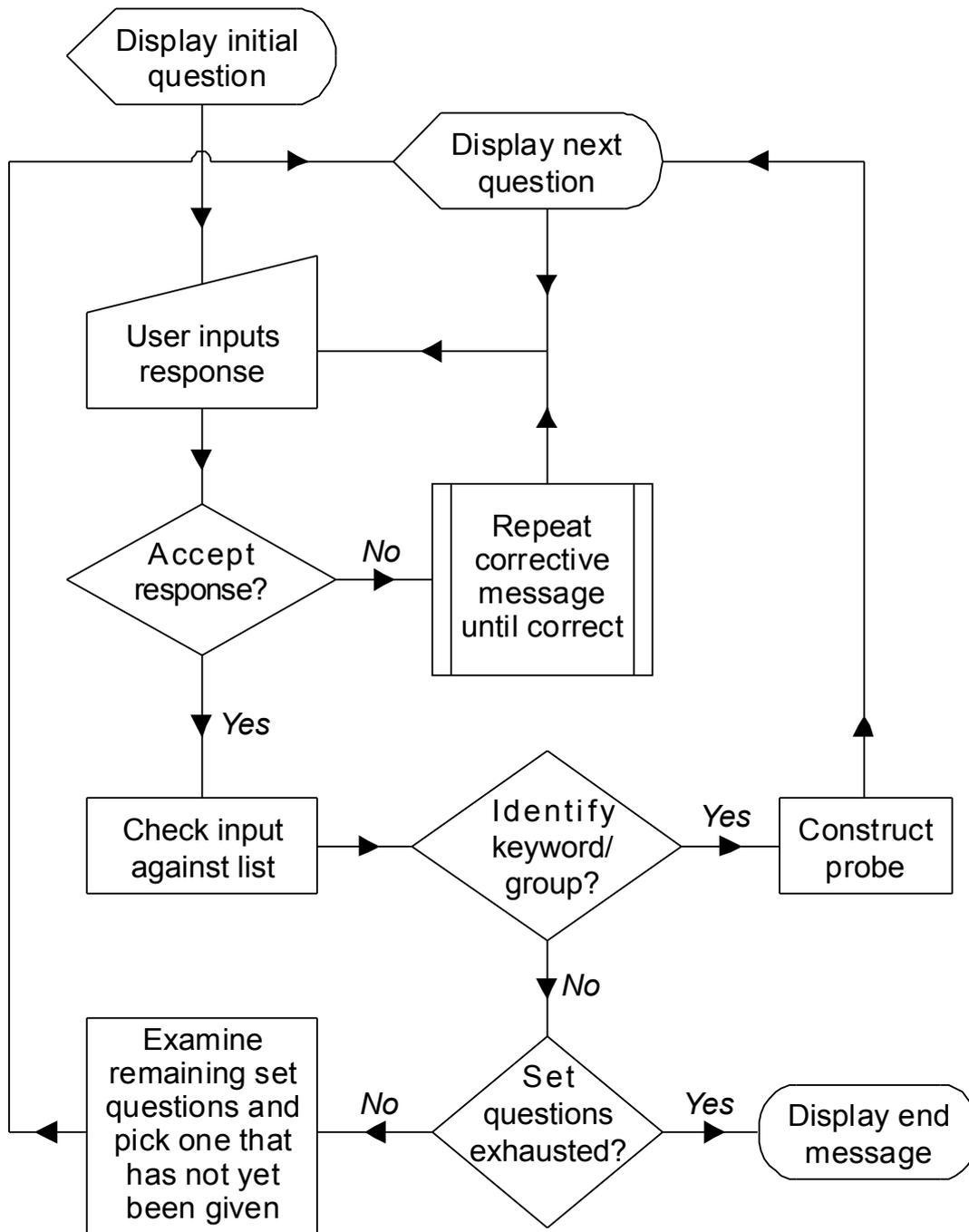
Respondent psychology – how does one set the context for the respondent?

Given that the context of the conversation has an impact on the success of the conversation, careful attention needs to be paid to how the software is introduced to the respondent. Numerous articles refer to people becoming attached to Weizenbaum's ELIZA program in the role of psychotherapist (though of course the original 'psychotherapist' context, used by Weizenbaum when introducing the ELIZA program, is inappropriate in a marketing research setting).

