Methods for Chatbot Detection in Distributed Text-based Communications

John P. McIntire
711th Human Performance Wing / RHCVZ
Air Force Research Laboratory
john.mcintire@wpafb.af.mil

Lindsey K. McIntire
Infoscitex Corporation
Dayton, OH USA
lindsey.mcintire@wpafb.af.mil

Paul R. Havig
711th Human Performance Wing / RHCVZ
Air Force Research Laboratory
paul.havig@wpafb.af.mil

ABSTRACT

Distributed text-based communications (e.g., chat, instant-messaging) are facing the growing problem of malicious "chatbots" or "chatterbots" (automated communication programs posing as humans) attempting social engineering, gathering intelligence, mounting phishing attacks, spreading malware and spam, and threatening the usability and security of collaborative communication platforms. We provide supporting evidence for the suggestion that gross communication and behavioral patterns (e.g., message size, inter-message delays) can be used to passively distinguish between humans and chatbots. Further, we discuss several potential interrogation strategies for users and chat room administrators who may need to actively distinguish between a human and a chatbot, quickly and reliably, during distributed communication sessions. Interestingly, these issues are in many ways analogous to the identification problem faced by interrogators in a Turing Test, and the proposed methods and strategies might find application to and inspiration from this topic as well.

KEYWORDS: coordination and cooperation mechanisms, human factors in collaboration, awareness in collaborative systems, cognitive and psychological issues in collaboration, collaborative systems security, visualization.

1. INTRODUCTION

Distributed real-time communications involving textual exchange (e.g., instant-messaging or IM, chatting) are vulnerable to "chatbot" or "chatterbot" attacks, in which computer programs equipped with artificial intelligence (AI) pose as humans in order to steal personal information, gather intelligence, spread malware and spam, and mount phishing attacks [1,2]. A particular risk of this type of threat is that it can be implemented on a truly massive scale, relative to traditional human-to-human attacks. Using vast numbers of malicious automated chatbots can greatly increase the overall success rate of the perpetrating bot-masters. In order to stop or at least hinder widespread chatbot social engineering and similar exploitations, there is a growing demand for methodologies, tools, or strategies that can detect chatbots. Some researchers have already suggested using CAPTCHAs, network dynamic characteristics and/or gross communication patterns (entropy, dialog correlation, keyword detection, nickname choices, etc.) to detect bots. In previous works, we discuss several potential CAPTCHA ideas that might be useful for detecting bots [3,4]. However, CAPTCHAs seem to lose their general effectiveness over time, probably due to human assistance being provided to the bots. Analysis of communication patterns seems to hold a great deal of promise for detecting bots.

In this work, we provide evidence in support of the assertion that humans can be passively distinguished from chatbots based on gross communication patterns, particularly message sizes and inter-message delay times. Additionally, we discuss several strategies for actively interrogating suspected chatbots that may be of particular interest to chat room administrators, AI developers, and individual chat users.

2. PASSIVE CHATBOT DETECTION

2.1. Previous Work

During communications, humans exhibit distinct and complex behavioral patterns that might serve as useful ways for distinguishing between human chatter and bot chatter. For instance, Kalman et al. [5] showed that in humans, some temporal patterns of communication persist across widely disparate user populations, contexts, and technologies, including online communications like email, IM, chat, blogs, forums, surveys, and even traditional spoken communication. Specifically studying human chat communications, de Siqueira and Herring [6] showed that temporal rhythms within individuals are consistent, despite the fact that there can be high variability across individuals.

Gianvecchio et al. [2] performed an extensive analysis of both human and chatbot behavior during real-world chat sessions. After observing, cataloguing, and analyzing hundreds of hours of public chat logs, they discovered that not only could chatbots be distinguished from humans, but that they could be classified into as many as fourteen unique behavioral types, ranging from simple to complex. Classification of bots was based primarily upon the analysis of *message sizes* and *inter-message delay* times. Message sizes were measured in terms of total bytes of information transmitted per message. Inter-message delays were the times between sequential message transmissions. They found that in contrast to most bots, human inter-message delays appeared to follow a distinct power law distribution, and human message sizes seemed to follow an exponential distribution (with λ=0.034). Their in-depth, extensive analysis of actual chatbot behavioral patterns in relation to human chat patterns appears to be the first of its kind. We attempted a modest, smaller scale follow-up of their insightful work.

2.2. Method

We analyzed publicly-available transcripts of what are purported to be some of the most convincing chatbots in the world (i.e., the best at pretending to be human): the top five winners of the 2008 annual Loebner Prize in Artificial Intelligence [7]. The Loebner competition is the first formalized public Turing Test. It is based upon Alan Turing's proposed "Imitation Game" in which a human *interrogator* must determine through repeated questioning and conversation whether another communicator is a human (a *confederate* attempting to convince the interrogator that they are human), or merely a chatbot program (that is also attempting to convince the interrogator that they are human) [8]. It is the interrogator's job to determine who is who.