Should we pretend that a human interviewer is carrying out the interview or should we make it clear that an A.I. or chatbot is involved? The first option has various problems, not the least of which is the ethical undesirability of deceiving respondents, along with the fact that the respondent may soon discover the truth.

Interestingly the ELIZA of the 1960s sometimes successfully carried out fairly long and apparently coherent conversations, even though subjects were made to understand that it was a program from the start, while in more recent times Loebner contestants have been expected (though admittedly with limited success) to maintain an extended dialogue with judges who are all forewarned that they might be conversing with an artificial intelligence. One would expect that when the human had no idea that they were conversing with a machine, it would take the human party rather longer to detect problems. In fact it seems that the possibility that a machine is doing the talking doesn't always occur to people. Honan (2000), for instance, cites anecdotal evidence of this, where an ELIZA program was hooked up to an AOL messenger service, users carrying on elaborate conversations without realising that they were conversing with a chatbot.

This highlights an interesting asymmetry between humans and A.I.s or chatbots – as we are automatically assumed to be what we really are (i.e. human) by the other party to a conversation! This is seldom the case with chatbots, which are rarely encountered in everyday life. One would normally either knowingly have to seek out a chatbot program to engage in conversation through a keyboard, or be informed that one was dealing with a chatbot. (In the Turing test, the human judge at least is made aware of the possibility, making the test harder to pass.) Similar cues also exist in the case of androids, such as voice quality and appearance, although these cues are becoming less noticeable as technology improves, resulting in improved interactions with humans (Yang, 2005). A notable recent example of this is Actroid® , an android receptionist with a realistic appearance, motions and voice synthesizer, and with an ability to engage in limited conversation (Reception Robot Actroid: Press release, 2005).

The second option, stating upfront that 'an artificial intelligence is going to ask you a couple of questions' may raise expectations to an unrealistic level. But also, judging by many of the transcripts available on the web, prior knowledge that a chatbot is carrying out the other end of the conversation often results in the chatbot's being subjected to a degree of interrogative hostility many humans wouldn't survive. See Platt (2001) for an example of this. Another example of less than helpful interaction is displayed in the Loebner contest, where the poor spelling and grammar sometimes deliberately used by Loebner contest judges must add to the potential for confusion.

Ultimately a third option was chosen. It was decided that describing *Elizabeth* as an 'automatic interviewer' would be an ideal compromise, as a reduced set of expectations is attached to the term 'automatic'. From gearboxes to pool cleaners, the term 'automatic' implies a limited type of intelligence. You would not expect to engage in a discussion about the latest news on Sky television with your car's automatic gearbox, however you would at least expect it to make approximately correct gear selections – a limited type of intelligence and at least a step forward from a manual

gearbox. In a similar way, *Elizabeth* in the automatic interviewer role is a step forward as compared to self-completion.

These suppositions, along with the 10 observations (i.e. recorded response scripts), have been used to formulate a number of perceptual and response hypotheses to suggest future research directions, which we outline in the following section.

Future research

The effect of respondent characteristics

Our observations, which are of course not reliably generalisable given the small sample, are nevertheless sufficient to suggest a number of hypotheses which could be tested in future. While far from exhaustive, it is hoped that these will increase understanding of the factors that may impact on the likelihood of success (as defined by our four criteria) in a currently non-existent area of application: the chatbot interview. As suggested previously, the following independent variables are likely to impact on the likelihood of interview success: 1) respondent characteristics, 2) the interview length, 3) the context presented to the respondent (e.g. conversation vs. research interview, human vs. chatbot), and 4) chatbot behaviour.

It is well known that respondent and interviewer characteristics impact on the nature of the responses, and even whether a response is given or not. Pickery and Loosveldt (1998), for instance, established that respondent gender, age, and education affected the number of 'no opinion answers'. These factors would therefore also be expected to impact on how forthcoming a respondent would be, irrespective of whether a chatbot conducted the interview or not. It is therefore important to attempt to isolate this source of variation.

Pickery and Loosveldt (1998) also established that the interviewer affected the number of 'no opinions', but were not able to identify which interviewer variables were important. In the case of an A.I. or chatbot interviewer, however, it is likely that certain characteristics unique to chatbots would impact on how a respondent perceived the interviewer and on the adequacy of respondent behaviour. Specifically, while it is assumed that chatbots will not present problems in terms of violating open loop control (for example by deviating from the researchers' wording), they may be more prone to redundancy by asking a question which has already been answered – a failure to meet the relevance criterion.

The hypotheses outlined below are divided between perceptual hypotheses and response hypotheses. Note that unless otherwise indicated, the subject is assumed to have prior knowledge that the interviewer is a chatbot.

Perceptual hypotheses

- H1: Subjects with strong vigilance and analytical traits are more likely to detect *irrelevant* questions.

This first hypothesis could be tested using a simple experimental design; for example, personality measures such as Cattell's 16PF vigilance and reasoning subscales (16PF world, 2005) could be used to divide subjects into experimental and control groups, followed by an after measure consisting of ratings of perceptions of question irrelevance.

- H2a: The longer the exchange on *a topic*, the more likely the A.I. or chatbot is to be seen as asking *irrelevant* questions.

- H2b: The longer the dialogue overall, the more likely the A.I. or chatbot is to be seen as asking *irrelevant* questions.

Length could be measured as the number of responses/questions given by the A.I. or in other ways such as the number of sentences or words. *Topic* is defined as being the object or subject of a sentence, for example the cell phone handset, or the service quality.

The rationale for H2a and H2b is that as a conversation progresses, and the information provided by the respondent builds up, the probability increases that a question will be asked which has already been answered, or that a question will show a lack of understanding of what went before. Though not only a problem limited to chatbots, it is more likely where a chatbot interviewer is involved, since a chatbot has only limited closed loop control ability (i.e. ability to understand respondent answers and adapt questions accordingly).

This could be illustrated using a probability tree, where using the multiplicative rule, the probability of asking something which has already been answered (either implicitly or explicitly) increases with every unit increase (e.g. each question and response). Suppose that some potential question has a 0.3 (i.e. 30%) probability of being answered each time the respondent speaks. Then after one such exchange, the probability that the question will *not* have been answered is 0.7, after two exchanges, 0.49 and so on. The thick red path in Diagram 2 represents this situation, in which after two responses, it is already more probable than not that the potential question will have been answered.

Diagram 2: The probability of a question remaining unanswered decreases with each answer
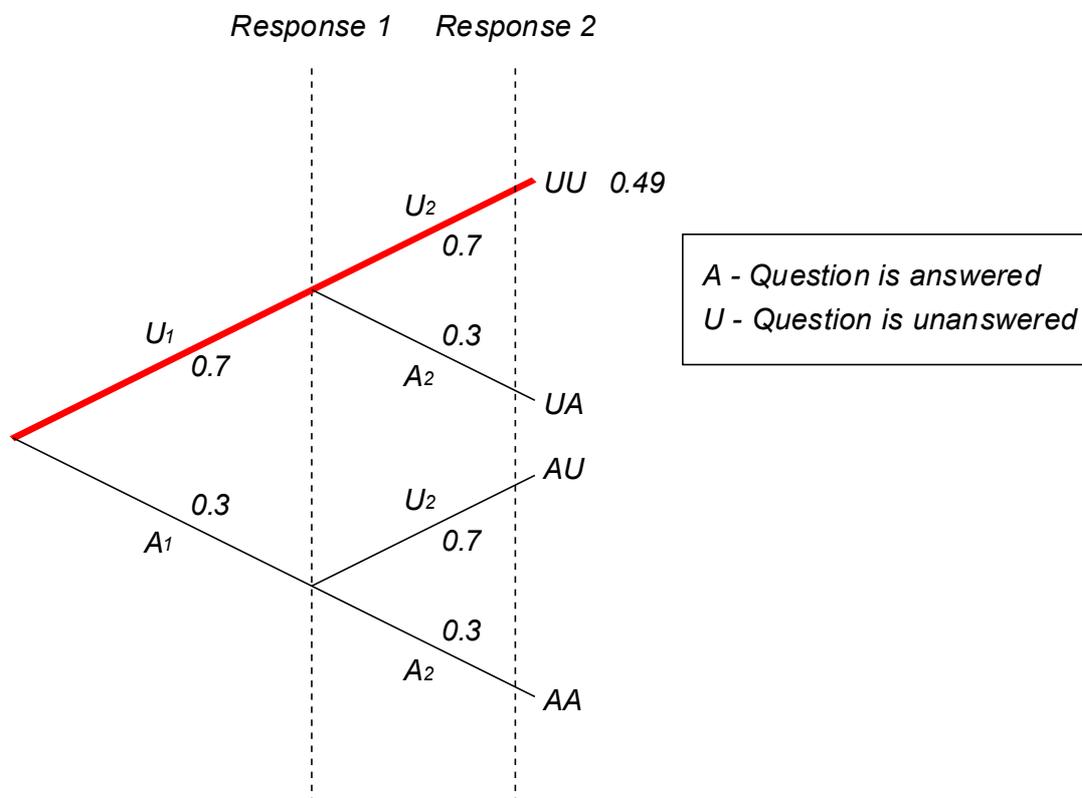
The implication of hypothesis H2a, if confirmed, is that a trade-off will exist in a chatbot-supervised interview. The trade-off is between the number of questions answered (which the researcher would like to maximise) and the risk of the chatbot beginning to be seen as irrelevant due to the lack of a closed control loop ability.