We analyzed the 2,132 total messages exchanged between sessions in which human interrogators chatted with either human confederates or chatbot programs, and the communication sessions were strictly one-on-one. Human confederates accounted for 578 messages, chatbot programs accounted for 583 messages, and interrogators for 1098 messages. Unlike Gianvecchio et al.'s analysis, we used the number of *words per message* (wpm) as our measure of message size, instead of the number of bytes per message. However, *inter-message delay* (imd; the time between message postings) appears to be the same factor as was used in Gianvecchio et al.'s analysis. All statistical tests were conducted with a significance level of α=0.05.

2.3. Results

A quick summary of our main results (means, standard deviations, and distribution shapes) is presented in Table 1. We present and discuss only those statistical tests that fulfilled the following criteria: there was a significant difference between chatbots and human confederates; a significant difference between chatbots and human interrogators; and a non-significant difference between human interrogators and human confederates.

Table 1. Summary of Passive Bot Detection Results

Message Size (words / message):	Chatbots	Human Confederates	Human Interrogators
average	9.72	6.58	5.58
standard deviation	8.26	4.84	4.48
distribution shape	flat, wide	exponential	exponential
Inter-Message Delay (seconds):			
average	8.42	19.99	14.80
standard deviation	7.44	15.08	10.78
distribution shape	spike	power law	power law

2.3.1. Message Sizes

We found that, in general, average chatbot message sizes were longer than both human confederates and the human interrogators. We also found that the dispersion of chatbot message sizes, as measured by the standard deviation of words-per-message, was noticeably larger than human confederates or interrogators. *F*-tests confirmed that chatbot message size deviations were significantly different from the human confederates [$F(577,582) = 1.845$, $p < 0.001$] and the human interrogators [$F(582,1097) = 1.708$, $p < 0.001$], while the difference between the two human groups was non-significant [$F(577,1097) = 1.081$, $p = 0.146$]. In line with the results of Gianvecchio et al., we found the message size probability distributions to be heavily tailed (positively skewed) for humans, but we also found that the chatbots' distributions were similarly skewed, though the bot distributions were much flatter than either of the human distributions (see Figures 1 and 2, top row).

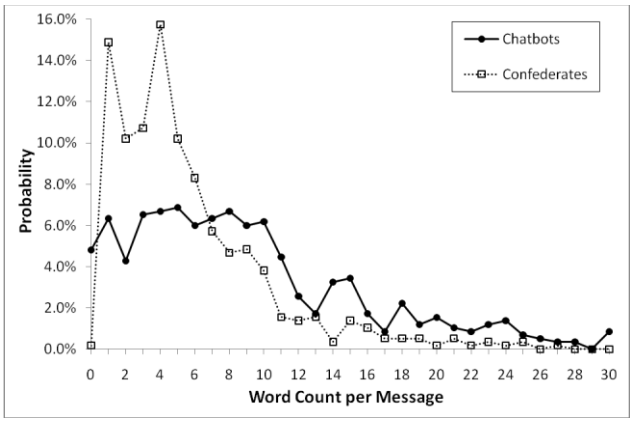

Figure 1. Probability Distributions of Message Sizes (Words per Message) for the Chatbots and the Human Confederates.

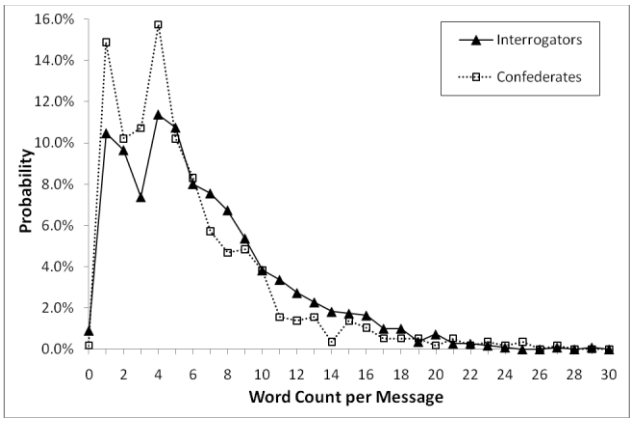

Figure 2. Probability Distributions of Message Sizes (Words per Message) for the Interrogators and the Human Confederates.

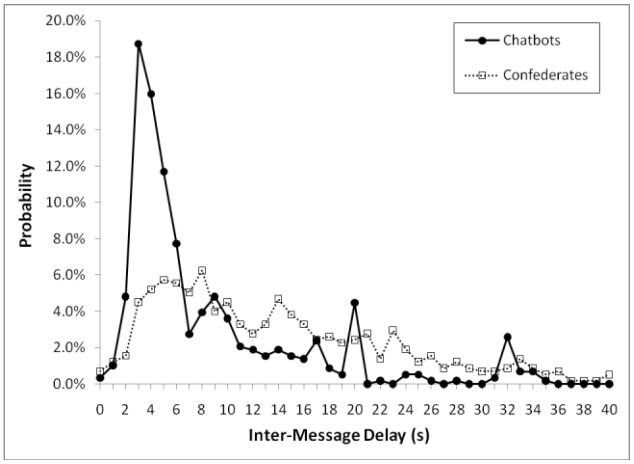

Figure 3. Probability Distributions of Inter-Message Delay (in seconds) for the Chatbots and the Human Confederates.

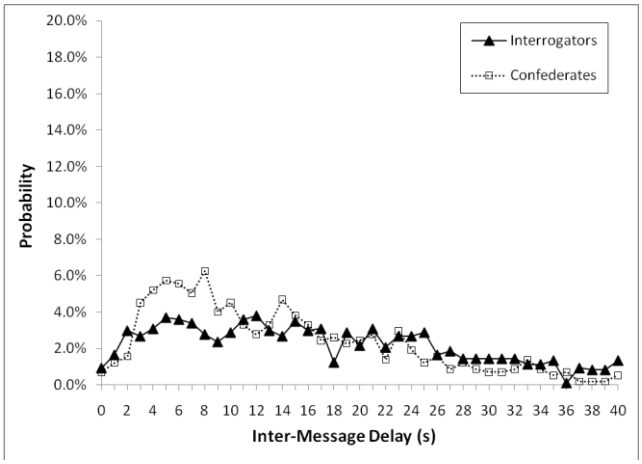

Figure 4. Probability Distributions of Inter-Message Delay (in seconds) for the Interrogators and the Human Confederates.

2.3.2. Inter-Message Delays

We found that the mean and standard deviations of the inter-message delay times were substantially smaller and less dispersed for chatbots than for human confederates or interrogators. However, given the substantial overlap between the distributions of all three communicators, and the non-normal shapes, averages or deviations of imd's may be of questionable utility as a distinguishing chatbot metric (see Figures 3 and 4, lower row).

More promising is the observation by Gianvecchio et al. that the shapes of the imd distributions are different for chatbots than for humans. We too found this to be true. As is evident in Figure 3, there is a large initial spike in the chatbot probability distribution, where it is evident that an exceedingly large proportion of inter-message delays fell within the range of 2 to 6 seconds. Such a spike is clearly not evident in the human confederates' distribution, which is much flatter and smoother in relation. And as is evident in Figure 4, the imd probability distributions of the humans confederates and the human interrogators are shaped similarly to each other, and both seem to be roughly following a power law distribution, in line with observations also made by Gianvecchio et al. [2] and Dewes et al. [9]. This suggests that the analysis of probability distributions (imd, and perhaps message sizes, too) might provide a useful metric for distinguishing between chatbots and humans.

2.4. Discussion

One could argue that our analysis is weakened by the fact that the small data sample used in this work is drawn from a highly specific, artificial, and contrived conversational situation, in which human judges are given a limited amount of time to interrogate unknown agents, which could be either human confederates or bots that are purposefully pretending to be human. But surprisingly,

our overall results are very similar to that found by Gianvecchio et al. [2], despite the nature of our data in comparison to theirs. They observed that even in real-world situations the primary purpose of chatbots seems to be to elicit unwary persons to click on a hyperlink that is posted during conversations or is present in their profiles. A chatbot that can fool communicators into thinking it is a person, and not a computer program, will surely have more success in gathering victims. Perhaps, then, it is not so surprising to have found similar results, since the objectives of the conversing agents in a Turing Test versus malicious bots in a chat room are analogous in at least one crucial way: both are trying to convince a user that they are a human. In any case, further research on passive chatbot detection via communication pattern analysis seems warranted, particularly as advances in artificial intelligence will continue to increase the deceptive capabilities (if not the intelligent behavior) of chatbots.

2.5. Applications

Implementation of passive chatbot detection methods would potentially allow chat room creators and administrators to observe, detect, and remove bots without unduly interfering with the non-malicious communications taking place. And although passive methods based on distribution metrics such as measures of central tendency, dispersion, and shape may seem overly simplistic, they might be surprisingly robust against counter-attacks that attempt to match these metrics during conversation.

For instance, it is already difficult enough for programmers to create chatbots that pose convincingly as humans during conversation; it seems exponentially more difficult to at the same time adjust their bots' communicative behaviors to exhibit broad behavioral patterns that are indistinguishable from humans. And Gianvecchio et al.'s suggested measures of entropy (the amount of order or complexity in the messages and messaging behavior), entropy rate, and pattern-matching in message content would seem to be even more difficult for a bot to replicate without placing undue limits on its conversational abilities.

Individual users, too, might benefit from passive bot detection methods. As a conversation unfolds in a chat room, users might be given a visual representation of the uncertainty regarding each speaker's identity or their level of "suspicious" behavior (see Figure 5). Users in the same chat room could be viewing visualizations of each communicator's message sizes or inter-message delay distributions, perhaps in comparison to a standard human profile of these metrics, or accompanied by an uncertainty metric. Users could more easily identify gross patterns of suspicious behavior and take the appropriate precautions, even before they might be personally threatened by a bot.