Diagram 3: Chatbot trade-off implication

```
┌─────────────┐                                  ┌─────────────┐
│  Number of  │         ⟵──────────⟶            │   Risk of   │
│  research   │     <      Trade-off      >      │ irrelevant  │
│  questions  │         ⟵──────────⟶            │  questions  │
│  answered   │                                  │ being asked │
└─────────────┘                                  └─────────────┘
```

The probability of *irrelevant* questions occurring will decrease if the chatbot is designed to ask questions covering fresh topics, and to minimise the number of probes triggered by keywords. In the present case, keywords were developed around each topic, and while they were programmed to trigger once only, any number of keyword-specific responses could be triggered by a specific keyword set. This means that more than one question relating to a topic could be triggered, so conceivably a number of questions could be triggered around a single topic. The likelihood of this can be controlled by adjusting the number of keywords which are set up to trigger a particular response.

The implication of H2b, if confirmed, is that the length of a dialogue should be kept fairly short to minimise the appearance of irrelevant questions.

- H3: Subject knowledge of whether they are communicating with an A.I. or chatbot is likely to moderate the statistical relationships described in H1 and H2/H2b.

This could either be measured by third parties' rating evidence of perceived irrelevance of questions in experimental and control groups transcripts, or by subjects themselves who identify the point in conversation at which they recall first having doubts about the relevance of the questions.

Response hypotheses

- H4a: Perceptions of irrelevance will increase the likelihood of hostility, dependent on whether the context is defined as a marketing research interviewing situation, or not defined at all.
- H4b: Increased hostility will decrease the relevance of the respondent's answers.
- H5: Introducing a chatbot by using terminology which creates lower expectations (such as 'automatic interviewer') will moderate the relationship between irrelevant questions and hostility.
- H6: Limited mimicry (operationally defined here as reflecting back a subject's statements, but with grammatical alterations to reflect change in person – i.e. not an exact restatement) will result in increased volume and improved answer relevance. This hypothesis is based on the chameleon effect, an example of this being reported by Poulsen (2005).
- H7: Whether the human party in a dialogue has prior knowledge that an A.I. or chatbot is the other party in a dialogue will moderate the relationships in H4a and H6.

H4a and H5 could be tested using third party ratings of evidence of hostility in experimental and control group transcripts. H6 could be tested in a similar way – using ratings of volume and answer relevance – split across an experimental group (exposed to mimicry) and a control group (not exposed).

Future applications: Pushing the boundaries

Are there other ways in which developments in artificial intelligence could assist in marketing research? Of course a number of dangers lie in attempting to forecast, as illustrated by the hugely over-optimistic predictions for natural language processing made in the 1960s (partly inspired by ELIZA), for 'knowledge-based systems' in the 1980s (when Japan's 'fifth generation project' began with a fanfare), and for the introduction of domestic robots over the last three decades (Platt, 2001); although it finally seems that domestic robots are becoming reality, in particular robotic vacuum cleaners..

So here we shall confine ourselves to considering the currently feasible advantages of artificial intelligence systems, and chatbots in particular, when compared with human intelligence, and how they might be used to assist without waiting for speculative future developments. Of course, the comparison is made against 'systems' with almost no intelligence whatsoever, except some simple skip logic, when considering the self-completion CAWI situation (however for obvious reasons of time and scope, we make no attempt here to explore the difficult question of how intelligence is to be defined).

Assisting human interviewers

Coping with pressure

The potential exists to assist interviewers with phrasing of probes, improving quality under pressure. In the middle of a highly structured quantitative interview it is no doubt difficult for interviewers to switch modes in order to probe open ends. Chatbots might be able to help by suggesting probes based on the responses typed in.