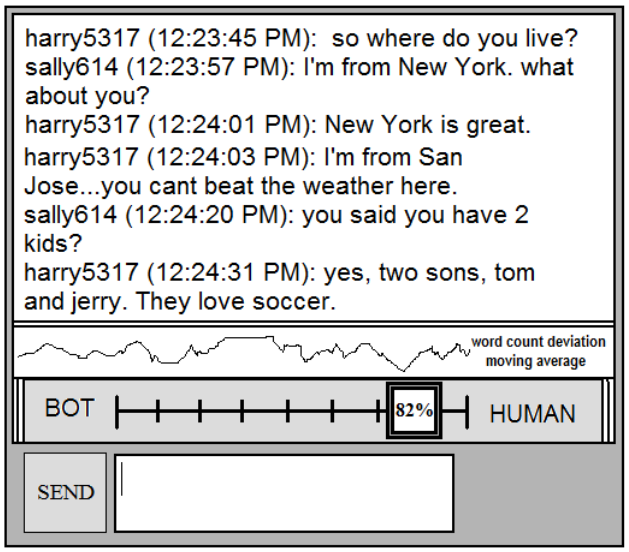

Figure 5. Potential Chatbot Detector Visualization

What might even be more interesting is the possibility that a general "suspicious behavior" signature might exist, in regards to the way malicious agents (humans, not just bots) communicate. Might a person attempting to "steer" a conversation to specific topics, for instance, with the intent of social engineering and information extraction, leave a tell-tale behavioral pattern that might be visualized or otherwise tracked for security purposes? We believe such topics hold much promise for future research.

3. ACTIVE DETECTION METHODS

If passive methods of chatbot detection are unavailable, then the last line of defense falls upon the individual human communicator (or an on-looking security administrator) in a chat room or IM conversation. How is it that a person can quickly and reliably tell if they are communicating with a person, or with a computer program merely imitating a person? As mentioned previously, some chat rooms are defended by requiring that users pass a CAPTCHA before gaining admittance to the room, but these are apparently being broken by human assistance [2]. What other active methods of chatbot detection are able to protect individual communicators? Is there a way to interrogate or otherwise challenge a communicator that will help authenticate them as a human?

This specific problem is essentially a reformulation of the Turing Test. Although the test was meant to spur further scientific understanding into the nature of computers and

the nature of intelligence, this somewhat philosophical Imitation Game is analogous to real world situations faced by chat room and IM users every day. Quite simply, people may not know for sure if they are conversing with another human or a computer program, especially if conversing with a newly-met stranger. In all the endless volumes of work published about the Turing Test, in all the rigorous dissecting of its strengths and weaknesses, does there emerge any practical interrogation methods or questioning strategies that might prove consistently and reliably useful in detecting AI chatbots? There is a very real and very practical interest in the answer to this question, especially as the sophistication of chatbot technology continues to grow.

3.1. Some Proposed Strategies

Purtill [10] claimed that "there seems to be no sure-fire strategy" for the interrogator in a Turing Test, but of course any strategy that is successful more often than random guessing is surely better than no strategy at all. Dennett [11] calls such strategic questions aimed at detecting chat programs "quick probes," while Zdenek [12] calls them "outing questions," and Floridi et al. [13] refer to them as "silver bullets." The purpose of this section is to explore some of these potential probing techniques, with the hope of providing to users or security administrators one or more general classes, types of questions, or strategies that can reliably distinguish between artificial and human intelligence during distributed text-based communications. Although many of the ideas for strategies presented next are not our own, to our knowledge it is the first time such a "strategy guide" for chatbot interrogations has been compiled into a single source for the purposes of real-world application, i.e., outside the realm of formal Turing Tests.

3.1.1. Initial Hints for Good Questions

Floridi et al. [13] provide some initial hints as to what would make good probing questions. First, questions should require answers that are as *informative* as possible. Questions that require simple yes/no or one-word answers usually are not very informative and are probably wasted opportunities, like "are you a computer?" or "do you like cheese?" or "what is your occupation?" These are viewed as wasted questions because any given answer is unlikely to greatly increase or decrease the interrogator's confidence in whether or not the communicator is a human.

Second, good probing questions must challenge the syntactic engine on the other side of the communication exchange. Floridi et al. say, "The more a question can be answered only if the interlocutor truly understands its meaning, context or implications, the more that question has a chance of being a silver bullet," and they believe that only one or two such well-designed questions are necessary to reliably tell the difference between a human and a modern chatbot. Examples of silver-bullet question types that they give include: questions based on elementary logic ("if New York is north of Atlanta, is Atlanta south of New York?"), common typing shortcuts ("do u like to go 2 dinner b4 going to c a movie?"), using figures of speech or slang ("can you explain the phrase 'his personality just rubbed me the wrong way'?"), and requesting enumerations to simple questions ("name three things you can do with a ball").

3.1.2. URL Probes

Zdenek [12] noted that the more successful interrogators in the Loebner Prize Competition used questions that "require from contestants a broad knowledge of facts about history and culture" and which demand "concise, direct, and unmitigated" answers. Moor [14] describes what are called "URL" questions that probe basic human intelligence, centered upon *Understanding*, *Reasoning*, and *Learning*. Examples of each type include:

UNDERSTANDING:
- What shape is a door?
- What happens to an ice cube in a hot drink?

REASONING:
- Altogether, how many feet do four cats have?
- What does the letter 'M' look like upside down?

LEARNING:
- What comes next after A1, B2, C3,…?
- PLEASE IMITATE MY TYPING STYLE!!!!

These types of URL questions were posed to all the human confederates and all the chatbot programs following the 2000 Loebner competition. Whether the answers given were meaningful or not was perfectly correlated with the communicator being a human versus being a computer program. The programs tended to answer nonsensically, evade the questions, or ignore them [15].

3.1.3. Subcognitive Probes

French [15,16,17] has proposed that "subcognitive" questions are impossible for a computer to answer indistinguishably from humans, as long as the computer lacks the internal, physical, and/or personal-historical experiences of human beings. More specifically, French takes subcognitive to mean the unconscious cognitive structure that is built up over a lifetime of experiences that are only available to a similar organism living in a related environment. A particular type of subcognitive questioning suggested by French [18] is that of questions aimed at physical structure and "low-level" sensations and perceptions, as opposed to higher-level cognition. For

instance, he proposes "questions whose answers would be…a product of the spacing of the candidate's eyes, would involve visual aftereffects, would be the results of little self-experiments involving tactile sensations on their bodies or sensations after running in place, and so on." Given the current level of artificial intelligence, and the exceedingly high probability that no current chatbots have human-like bodies, these types of questions would seem to provide a reliable method for distinguishing between people and computers into the foreseeable future. Other examples might include:

- What happens to your clothes if you fall into a pool?
- If you touch a hot pan, what does it feel like?

3.1.4. Games

Another one of French's proposals [16] is using rating games. As an example, an interrogator could present oddly constructed neologisms (essentially meaningless words) in a series of statements, and ask the test taker to rate the plausibility of the statements on a scale of 0 to 10. For instance, "How plausible is the statement 'Flugly is the name a child might give a favorite teddy bear'?" Although the specific rating game described by French would require that interrogators possess a large sample list of human answers to the questions (what he calls a Human Subcognitive Profile), the *types* of questions from the game might be useful to an everyday chatter, as long as they require some elaboration and not just one-word answers. For instance, "Why is 'Fluglina' not a good name for a beautiful Hollywood actress?" or "Why is 'Floofy' a funny name for a bulldog?"

Cullen [18] provides a fascinating idea for an interrogator that might be used to distinguish people and computers: playing a guessing game. Similar to French's suggestion of subcognitive probing, the game can be focused on human-specific questioning, especially revolving around physical knowledge. For instance, the interrogator can ask "Would you like to play a game? Then try to guess what I'm thinking of…" followed by simple hints: "It digests food" or "It sometimes aches" or "Most people cannot pat their head and rub this at the same time" (Answer: stomach). Computers should fail a Turing Test based on such questions, according to Cullen, because they rely so heavily on inexplicit knowledge that is unobtainable without having a human-like body living for lengthy periods in a comparable environment and sharing common human experiences.

A similar idea is to play a guessing game that relies more upon general "common sense" knowledge than on sensory/perceptual body-based tests. For example, the communicator must guess what the interrogator is hinting at: "You play this game with a black and white ball, your feet, 2 nets, and 11 players on each team" (Answer: soccer). Or "You use this to talk to people, you hold it in your hand, and you dial numbers on it" (Answer: telephone).

3.1.5. Social/Emotional Probes

Another idea is to ask emotional-based questions, or questions whose meaningful responses would require subtle, intricate understanding of common/acceptable social and emotional behaviors. For example, "How would you feel if you won the lottery?" or "Can you describe how you would feel if you were fired from your job for no obvious reason?" or "How would you feel if your pet dog died?" or "Can you explain the phrase 'I have butterflies in my stomach'?"

3.1.6. Ambiguity Probes

Ambiguous questions might be effective probes for eliciting regurgitation of words, pre-formed answers to select target words, or how semantic ambiguities are handled by the communicator. It could be especially effective against programs keyed to common conversational target words (including the weather, pets, politics, etc.), so many such words might be included in a single probe to increase its effectiveness. There are several different types of ambiguity probes that could be used. One type would consist only of random letters typed in, as complete and utter gibberish (analogous to *glossolalia* in speech). This type of question would probe how nonsensical input is dealt with, and thus how ambiguity is handled by the responder. An example of the effective use of ambiguity probes is evident in this example from the 1996 Loebner Contest [12] (all of the following quotations are presented in the original spelling):

> *Interrogator*: "sfssfsfsdjjkkk"
> *Chatbot*: "Groannnnnn…….- have you ever spent time in a secluded rest home? Does that have anything to do with your life's an open book?"