Greater control

Probing that is left to the discretion of the interviewer often leads to open ends which are difficult to analyse, because one cannot be certain of what was asked without wading through transcripts or recordings. A chatbot can be programmed with set questions which can be repeated across interviews, and in a specific sequence, interspersed between probes. A computer-based interface also allows for precise transcripts to be kept and easily collated.

Better memory

Knowledge of previous interviewees' responses to open ends while a survey is in progress can enhance an interviewer's ability to probe. Learning chatbots offer a potential solution, and the *Elizabeth* system already has facilities to assist with the implementation of this, although effective design of a learning structure is a complex matter.

During the execution of a survey, interviewers could facilitate the process by probing areas brought up by other respondents. In new markets, it is often the case that human interviewers must learn as they go along. Artificial intelligences should theoretically be capable of remembering far more than human beings, taking into account what hundreds or even thousands of respondents have previously answered. The challenge will be to make sense of these responses and incorporate them appropriately in later exchanges.

Replacing human interviewers

Duplicating expertise

A standalone chatbot interviewing system could exploit some of the suggestions made above (under 'Better memory') to duplicate the developing expertise of a specialist interviewer, enabling lessons learned during a survey, and questions designed on that basis, to be rapidly and reliably deployed to many simultaneous chatbot interviewers. In this way, survey development – even on a very large scale – can move at the speed of the fastest, rather than the slowest, while maintaining coherence and consistency.

Specific applications

Beyond the limited application to open-ended questions in this exploratory study, more challenging types of qualitative research could be explored, applying chatbot technology or other A.I. systems to enable efficient online collection of qualitative data.

Techniques which may one day be applied by an A.I. or chatbot include sentence completion, personification, and thematic apperception tests (which although they can be left to self-completion, yield better results when respondents are guided). Focus groups are likely to be far harder to deal with, given the number of participants, while in-depth interviews are likely to require far too general an understanding of the world for current technology to deal with.

Stretching the imagination

Complex plan-based systems, such as that described by Allen *et al.* (2001), may offer potential in future to facilitate role playing. Such systems hold a model of the task or the world in which the task would be carried out, and work with users using natural language to develop a plan to carry out the task. Examples include kitchen design planning, and rescue and evacuation planning. One could well imagine such a system yielding useful information on how consumers make decisions, particularly for complex high-involvement purchases, where large amounts of time are spent thinking and planning before choosing a product or service. This transcript from Allen *et al.* (2001), of a user and artificial intelligence planning to save a woman in need of medical attention, hints at the potential of such a system in collaborative role-based problem solving.

(1) USR: We need to get the woman in Penfield to Strong.
(2) SYS: OK
(3) USR: What vehicles are available?
(4) SYS: There are ambulances in Pittsford and Webster.
(5) USR: OK. Use one from Pittsford.
(6) SYS: Do you know that Route 96 is blocked due to construction?
(7) USR: Oh
(8) USR: Let's use the interstate instead.
(9) SYS: OK. I'll dispatch the crew.

An example of a marketing research application would be the enterprise software solutions market, where qualitative research would either involve focus groups or in-depth interviews using very knowledgeable interviewers. Business buyers wanting to choose a new enterprise-wide software system would go through a number of complex stages before choosing a system, ranging through issues such as compatibility with existing software systems, training, price and licensing structures, benefits, and supplier stability. A plan-based system could not only assist a buyer, but gather valuable insights into how a buyer would go about buying such a system, and what types of information they are looking for.

In summary

This paper has demonstrated the exciting potential of artificial intelligence, and in particular relatively simple chatbot technology, as an interviewing aid or interviewer substitute. A chatbot called *Elizabeth* was successfully programmed to ask the open-ended question 'Why did you choose your current cell phone network operator?', and to then probe the respondents' answers to obtain richer information. This has particular application when CAWI and CAPI data collection methods are used, as self-completion often results in superficial responses to open-ended questions. Future research directions and areas of application were then explored.

References

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2001), 'Towards conversational human-computer interaction', *AI Magazine* 22 (4), pp. 27-38.
http://www.cs.rochester.edu/research/cisd/pubs/2001/allen-et-al-aimag2001.pdf
Accessed: 6 January 2006.