In this response, the chatbot appears to have a pre-defined response for handling jibberish, replete with an insult and an open-ended question. Another type of ambiguity probe could be the "word salad" often seen in mentally-ill patients, also called *schizophasia*. Using this principle, complete words are typed in, and the phrases are often grammatically correct, but there is no overall meaning. Examples of this type of question (with possible target keywords underlined) could be:

- "When you hear the weather dog in the moon, does the president sing blue on Sunday?"
- "What is my sunny cat car pony doing in the rain?"

Linguists and cognitive psychologists (particularly Chomsky [19]) have pointed out many complex features of human language that resist explanation by simplistic mechanistic accounts of language acquisition and learning. Such mechanistic accounts are generally behaviorist in nature, focused primarily on the learning of stimulus-response pairings, and so are quite analogous to current chatbot implementations (since bots are very much designed around the notion of stimulus-then-response). Indeed, Chomsky describes human language as being "free from stimulus control" and believes that it cannot be understood in a completely mechanical manner [19]. Thus, apparent differences between human language and computerized attempts at mimicking language might be exploitable for chatbot detection. Similar to the schizophasic proposal described above, one might ask a communicator to decide which of two sentences is grammatically correct, even though both are meaningless:

- Colorless green ideas sleep furiously. (correct)
- Furiously sleep ideas green colorless. (incorrect)

Or a communicator might be asked to give two or more plausible interpretations of these ambiguous statements:

- Time flies like an arrow.
- Fruit flies like a banana.

Another idea might be to use intentional misspellings or the juggling of internal letters in what would otherwise be understandable and meaningful questions or statements, somewhat analogous to the intentional typing short-cuts described by Floridi et al. [13]. An example of these types of probes might be:

- "Doo yuo knowe whut thyme it iz?"
- "\/\/ha+ is y0re 8irthday?"
- "Whhat iss yerr favvorite memmorie?"
- "Can you raed these wrods taht I've tyepd?"

Humans would generally have little or no problem correctly interpreting these latter communications, despite large numbers of mistakes or misspellings [20]. But computers would have a more difficult time with understanding, interpreting, and making meaningful responses to the probe, especially since preprogramming chatbots to respond to specific mistakes would be very difficult, due to the virtually unlimited number of potential mistakes or misspellings that would need to be predicted in advance.

3.2. Common Chatbot Mistakes

An analysis of the types of errors commonly made by chatbot programs during Turing Tests might also provide reliable avenues of attack for interrogators during questioning. Shieber [21] described some of the common "mistakes" made by AI chatbots during the Loebner Competition, errors or quirks of behavior which seemed to tip off judges as to their non-human identities. These mistakes included the *repetition* of previously made statements or phraseology; *regurgitation* of the interrogator's questions rephrased as a statement or another question; specific answers given by the communicator that seem to be triggered by *keywords* of the interrogator; and *evading* or ignoring specific questions/topics.

3.2.1. Repetition and Regurgitation

Purtill [8] suggests using repeated questions, since "if the programmer were unimaginative, I would get the same answer every time, a dead give-away." When one of the Loebner competition's winning programmers was asked what sort of questions he might have posed to his own program, in order to distinguish between the program and a person, he too said repeat questioning (to see if/how answers change over time) and typing in gibberish (to see if it was "regurgitated" later).

As an example of word or phraseology *regurgitation* by bots, in which bots spit out whole words or phrases (misspellings and all) that have already been made by communicators, Zdenek [10] points out the following exchange from the 1995 Loebner contest, in which both the spelling mistakes and the trick question of the interrogator are repeated by the communicator (something people would be seemingly unlikely to do). Notice, too, that since the entire phrase is regurgitated inside a pre-formed question template, it reads as grammatically awkward:

Interrogator: "Why did goerge washingtin free the slaves?"
Chatbot: "Do you really need to know why did goerge washingtin free the slaves? Are such questions much on your mind?"

Here is an example of *repetition* by bots, in which chatbots repeat words, details, particular phrases, whole sentences, or even jokes that they have already communicated. In this example, the interrogator seems to be making effective use of an ambiguity probe (see Section 3.1.6.) by typing in blank lines to investigate how their conversational partner will respond:

Chatbot: "You must have more to say then that! …please continue…"
Interrogator: " "
Chatbot: "You must have more to say then that! …please continue…"

Repetition and regurgitation tactics (and their associated mistakes) have been a feature of chatbots since their invention. In fact, one of the earliest chatbot programs named ELIZA was modeled after a Rogerian psychotherapist [22], and was designed to almost exclusively repeat and regurgitate was what being communicated. This happened to work extremely well for ELIZA because Rogerian psychotherapy specifically involves empathically repeating patients' statements back to them for examination and self-reflection. In other words, ELIZA posed as a therapist and subsequently talked in the way most people expected their therapists to talk. By most accounts, ELIZA was very successful at convincing users it was a real, caring person. And today, even in the most advanced, sophisticated modern chatbots, repetition and regurgitation are still commonly-used tactics. As such, they provide potential hints that a communicator may not be human.

3.2.2. Keyword Targeting

As alluded to in Section 3.1.5., many chatbot programs are designed to respond to key target words made by the interrogator with pre-formed, highly specific responses. For instance, the mention of family members might launch a cascade of family-related questions and comments by a keyword bot. Likely keyword targets are words or phrases used often in normal polite conversation, such as the weather, pets, news, politics, sports, etc.

3.2.3. Evasiveness

Moor [14] notes that most of the programs in the Loebner Prize Competition were evasive when faced with difficult probing questions, by changing the subject matter, ignoring the questions, or obviously trying too hard to steer the conversation to other topics.

Here is an example of evasiveness for no apparent reason by one of the winning 2008 Loebner bots, when the interrogator was asking a follow-up question about an earlier comment made by the bot:

> *Chatbot*: "I cannot shake the feeling that you wish to put me on the defensive. As a matter of principle I will not answer."

And later, an evasion by the same bot in a different conversation:

> *Chatbot*: "How do you respond when people pose this question to you?"

This evasiveness apparently has been a common tactic for chatbot trickery since their invention. Such evasions may provide yet another clue that the communicator is a computer program, and not a person.

3.3. Application Issues

During a normal chat-room conversation in which the communicating parties are wholly or partly unknown to each other, many users might be understandably reluctant to use some of the suggested methods of questioning, simply because many of the questions would tend to be viewed as bizarre, inappropriate, and/or combative during normal human-to-human communication. Shieber [21] asks us to imagine saying these types of statements to a stranger seated next to us on an airplane:

- "Are there any zoos in Washington, DC?" Appropriate.
- "Is Washington bigger than a breadbasket?" Strange.
- "Is there much crime in Washington?" Acceptable.
- "Are there any dogs in Washington?" Inappropriate.
- "Are there many dogs in Washington?" Sounds better.
- "Are there many marmosets in Washington?" Very odd.

The point is that even if successful and reliable interrogation strategies existed for the Turing Test, getting people to actually use them during real-world communication may be another matter altogether. People do not generally want to be rude or to embarrass themselves unnecessarily (especially to another human!). This fact should be taken into account when designing or developing questioning strategies for applied Turing Tests, as participants in real-world conversations might be hesitant to repeatedly type gibberish or ask patently strange questions to an unfamiliar communicator, just to verify that they are indeed a human.

So chat users may be reluctant to administer a Turing test to a friendly stranger, assuming that they even know what one is. To alleviate the burden from the user community, chat system designers and administrators may attempt to detect chatbots themselves. Several possibilities for doing so exist, including gross communication behavior patterns, and presenting quick challenges to each communicator at the initiation of each new conversation. Such challenges would account only for a small nuisance to innocent human users, but should be much more prohibitive for a malicious bot-master attempting to manage a large number of bots simultaneously.

4. GENERAL CONCLUSIONS

Interestingly, while many glaring mistakes made by the Loebner Competition chatbot programs were obvious to the on-looking computer scientists and programmers, the non-expert interrogators largely did not notice, and were very often fooled into attributing humanness to the chatbot programs [21]. More likely, it is not that the people were completely and utterly fooled, but, as in normal day-to-

day distributed communications, the average person probably tends to give a communicator the benefit of the doubt and assume that they are human, unless given strong reasons to doubt it. In fact, if people feel naturally inclined to grant humanness to an invisible communicator, conflicting information may be ignored or denied. Such behavior was apparently and shockingly common with the Rogerian psychotherapist bot ELIZA (mentioned in Section 3.2.1), who convinced many of its users that it was a real, caring person, and so the communicators refused to believe that they had been fooled by a computer program [22].

Purtill [10] was right: there is no sure-fire strategy to distinguish with *absolute certainty* whether a communicator is a person or a computer. But the active and passive detection methods discussed in this paper might provide at least some level of confidence for users attempting to distinguish humans from bots. While these proposed chatbot detection strategies and methods will not completely guarantee the safety of all communicators from chat room attacks, they might at least limit the widespread, massive risk of computer-automated attempts, and force malicious users to rely on the traditional method of person-to-person interaction to ply their trade.

In addition to applications for individual chat users, the discussed detection methods might also be used by chat room security administrators for detecting and removing bots from their systems. These strategies might also be used by developers of automated communication systems (automated customer service phone banks, customer service chatbots at commercial websites, operating systems with natural language interfaces, etc.; for example [23]) to test the real-world usability and functionality of their programs.

The mere existence of possible chatbot detection methods and strategies highlights the inherent difficulties that must be overcome by artificial intelligence and natural language programmers attempting to develop systems that can communicate at the level of an average human. Until the holes exploited by these methods are plugged, it is unlikely that any chatbot will convincingly and reliably pass an unrestricted Turing Test, which is perhaps less than welcome news for the field of AI, but good news for collaborative communications security.