Grice, H.P. (1975), 'Logic and conversation', in P. Cole and J. Morgan (eds), *Syntax and Semantics*, Vol. 3 (London: Academic Press), pp. 41-58, and reprinted in H.P. Grice, *Studies in the way of words* (Harvard University Press, 1989), pp. 22-40.

Grossman, L. (1998), 'Get smart: how intelligent can artificial intelligence be?', *Time-Digital Magazine*, October 1998,
http://www.psych.utoronto.ca/~reingold/courses/ai/cache/time-digital.html
Accessed: 6 January 2006.

Ivis, F.J., Bondy, S.J., and Adlaf, E.M. (1997), 'The effect of question structure on self-reports of heavy drinking: closed-ended versus open-ended questions.', *J Stud Alcohol* 58 (6), pp. 622-4.

Honan, M. (2000), 'AOLiza', Macworld,
www.macworld.com/2000/08/25/aoliza.html
Accessed: 6 January 2006.

JGuide: Stanford guide to Japanese information sources (2005), Stanford University,
http://jguide.stanford.edu/site/overviews 284.html
Accessed: 16 November 2005.

Millican, P.J.R. and Clark A., eds (1996), *Machines and Thought: The Legacy of Alan Turing*, Oxford: Oxford University Press.

Millican, P.J.R. (2003-5), '*Elizabeth*'s home page', Leeds Electronic Text Centre,
http://www.etext.leeds.ac.uk/elizabeth/
Accessed: 6 January 2006.

Mullins, J. (2005), 'Whatever happened to machines that think?', *New Scientist* 186 (2496), 23 April 2005.

16PF World (2005), '16PF Questionnaire: Primary Factors – Definitions',
http://www.16pfworld.com/primaryfactors.html
Accessed: 10 November 2005.

Pickery, J. and Loosveldt, G. (1998), 'The impact of respondent and interviewer characteristics on the number of "no opinion" answers', *Quality and Quantity* 32, pp. 31-45.

Platt, C. (2001), 'Soul in the machine', *Yahoo! Internet Life* 7 (8), pp. 86-90.

Poulsen, K. (2005), 'A.I. seduces Stanford students', *Wired News*, 1 May,
http://www.wired.com/news/culture/0,1284,67659,00.html
Accessed: 16 November 2005.

Reingold E. and Nightingale J. (1999), 'Cyc', University of Toronto,
http://www.psych.utoronto.ca/~reingold/courses/ai/cyc.html
Accessed: 25 October 2005.

Saygin, A.P. and Cicekli, I. (2002), 'Pragmatics in human-computer conversation', *Journal of Pragmatics* 34 (3), pp. 227-258.

Shieber, S.M. (1994), 'Lessons from a restricted Turing test', *Communications of the ACM* 37 (6), pp. 70-8.
http://www.eecs.harvard.edu/shieber/Biblio/Papers/loebner-rev-html/loebner-rev-html.html
Accessed: 4 November, 2005.

Shieber, S.M., ed. (2004), *The Turing Test: Verbal behaviour as the hallmark of intelligence*, Cambridge, Massachusetts: MIT Press.

Stephens, K.R. (2005), 'What has the Loebner contest told us about conversant systems?',
http://www.behavior.org/computer-modeling/index.cfm?page=http%3A/
/www.behavior.org/computer-modeling/stephens/stephens1.cfm
Accessed: 6 January 2006.

Turing, A.M. (1950), 'Computing machinery and intelligence', *Mind* 59 (236), pp. 433-60.

Van der Zouwen, J. (2001), 'Cybernetics and interviewing', *Kybernetes* 30 (9 & 10), pp. 1064-71.

Van der Zouwen, J. and Smit, J.H. (2005), 'Control processes in survey interviews: a cybernetic approach', *Kybernetes* 34 (5), pp. 602-16.

Weizenbaum, J. (1966), 'ELIZA – A computer program for the study of natural language communication between ,man and machine', *Communications of the ACM* 9 (1), pp. 36-45.

Yang, J. (2005), 'Asian Pop Robot Nation: Why Japan, and not America, is likely to be the world's first cyborg society',
http://www.sfgate.com/cgi-bin/article.cgi?f=/g/a/2005/08/25/apop.DTL
Accessed: 6 January 2006

Appendix: Useful websites

American association for artificial intelligence	http://www.aaai.org/
A.I.	http://www.a-i.com/
CYC	http://www.cyc.com/
Elizabeth's website	http://www.etext.leeds.ac.uk/elizabeth/