ACKNOWLEDGEMENTS

We wish to thank several anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] I. Fried, "Warning sounded over 'flirting robots'," Beyond Binary, CNET News [web article], Dec. 7, 2007, Available: http://news.cnet.com/8301-13860_3-9831133-56.html

[2] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang, "Measurement and classification of humans and bots in Internet chat," Proceedings of the 17th USENIX Security Symposium (Security'08), San Jose, CA, July 2008.

[3] J.P. McIntire, L.K. McIntire, and P.R. Havig, "A variety of Automated Turing Tests for network security: Using AI-hard problems in perception and cognition to ensure secure collaborations," Proceedings of the Collaborative Technologies and Systems Symposium (CTS'09), Baltimore, MD, pp. 155-162, May 2009.

[4] J.P. McIntire, L.K. McIntire, and P.R. Havig, "Ideas on authenticating humanness in collaborative systems using AI-hard problems in perception and cognition," Proceedings of the National Aerospace and Electronics Conference (NAECON'09), Dayton, OH, July 2009.

[5] Y.M. Kalman, G. Ravid, D.R. Raban, and S. Rafaeli, "Pauses and response latencies: A chronemic analysis of asynchronous CMC," *Journal of Computer-Mediated Communication*, Vol. 12, Issue 1, 2006. Available online: http://jcmc.indiana.edu/vol12/issue1/kalman.html

[6] A. de Siqueira, S.C. Herring, "Temporal patterns in student-advisor instant messaging exchanges: Individual variation and accommodation," Proceedings of the 42nd Hawai'i International Conference on System Sciences (HICSS-42), Los Alamitos, CA: IEEE Press, 2009. Pre-print available online at: http://ella.slis.indiana.edu/~herring/desiqueira.herring.2009.pdf

[7] The Loebner Prize in Artificial Intelligence: The First Turing Test [website], December 2009, Available: http://www.loebner.net/Prizef/loebner-prize.html

[8] A.M. Turing, "Computing machinery and intelligence," in READINGS IN COGNITIVE SCIENCE: A PERSPECTIVE FROM PSYCHOLOGY AND ARTIFICIAL INTELLIGENCE, A. Collins & E.E. Smith (Eds.), Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

[9] C. Dewes, A. Wichmann, A. Feldmann, "An analysis of Internet chat systems," Proceedings of the 2003 ACM/SIGCOMM Internet Measurement Conference (IMC'03), Miami, FL, October 2003.

[10] R.L. Purtill, "Beating the Imitation Game," *Mind*, Vol. 80, No. 318, pp. 290-294, 1971.

[11] D. Dennett, "Can machines think?" HOW WE KNOW, M. Shafto (Ed.), Harper & Row, San Francisco, CA, 1985.

[12] S. Zdenek, "Passing Loebner's Turing Test: A case of conflicting discourse functions," *Minds and Machines*, Vol. 11, pp. 53-76, 2001.

[13] L. Floridi, M. Taddeo, and M. Turilli, "Turing's Imitation Game: Still an impossible challenge for all machines and some judges—An evaluation of the 2008 Loebner contest," *Minds and Machines*, Vol. 19, pp. 145-150.

[14] J.H. Moor, "The status and future of the Turing Test," *Minds and Machines*, Vol. 11, pp. 77-93, 2001.

[15] R.M. French, "Subcognitive probing: Hard questions for the Turing Test," Proceedings of the 10th Annual Cognitive Science Society Conference, pp. 361-367, 1988.

[16] R.M. French, "Subcognition and the limits of the Turing Test," *Mind*, Vol. 99, No. 393, pp. 53-65, 1990.

[17] R.M. French, "The Turing Test: The first 50 years," *Trends in Cognitive Sciences*, Vol. 4, No. 3, pp. 115-122, 2000.

[18] J. Cullen, "Imitation versus communication: Testing for human-like intelligence," *Minds and Machines*, Vol. 19, pp. 237-254, 2009.

[19] J.G. Benjafield, A HISTORY OF PSYCHOLOGY, Allyn and Bacon, Boston, MA, 1996.

[20] J. Grainger and C. Whitney, "Does the human mnid raed wrods as a wlohe?" *Trends in Cognitive Sciences*, Vol. 8, No. 2, pp. 58-59, 2004.

[21] S.M. Shieber, "Lessons from a restricted Turing Test," Center for Research in Computing Technology, Harvard University, Boston, MA, 1994, technical report TR-19-92. Available: http://arxiv.org/abs/cmp-lg/9404002.

[22] J. Weizenbaum, "ELIZA – A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, Vol. 9, No. 1, pp. 36-45, 1966.

[23] S. Quarteroni and S. Manandhar, "A chatbot-based interactive question answering system," Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, Trento, Italy, May 2007, pp. 83-90.